Going beyond archiving—
A collaborative tool for typological research

Alexander Borkowski and Andrea C. Schalley

Introduction[1]

> In his *Einleitung in das gesamte Sprachstudium* ('Introduction to the general study of language', 1810/11), Wilhelm von Humboldt complains about the lack of a general comparative grammar and the abundance of judgements that all too obviously lack the firm ground of properly established leading ideas. Almost two hundred years later we know much more about language, but the project of a general comparative grammar based on the firm ground of properly established leading ideas is still far from completion and the issue has become much more urgent given the increasing speed of extinction of languages in the course of globalization. (Zaefferer 2006: 113)

The provision of a 'general comparative grammar' has indeed been an important aim of linguistic typology, as demonstrated by, e.g., the Lingua descriptive studies questionnaire (Comrie and Smith 1977), and more recently, efforts such as *The World Atlas of Language Structures (*WALS*)* (Dryer and Haspelmath 2011) and the *Typological Database System (*TDS*)* (Dimitriadis *et al.* 2009) (cf. also §3). While we, as Zaefferer comments, might know a lot about language in general by now, and might know a lot about specific languages, it is still obvious to the typologist that we are not even close to being able to compare languages in a systematic, efficient, data-driven way, and we are not able to flexibly test hypotheses against a broad set of typological data. Given the incredible language diversity around the world, such a task can only be addressed through the intelligent use of computational technologies that are available to us nowadays, allowing not only for faster and more flexible dissemination paths and tailored access to data and its analysis, but also for concerted efforts in collaboratively collating relevant typological data and knowledge.

In this paper, we will introduce a new collaborative tool for typological research, TYTO, concentrating on the question of how sustainability can best be achieved for

[1] We would like to thank two anonymous reviewers for very constructive comments on an earlier version of this paper.

such a tool. How does the tool's design support data sustainability? What design aspects are crucial for this? The paper is structured as follows: We will describe the background and aims of the tool in §1, including a brief overview of the workflow. §2 will then concentrate on how sustainability is addressed, discussing a number of different facets of sustainability. This includes a discussion of data storage formats, knowledge base design, user interface and workflow modelling, and system output. In §3 we will contextualise it in relation to other similar projects, before concluding in §4 with an outline of some problems and a brief outlook on envisaged future development.

1. A collaborative tool for typological research—overview

The development of TYTO started as part of the project 'Social Cognition and Language: The design resources of grammatical diversity', funded by the Australian Research Council (DP0878126). One of the aims of the project is to develop a sophisticated model of the diverse solutions the world's languages have evolved for expressing 'social cognition', i.e. for the capacity to represent and reason about agents and events in our social universe, and to interact with others by building a shared mental world (cf. e.g. Goody 1995; Enfield and Levinson 2006). This model is to be presented in such a way that other scholars (linguists and potentially scholars from other disciplines, such as anthropology, sociology, and artificial intelligence) can readily access it and query the data. This is best achieved through a knowledge base that is both downloadable from the Internet and can be directly queried on the Internet. TYTO was conceived to provide the required infrastructure for this research project[2].

TYTO is a computational system designed to support typological work and linguistic theory building by integrating cross-linguistic data analyses. It is grounded in capturing meaning rather than structural representation. And while it is decidedly neither a fieldwork tool (in the sense of a tool to support the processing of primary data) nor a grammar-writing tool, we hope that TYTO's ontology-based approach will support some aspects of these tasks as well.

[2] The TYTO software is being developed in the programming language Java. It consists of a set of code libraries which form the basis for plugin modules for the ontology editor Protégé (BMIR 2011), the reporting framework JasperReports (JasperForge 2000-2010b), and the report designer iReport (JasperForge 2000-2010a). A server program manages the collaboration and archival aspects of the system.

> Current progress in linguistic theorizing is more and more informed by cross-linguistic investigation. Comparison of languages relies crucially on those concepts which are essentially the same across human minds, cultures, and languages, and which therefore can be activated through the use of any human language. These instances of mental universals join other less common concepts to constitute a complex structure in our minds, a network of cross-connected conceptualizations of the phenomena that make up our world. Following more and more widespread usage we call such a system of conceptualizations an *ontology*, and we submit that the most reliable basis for any cross-linguistic research lies in the common core of the different individual human ontologies. This is the basic tenet of all approaches that can properly be called ontology-based linguistics or *ontolinguistics* for short. (Schalley and Zaefferer 2007: 3)

TYTO thus encompasses a knowledge base, a system containing semi-structured information on and from the languages and societies of the world. This knowledge base is sub-divided into several ontologies, each covering a different aspect of the model. These aspects include linguistic example data and metadata (both data source information as well as general metadata such as contributor information), linguistic description (i.e. cross-linguistic form and function information), language background information (e.g. family, size, vitality, but also geographic region [linguistic and political] and society information [economy, religion, tradition]), as well as aspects of human cognition—for the purposes of the ARC project mentioned above in particular social cognition. The latter is our primary contribution to the research project, while the former are foundational for that aim. Some of these foundational ontologies (such as the linguistic description ontology) form a re-usable framework for linguistic description and typological analyses beyond the domain of social cognition and hence provide a general framework for further work in the field.

There are two main workflows relevant for typologists willing to contribute to the project, data entry and data synchronisation. Data entry refers to the actual entry of linguistic example data by a single user into a local copy of TYTO, while data synchronisation is the task of integrating the newly entered data into the collaborative work of several users. We will address these workflows in turn. Linguistic analyses follow a particular pattern: They typically consist of interlinear glossed text to convey the basic analysis of the example data as well as of further explanations in prose. TYTO aims to capture the meaning of such linguistic analyses by linking elements of the linguistic example description entered into the system to elements in the knowledge base which represent meaning. How is this integration and the linking achieved? Partly automatically, in that data entered in interlinear glossed format is parsed on the basis of the Leipzig Glossing Rules (Bickel, Comrie, and Haspelmath 2008), with some additions as necessary for computational processing. Yet this is only the first step, as there usually is a wealth of information which cannot be captured by automated processing of a morphological analysis. In the second step the results of the automated processing (the aforementioned links between example elements and meaning concepts) will be checked by the typologist and where necessary manually revised and amended. This includes identifying further phenomena conveyed by the linguistic example and adding these to the system in a textual description way where necessary. In particular with regards to semantic and functional markup, expert contributions are indispensable. We believe that these cannot be automated to the level required for our purposes[3]. Schalley (in press) comprises a detailed example of the data entry procedure which is beyond the scope of this paper.

The data entry workflow results in a gain of accessible, analysed typological information available in the knowledge base, of linked data in linguistics ready for exploitation such as in the testing of linguistic hypotheses. This testing allows the knowledge base to be queried via a reporting system that permits the user to control any set of variation dimensions (for which there is actual data in the knowledge base) within a query, which is a unique characteristic of TYTO: to our knowledge, TYTO is the first tool that allows such flexible querying across a typological knowledge base. For instance, questions such as how many or which languages exhibit social group distinctions by way of affixes (e.g. showing an in-group classification in suffixes attached to nouns) can be answered, and additionally (i) a list of the languages exhibiting this and (ii) a list of the example sentences contained in the knowledge base can be obtained.

[3] ODIN, the Online Database of Interlinear Text (http://odin.linguistlist.org/), does attempt to automate the collection of interlinear text, but stops short of assigning meanings and integrating the results in a cross-linguistic ontology.

While the data entered by a single user is immediately available locally for querying and reporting, TYTO is targeted at collaborative work carried out by a number of researchers. The second main workflow of data synchronisation begins with sending the newly entered data to a TYTO server and thus updating the knowledge base which is shared with other researchers. It is similar to uploading a file to a web server, but due to the nature of collaborative work, the TYTO server cannot simply overwrite what is there. It instead attempts to automatically merge the new work done by the submitting user with all other work done in the meantime by other collaborators. Should this fail due to conflicting information, a manual merge process can be performed by the TYTO developers in communication with the submitting contributors. Once the merge process has been completed, the TYTO software will update the local copy of the knowledge base with the newly merged data from all contributors.

This user involvement on many levels, as we will argue later, is a critical factor in keeping the knowledge base and the contained data ‘alive’. Particular attention has thus been given to how to best set up the collaborative aspect of the tool and the human-machine interaction. We will discuss more details of this together with the sustainability question in §2.3 and 2.4.

Prasarnphanich and Wagner (2008: 126) point out that there are two major challenges for collaboration systems: ‘the start-up problem (insufficient initial contribution) and discontinuity problem (no continued contribution to grow [the] knowledge base and keep it alive and up to date)’. While the latter has to do with sustainability and will be discussed in §2, the former has been identified as a challenge for our project, as our knowledge base presents a case which Oliver *et al.* (1985: 542) label as ‘accelerating production functions’:

> [A]n accelerating function describes a situation where successive contributions to the collective good yield progressively larger payoffs. As a result, the collective good suffers daunting start-up challenges, but if initial contributions can be obtained, optimization and sustainability can be achieved due to increasing marginal rates of return. (Prasarnphanich and Wagner 2008: 127)

The more analysed linguistic data is already available in the knowledge base, the more valuable it will be for scholars to contribute their own results, as their data will then be found in any relevant queries and reports generated by the system. This means that they will be able to compare their data with all the data already in the system. As mentioned above, TYTO is used for modelling the social cognition domain by the ARC project and hence does not start out as an empty framework. This should provide a sufficient 'critical mass' of data (Prasarnphanich and Wagner 2008) to get TYTO past the start-up problem and to gain interest from the scientific community. We are however prepared to enter further cross-linguistic data from the literature ourselves (also to stabilise the knowledge base structure to some extent), thereby creating a crucial incentive for other scholars to contribute to the ensuing shared effort. For more information on the knowledge base in general and the submission and reporting processes, see Schalley (in press).

Based on TYTO and its flexible and effective computational infrastructure, our aim therefore is to build a knowledge base of analysed linguistic data, data that can be revisited, to the analysis of which pieces of information can be added, and which is integrated into a highly interrelated network of cross-linguistic information. The knowledge base allows for both semasiological and onomasiological entry points into the data (cf. Schalley 2011 and Schalley in press). The building of this knowledge base is expected to take place over a long period of time, and due to the built-in ease of extensibility (cf. §2.2) it is hoped that sooner rather than later the knowledge base will comprise more than just the initial domains and those initial domains will be further refined in the process. Examples of this would be the domain of spatial cognition (an additional domain), or the ontology for the description of constructional signs (a refinement), respectively.

2. Facets of digital sustainability

Within the debate around digital sustainability the following questions take centre stage: How can we ensure that the digital data, i.e. in this case the TYTO ontologies and software, are properly archived, maintained and continue to be available, useful, and grow in the future? How can we ensure that the processes driving the project are sustainable and appropriately managed? Looking for answers to these questions we turned to best practice documents in the area of information archiving both from linguistics (Bird and Simons 2003) and beyond. The Reference Model for Open

Archival Information Systems (CCSDS 2002[4]) proved to be particularly insightful, especially as it provides a comprehensive overview of the general structures and processes involved in digital archiving. Importantly, it expresses a link between the information preserved in the long term in an archive and the community it is being preserved for:

> To avoid confusion with simple ‘bit storage’ functions, the reference model defines an Open Archival Information System (OAIS) which performs a long-term information preservation and access function. An OAIS archive is one that intends to preserve information for access and use by a Designated Community. (CCSDS 2002: 2-1—2-2)

We argue that without the existence of the Designated Community of users any attempt at archiving information will ultimately fail in the long term. Fostering the Designated Community is a pre-condition for keeping the data alive. To this end we aim for user orientation and involvement in all areas of the design and development of the TYTO software and the comprised data, in particular in the areas of data storage formats (cf. §2.1), knowledge base design (cf. §2.2), user interface and workflow modelling (cf. §2.3), and system output (cf. §2.4).

Intertwined with user involvement, other factors that play a role in digital sustainability are *longevity*, *standards conformity*, and *accessibility*. *Longevity* addresses the conceptual aspects, *standards conformity* the technical aspects, and *accessibility* the physical aspects of long-term preservation. We understand data longevity as a characteristic of information to be independently understandable, i.e. understandable to the Designated Community without assistance of the experts who produced the information (cf. CCSDS 2002: 3-1). Care has to be taken to keep the software and the knowledge base meaningful, by way of comprehensively documenting TYTO, which is one of the important tasks that we are undertaking. Standards conformity simplifies the curation of the data, due to well-defined processes and formats, which themselves are being curated elsewhere by relevant standardisation bodies. Also, it facilitates interoperability with other systems, as could potentially be required in the case of data hand-overs or the integration of the data into other archives. In the face of changing technologies, standards conformity is likely to offer

[4] Identical to ISO standard 14721:2003 Space data and information transfer systems—Open archival information system—Reference model.

standardised migration paths to new technologies and hence is less likely to require custom-built migration solutions (or make the data obsolete). TYTO embraces standards conformity, especially in aspects of archiving, by aiming to implement the OAIS reference model and following other best practices as evident e.g. in its choice of file formats (cf. §2.1 below). Accessibility ensures that the software and knowledge base continue to be physically available. A general strategy to achieve this is the mirroring of the information. Our approach slightly diverges from this in that we disseminate the complete package (both software and knowledge base) as widely as possible, including updates.

We will, in the following, examine design aspects of TYTO and how they address sustainability but, for the purposes of this paper, we will not discuss other aspects of sustainability such as legal issues (e.g. licensing) in depth.

2.1. Data storage formats and location

We aim for the digital data we create to be accessible and available in the future. As part of our project we are creating the collaborative software tool TYTO as well as a linguistic knowledge base. These are designed to be two separate but related kinds of digital artefacts, each with its own characteristics and lifetime. Viewed as digital data, the TYTO software itself is being stored and provided in two forms: As (i) an executable program that runs on all major operating systems as well as in (ii) source code format (i.e. text readable by programmers from which the executable program is being generated) under an open license which allows for sharing and re-distribution of the software and at the same time invites contributions from interested developers. As TYTO is undergoing constant improvement and is potentially going to go through several major revisions, the lifetime of its current form is expected to be significantly shorter than that of the linguistic knowledge base and hence the focus lies more on accessibility and use of current technologies in software development than on standards conformity[5]. With regard to storage location, the software is designed to be installed

[5] For example, TYTO is currently not being developed in an ISO standardised programming language. From the point of view of the OAIS logical model for archival information (CCSDS 2002, §4.2.1) this implies that in order to fully archive TYTO (‘Information Object’ in OAIS terms) the programming language documentation (‘Representation Information’) required to interpret the TYTO program source

locally on the contributors' computers rather than being hosted centrally e.g. as web application. This permits use while on fieldwork—disconnected from the Internet—and applies to the server component that manages the data integration and archival processes as well, so interested parties can use the complete TYTO system independently.

TYTO always works with local copies of data. This ensures that all users of the software can always access a complete copy of the knowledge base and of the reports. The data is stored and distributed in the form of plain files whose contents are various flavours of XML (Bray *et al.* 2008). These files are easier to distribute and to share with other users than would files in a lesser known or proprietary format or data held in a relational database system. They also have a longer life expectancy due to the underlying standardisation.

Reports—which contain queries to run and information as to where to place and how to format the results of these queries—used as input by the analytical part of the software (cf. §2.4) are stored in an open XML file format[6]. Yet, the reporting system itself allows for a wide range of output formats including the Portable Document Format (PDF). These generated output documents are however not deemed to be data managed by the project.

The linguistic knowledge base consists of a set of statements made using (a subset of) First Order Logic; these are expressed in a well-known knowledge representation language, the Web Ontology Language (OWL). TYTO stores the knowledge base in the RDF/XML format required to be supported by all OWL software tools (Horrocks *et al.* 2009).

Ultimately we would like our data to be useful to and used by other parties. Hence, we will provide several ways to obtain and use the data without having to install the TYTO software (cf. §2.4). For these purposes we are also in the process of evaluating standardised persistent identification schemes such as Handle (CNRI 2011) or DOI (IDF 2011) for use with our data sets.

code ('Data Object') was not already archived by another reliable organisation such as ISO and hence would have to be included in the TYTO archive as well.

[6] The JRXML format is openly specified, but dependent on the JasperReports reporting engine.

2.2. Knowledge base design

The main focus of the current research project is to create a digital model of human social cognition concepts based on typological analysis of a broad range of languages. We are doing this in the form of the linguistic knowledge base mentioned so far. The sub-structuring of the knowledge base (cf. §1) is reflected in the partitioning of the storage into several files which include each other using an OWL mechanism intended for this purpose.

One facet of the knowledge base design is the realisation that the linguistic example data is something that should be opaque to the knowledge base itself. Our initial approach of actually describing the structure of the example sentences in the knowledge base was deemed to be unworkable by users as there was a lot of data entry and interaction with the graphical user interface required for even the most basic examples. We opted for encapsulating each linguistic example in a self-contained XML data fragment whose structure is based on the general model for interlinear text proposed by Bow *et al.* (2003) and specified by a separate XML schema. This XML fragment is contained in the knowledge base and hence knowledge base statements can refer to it (e.g. which examples does this morpheme occur in?; what are the examples for a given language?), but its sub-structure is not accessible to the knowledge base. This in turn led to the development of a stand-alone software component for accessing and managing linguistic example data which is being used by the reporting system also and potentially can be re-used for other projects. We took a similar approach to source and bibliographic information.

With regard to content, the initial release of the knowledge base is not going to be an empty framework but will contain a ‘critical mass’ of data to gain interest from the scientific community. We believe this is a crucial incentive for other scholars to contribute to this shared effort and hence a measure to prolong the lifetime of the knowledge base.

We are going to provide two different editions of the knowledge base; one consists of all submitted data that passes a set of automated consistency checks. In particular, OWL reasoner software checks the knowledge base for logical inconsistencies resulting from integrating submitted data. The other edition consists only of data from the automatically checked edition which has undergone another quality control process in the form of a peer-review. The reason for this approach is to (i) support 'hot' ongoing development and sharing of results among the contributors and hence increase user (i.e. producer) satisfaction while (ii) maintaining a higher-quality version for users (i.e. consumers) from outside the project or even outside the discipline. We expect this to also aid in keeping and gaining users for the system.

2.3. User interface and workflow modelling

That TYTO decidedly goes beyond archiving and hinges on user interaction and involvement has naturally impacted on exactly those two parts of the system that the user interacts with: with the data entry system (where the user is a contributor or knowledge 'producer') and with the querying and reporting system (where the user is a knowledge 'consumer')[7].

Initially, the producers and consumers of our system will be identical and it is our aim to involve typologists and other linguistic researchers who may start out as consumers to become producers as well. To facilitate this we aim for a system which (i) does not require linguistic users to be trained extensively in system usage, (ii) allows linguists to deploy their standard methods of data entry (e.g. interlinear glossing), and (iii) provides contributors with immediate integration of their own with previously entered data and access to the resulting analysis (i.e. querying) and research potential.

Every user of TYTO has all the data they require locally available on their computer. They can create reports using a graphical report designer interface. They can then compile any report document available locally into a number of different output formats using the data in their own local copy of the knowledge base (cf. §2.4). These tasks are relatively simple to perform and make use of third party software components.

When it comes to data input however, we realised very early on in the project that the user interface of standard tools to work with ontologies (e.g. the Protégé ontology editor, cf. BMIR 2011) is not geared towards linguistic data entry. It took a lot of time to enter tiny amounts of data (e.g. a morpheme consisting of maybe just one letter) in numerous places and to then link all these tiny amounts of data. This was due

[7] We understand the terms 'producer' and 'consumer' as defined in the reference model for Open Archival Information Systems (CCSDS 2002).

to the highly sub-structured nature of linguistic data and its analysis (e.g. morphemic analysis). At the same time such a data entry process was completely unintuitive for linguists. Hence, we decided to make provisions for linguistic data to be entered in the standardised form of interlinear glosses (Bickel, Comrie and Haspelmath 2008). While there were small adjustments necessary to disambiguate the input format for the computer, overall it remained fairly close to 'normal' data entry for the linguists.

Once the initial data entry of the interlinear glossed data is complete, TYTO will, as part of the workflow, display lists of the different parts of the linguistic analysis (like individual glosses, morphemes, or syntactic constructions) and, for the current application domain, the taxonomy of social cognition concepts. It will also look up existing links between linguistic analysis parts and social cognition concepts and display those as well. The user can then add further links between these two areas and thus explicate the meaning of the language data. Now the data is immediately available locally for further testing such as through the reporting and querying system (cf. §2.4).

A contributor can at any later point in time initiate the data synchronisation process. This process ingests the data into the shared knowledge base in several steps. After agreeing to the contribution license, the data will be received by the TYTO server process which will attempt an automated integration of the data into the work-in-progress edition (cf. §2.2). If this automated ingest is successful, the new version will be the basis for automatic ingest of any further submissions by any contributor[8]. Should the automated ingest fail, the contributor's version will be branched off the main line of development and a manual ingest process will be started by notifying the contributor and the curator about the issues with the data that need to be resolved. In case of problematic submissions, once the manual ingest is complete, the contributor's local knowledge base will be brought up to date with the current work-in-progress edition again. (As mentioned above, we will also provide an edition of the knowledge base that incorporates contributions which have undergone an additional quality control process including peer-review.)

[8] The TYTO client software will, as part of the TYTO server's reply, receive the updated knowledge base as basis for all further submissions.

2.4. System output

TYTO is designed to allow users to find answers to a broad range of questions from linguistic typology. To this end it supports a powerful query mechanism based on the underlying semantic structure of the knowledge base. This underlying structure 'is a collection of triples, each consisting of a subject[9], a predicate and an object' and '[t]he assertion of [a] triple says that some relationship, indicated by the predicate, holds between the things denoted by subject and object of the triple.' (Klyne and Carroll 2004). The query language SPARQL (Prud'hommeaux and Seaborne 2008) used by TYTO works by searching for triples (where subject, predicate, and/or object may be left as variables rather than actual values) and then possibly aggregating the resulting matches. This is a very generic mechanism that allows for extremely flexible queries. In a very simplified example, assume that our knowledge base contains triples where the subjects represent languages, the objects represent examples and there is one predicate linking these languages and examples. One could now specify several kinds of queries: (i) 'What examples are there for a given language?' (by specifying a triple search where the triple object is a variable), (ii) 'What language(s) is this given example an example for?' (by specifying a triple search where the triple subject is a variable), (iii) 'What are all the languages that have examples in the knowledge base?' (by specifying a triple search where both subject and object are variables), or (iv) 'What are the ten languages with the most examples in the knowledge base?' (by combining the query in (iii) with the builtin counting, grouping, and sorting features of the query language). If we take into account that the query language can combine several triple searches and the knowledge base contains a diverse range of triples, the system is going to be able to answer questions such as 'What is the maximum size (i.e. number of active speakers) of all languages in the knowledge base which exhibit grammatical kinship markers?' It allows users to phrase and devise queries themselves as well as store and share those queries.

On top of the query layer TYTO includes a reporting layer providing a way to produce formatted output. The reporting layer takes a report, executes the queries contained therein, and fills in the document with the query results which can then be

[9] Note that the terms *subject*, *predicate*, and *object* do not refer to the linguistic notion.

saved in a wide range of formats: Major commercial and open source word processing and spreadsheet packages, PDF, and generic formats such as HTML, XML, CSV, RTF, and plain text. This greatly facilitates inclusion of query results in research outputs. As users can create these reports themselves or on behalf of other users and share them with the community we expect to see high quality reports emerge and again will provide initial ones as contribution incentive.

We envisage several ways to accommodate researchers who might want to obtain and use the data without having to install the TYTO software by using an online querying and reporting facility with the same functionality as local installations of TYTO, but allowing users to choose which edition and version of the knowledge base they would like to use as input data, and a download area on a publicly accessible web server providing access to the complete set of knowledge base and report data files—including earlier but then obsolete versions—as well as the TYTO software itself.

3. Related projects

Two related projects have already been mentioned in the introduction in Section 1, *The World Atlas of Language Structures* (WALS) (Dryer and Haspelmath 2011) and the *Typological Database System* (TDS) (Windhouwer and Dimitriadis 2008, Dimitriadis *et al.* 2009). In this section, we will briefly contextualise TYTO in relation to those two projects, as well as in relation to two further projects, the *Generalized Ontology for Linguistic Description* (GOLD) (introduced in Farrar and Langendoen 2003) and the *Cross-linguistic Reference Grammar* (CRG) (Comrie *et al.* 1993, Zaefferer 2006). We omit from the comparison tools which share some aspects of TYTO but differ in their primary purpose such as the grammar authoring tool Galoes (Galoes n.d.; Nordhoff 2008), the language documentation tool Fieldworks Language Explorer (FLEx) (SIL International 2010; Butler and van Volkinburg 2007), and the interlinear glossed text tools TypeCraft (TypeCraft n.d.) and EOPAS (EOPAS n.d.; Schroeter and Thieberger 2006). The following is not a comprehensive comparison but will highlight selected aspects.

We will start with one of the first projects, the CRG. Its aim is to provide a general format for reference grammars that (i) guarantees an adequate and comprehensive description of any human language under consideration (including sign languages), and (ii) ensures that the description is organised along the same lines for every language, thereby allowing cross-linguistic comparison in a systematic way[10]. Linguistic description is essentially structured along an AND-OR tree[11] and the CRG is specifically targeted at grammatical description, without explicit focus on conceptual or semantic categories, as comprised by TYTO. CRG was one of the first projects that tried to build a system that allows the collation of an impressive body of knowledge on specific languages into an electronic format and to make this available online. While it has been implemented to near-completion, CRG has not been activated, and hence its aim of collating actual grammatical descriptions of diverse natural languages has unfortunately not come to fruition to date. Although TYTO does not aim at collating grammatical descriptions of diverse languages, there is a clear overlap and there are clear differences in the approaches taken, as outlined in Table 1.

CRG	TYTO
predefined AND-OR tree for linguistic description	user-driven, data-driven development of linguistic description apparatus
aims at integrated comprehensive grammatical descriptions of languages (primarily semasiological entry point)—one knowledge base	aims at integrated information on what languages code in which way (semasiological and onomasiological entry points)—one knowledge base
elaborate example data structures based on interlinear glossing with translations, as featured in typological literature (with a detailed interlinear representation format, containing up to 13 representation levels, cf. Zaefferer 2006)	example data structures based on basic interlinear glossing with translations plus additional information entered via linking of highlighted elements or form fields (cf. §1 and Schalley in press)
users not specifically accommodated	specific focus on attractiveness for users, in particular through the targeted input system and the flexible querying possibilities (both contributors and ‘consumers’, cf. §2)

Table 1: Comparison of CRG and TYTO

[10] This information was previously taken from the project's website at http://www.crg.lmu.de/, but this web page has recently gone offline.

[11] An AND-OR tree is a formalism from artificial intelligence for decomposing information or problems into conjunctions and disjunctions of sub-information or sub-problems.

To stay with those systems that aim at producing one knowledge base, GOLD, the Generalized Ontology for Linguistic Description (GOLD 2010), will be addressed next. GOLD is an ontology for descriptive linguistics. It gives an account of the most basic categories and relations used in the scientific description of human language and is intended to capture the knowledge of a well-trained linguist. It can thus be viewed as an attempt to codify the general knowledge of the field. What GOLD aims at, using a top-down approach, is what TYTO strives for in the data-driven linguistic description part of the knowledge base, i.e. TYTO intends to generate its own ontology for linguistic description. See Table 2 for a comparison between GOLD and TYTO.

GOLD	TYTO
top-down approach for terminology development, terms taken from, e.g., the SIL International's online glossary of linguistic terms (Loos *et al.* 2004) and standard linguistic sources such as Crystal (1997) (but there is the option to contribute data via the 'submit issues' function[12])	bottom-up approach in terminology development, terms contributed by contributors through the submission of linguistic data information; collaborative work; those morphological categories pre-set that are required for computational processing of interlinear glossed data (cf. §1)
concepts contained in ontology for linguistic description not extensively cross-linked; mainly taxonomic structure (although this seems to change)	concepts contained in ontology for linguistic description cross-linked
current focus on grammar (semasiological entry point)	allows for both semasiological and onomasiological entry points
not all linguistic terminology linked to example data	all linguistic terminology linked to actual example data
data storage: OWL/XML files	data storage: OWL/XML files

Table 2: Comparison of GOLD and TYTO

GOLD faces the challenge of achieving some consensus around linguistic terminology, and its history suggests that this is not easily accomplished. We also believe that well-defined typological terminology and its consistent usage are crucial for providing a tool that allows flexible and meaningful language comparison. We

[12] Cf. http://linguistics-ontology.org/issue.

hence attempt to foster such consistent use of terminology, e.g. through exploiting the type-instance distinction available in the knowledge base to capture the difference between general categories and subtle language-specific differences (cf. Haspelmath's 2010 distinction into comparative concepts and descriptive categories, and also §4).

The *World Atlas of Language Structures* (WALS) 'is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors (many of them the leading authorities on the subject).' (Dryer and Haspelmath 2011) Currently, extensive cross-linguistic information on 192 features such as 'Tone', 'The Position of Negative Morphemes in SOV Languages', or 'Finger and Hand' is available. 'WALS Online now includes 76492 datapoints for 2678 languages. The feature with the most languages (Order of Object and Verb) now has data for 1519 languages' (Dryer and Haspelmath 2011). For a comparison of WALS with TYTO, cf. Table 3.

WALS	TYTO
primarily based on invited expert information (chapters were written by experts); for some features examples available, but not data-driven as such; not collaborative system	data-driven; examples underpin the development of the knowledge base; collaborative system (allows anyone to contribute and revise information, similar to a Wiki)
in public query interface only two features combinable; searches possible with respect to language, region and features only; query output is restricted to overview data plus language specific examples in some instances	in public query interface any combination of variation dimensions combinable; different output (e.g. number of languages vs. list of languages that have a feature vs. list of examples given for a feature, amongst others) and different output formats supported (such as Word, PDF, html)
features, languages, and datapoint information downloadable in matrix format; download option does not include, e.g., example information and references accessible in online interface	everything will be made available for download; this includes the software as well as the different knowledge base parts (linguistic data, linguistic description etc.)
links cross-linguistic information to a world map ('world atlas') and hence gives good overview of distribution of features across the world	no link to actual map given, only geographic regions listed for languages within the knowledge base

Table 3: Comparison of WALS and TYTO

WALS encompasses an impressive amount of cross-linguistic information, while TYTO is currently still in its infant stages and will have to deal with the start-up problem as indicated above. While WALS provides comprehensive information with respect to their features using a map (and hence well illustrates the distribution of features across the world), TYTO allows for querying that is a lot more flexible and can, once up and running, generate relevant data for a large number of specialised research queries.

The last related project to discuss is the Typological Database System (TDS):

> The Typological Database System (henceforth TDS) is a web-based service that provides integrated access to a collection of independently created typological databases. Thus it is not an original data collection, but an interface to the data contained in its component databases. (Dimitriadis *et al.* 2009: 155)

TDS is hence different from the other projects discussed so far and from TYTO in that its aim is to make existing knowledge available and cross-link typological data for unified querying. It is, however, from a technical point of view, the related project that is closest to TYTO. For a comparison of TDS with TYTO, cf. Table 4.

TDS	TYTO
not knowledge base itself but interface to different collections	original knowledge base with re-usable components
uses Semantic Web technologies	uses Semantic Web technologies
data storage through XML; no relational database as back end	data storage through XML; no relational database as back end
unified querying supported with the help of an integrated ontology	unified querying centre-piece, based on underlying ontological structure
bottom-up ontology development	bottom-up ontology development
does not try to resolve conflicting typological analyses and terminology; conflicting information is included in the global ontology and unified under broader categories	tries to resolve conflicting typological analyses and terminology, although some leeway is given through the application of different strategies as indicated above

Table 4: Comparison of TDS and TYTO

> The component databases of the TDS ‘add up to some 1200 different descriptive properties, about more than 1000 languages. (Because of the heterogeneous nature of the collection, most properties are only filled for a fraction of the languages). Most of the data is in the from [sic] of high-level ‘analytical’ properties, but there are also a few collections of example sentences (with glosses) illustrating particular phenomena.’ (Project website, http://languagelink.let.uu.nl/tds/index.html)

4. Problems and outlook

As Zaefferer (2006: 113) noted in our introductory quote, the project of a general comparative grammar based on the firm ground of properly established leading ideas is still far from completion. To date, linguistic terminology captures a wide variety of ideas, yet no consensus has been reached as to which of these constitute the leading ideas in the field. In fact, one often finds confusion and disagreement within linguistic terminology (Nickles *et al.* 2007). One primary debate centres around the question of whether language-specific descriptive categories can be generalised for typological purposes. Haspelmath (2010) argues instead for a separation of ‘comparative concepts’ for typological purposes and language-specific ‘descriptive categories’ for the description of a particular language. We take a similar approach in that we make use of a layer of abstraction that is inherent in the TYTO system, in form of the type-instance distinction. Linguistic concepts (Haspelmath’s comparative concepts) will be defined on the type level within an ontological hierarchy, while the concrete language-specific realisations of a given type (Haspelmath’s descriptive categories) will be captured with their subtle differences as instances of this type.

A much-discussed example in the literature is the term ‘evidentiality’ (cf., e.g., Behrens in press, Boye and Harder 2009, Brugman and Macauley 2010 for discussions on the different interpretations of the term ‘evidentiality’ in the literature). While the ontology (and hence the type level) will contain conceptual categories such as HEARSAY, the specifics of hearsay categories in different languages will be recorded on the instance level, with language-specific instances linking to the HEARSAY type and other ontological types as required (e.g. some hearsay evidentials may be linked to concepts of grammaticalisation, whereas others may not, depending on whether the linguistic expression coding the concept of HEARSAY is grammaticalised or not).

An issue in the area of standardisation is related to how we combine the knowledge base, linguistic example information, and bibliographic reference data. Currently, our approach is to embed the linguistic example information and the bibliographic reference data in an opaque way (cf. §2.2) into the knowledge base. Yet, this is a case of a 'Content Information-specific software':

> Software is needed for efficient access to Digital Content Information. However, maintaining Content Information-specific software over the Long Term has not yet been proven cost effective due to the narrow application of such software. The danger of information loss is great when such software is relied upon for information preservation and understanding because it may cease to function under only small changes to the hardware and software environment. (CCSDS 2002: 3-4)

While we took measures to mitigate the problem by making that particular part of the software re-usable for other projects, we are still going to investigate more standard-conforming ways with a view to eliminate the problem altogether. At present, we are looking at implementing the Text Encoding Initiative's guidelines (TEI Consortium 2011), which would result in one TEI-conformant document not requiring 'Content Information-specific software' anymore.

Future areas of development include the potential integration with tools that already allow for linguistic data entry such as ELAN (ELAN n.d.; Wittenburg *et al.* 2006) and Toolbox (SIL International 2011). While the current project focuses on the domain of social cognition, and hence is being driven from an onomasiological perspective, TYTO is versatile enough to be applied to any other such cognitive domain. What is more, TYTO can also cater for a semasiological approach in that linguistic structures across languages can be the main focus of investigation. It is this huge research potential and its orientation towards its users that will hopefully make TYTO a very valuable and useful typology tool.

References

Behrens, Leila. in press (to appear 2012). Evidentiality, modality, focus and other puzzles: Some reflections on metadiscourse and typology. In Andrea C. Schalley (ed.), *Practical Theories and Empirical Practice.* Amsterdam/New York: John Benjamins.

Bickel, Balthasar, Bernard Comrie and Martin Haspelmath. 2008. *Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-Morpheme Glosses.* [http://www.eva.mpg.de/lin-gua/resources/glossing-rules.php]. Accessed on 28/10/2011.

Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79: 557-582.

BMIR (Stanford Center for Biomedical Informatics Research). 2011. *The Protégé Ontology Editor and Knowledge Acquisition System.* [http://protege.stanford.edu/]. Accessed on 28/10/2011.

Bow, Cathy, Baden Hughes and Steven Bird. 2003. *Towards a General Model for Interlinear Text.* [http://emeld.org/workshop/2003/bowbadenbird-paper.pdf]. Accessed on 28/10/2011.

Boye, Kaspar and Peter Harder. 2009. Evidentiality: Linguistic categories and grammaticalization. *Functions of Language* 16.1: 9-43.

Bray, Tim, Jean Paoli, C. Michael Sperberg-McQueen, Eve Maler and François Yergeau. 2008. *Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation, 26 November 2008.* [http://www.w3.org/TR/2008/REC-xml-20081126/]. Accessed on 28/10/2011.

Brugman, Claudia and Monica Macaulay. 2010. *Characterizing evidentiality*. Poster presented at the LSA annual meeting in Baltimore, MD, January 9.

Butler, Lynnika and Heather van Volkinburg. 2007. Review of Fieldworks Language Explorer (FLEx). *Language Documentation & Conservation* 1.1: 100–106.

CCSDS (Consultative Committee for Space Data Systems) (eds.). 2002. *CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Standards.* [http://public.ccsds.org/publications/archive/650x0b1.PDF]. Accessed on 2011-10-28.

CNRI (Corporation for National Research Initiatives). 2011. *The Handle System.* [http://www.handle.net/]. Accessed on 2011-10-28.

Comrie, Bernard and Norval Smith. 1977. Lingua Descriptive Studies: questionnaire. *Lingua* 42: 1-72.

Comrie, Bernard, William Croft, Christian Lehmann and Dietmar Zaefferer. 1993. A framework for descriptive grammars. In André Crochetière, Jean-Claude Boulanger and Conrad Ouellon (eds.), *Actes du XVe Congrès International des Linguistes/Proceedings of the XVth International Congress of Linguists.* Sainte-Foy: Les Presses de l'Université Laval. 159-170.

Crystal, David. 1997. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Dimitriadis, Alexis, Menzo Windhouwer, Adam Saulwick, Rob Goedemans and Tamás Bíró. 2009. How to integrate databases without starting a typology war: the Typological Database System. In Martin Everaert, Simon Musgrave, and Alexis Dimitriadis (eds.), *The Use of Databases in Cross-Linguistic Studies*. Berlin: Mouton de Gruyter, 155-207.

Dryer, Matthew S., and Martin Haspelmath. (eds.) 2011. *The World Atlas of Language Structures Online.* Munich: Max Planck Digital Library. [http://wals.info/]. Accessed on 2011-10-28.

ELAN. n.d. *Eudico Language Annotator. Language Archiving Technology (ELAN).* [http://www.lat-mpi.eu/tools/elan/]. Accessed on 2011-10-28.

Enfield, Nick J. and Stephen C. Levinson. (eds.) 2006. *Roots of Human Sociality: Culture, Cognition and Interaction.* Oxford: Berg.

EOPAS. n.d. *ETHNOER Online Presentation and Annotation System (EOPAS)*. [http://app.eopas.org/]. Accessed on 2011-10-28.

Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7.3: 97-100.

Galoes. n.d. [http://www.galoes.org/]. Accessed on 2011-10-28.

GOLD *(Generalized Ontology for Linguistic Description)*. 2010. [http://www.linguistics-ontology.org/gold.html]. Accessed on 2011-10-28.

Goody, Esther N. (ed.) 1995. *Social Intelligence and Interaction*. Cambridge: Cambridge University Press.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86.3: 663-687.

Horrocks, Ian, Markus Krötzsch, Michael K. Smith and Birte Glimm. (eds.) 2009. *OWL 2 Web Ontology Language Conformance. W3C Recommendation, 27 October 2009.* [http://www.w3.org/TR/2009/REC-owl2-conformance-20091027/]. Accessed on 2011-10-28.

IDF (The International DOI Foundation). 2011. *The Digital Object Identifier (DOI®) System.* [http://www.doi.org/]. Accessed on 2011-10-28.

JasperForge. 2000-2010a. *iReport: The Report Designer for JasperReports and JasperServer.* [http://jasperforge.org/projects/ireport]. Accessed on 2011-10-28.

JasperForge. 2000-2010b. *JasperReports: Open Source Java Reporting Library.* [http://jasperforge.org/projects/jasperreports]. Accessed on 2011-10-28.

Klyne, Graham and Jeremy J. Carroll. (eds.) 2004. *Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 10 February 2004.* [http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/]. Accessed on 2011-10-28.

Loos, Eugene E., Susan Anderson, Dwight H., Day, Paul C. Jordan and J. Douglas Wingate. (eds.) 2004. *Glossary of Linguistic Terms.* SIL International. [http://www.sil.org/ linguistics/GlossaryOflinguisticTerms/contents.htm]. Accessed on 2011-10-28.

Nickles, Matthias, Adam Pease, Andrea C. Schalley and Dietmar Zaefferer. 2007. Ontologies across disciplines. In Andrea C. Schalley and Dietmar Zaefferer (eds.), *Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts.* Berlin/New York: Mouton de Gruyter. 23-67.

Nordhoff, Sebastian. 2008. Electronic reference grammars for typology: Challenges and solutions. *Language Documentation & Conservation* 2.2: 296–324

Oliver, Pamela, Gerald Marwell and Ruy Teixeira. 1985. A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *The American Journal of Sociology:* 91.3: 522-556.

Prasarnphanich, Pattarawan and Christian Wagner. 2008. Creating critical mass in collaboration systems: Insights from Wikipedia. In *Proceedings of the Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)*, 126-130.

Prud'hommeaux, Eric and Andy Seaborne. (eds.) 2008. *SPARQL Query Language for RDF. W3C Recommendation 15 January 2008.* [http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/. Accessed on 2011-10-28.

Schalley, Andrea C. 2011. Semasiology 'versus' onomasiology? Paper presented at *ALS 2011 (Annual Conference of the Australian Linguistic Society)*, Australian National University, Canberra, December.

———— in press (to appear 2012). Many languages, one knowledge base: Introducing a collaborative ontolinguistic research tool. In Andrea C. Schalley (ed.), *Practical Theories and Empirical Practice.* Amsterdam/New York: John Benjamins.

Schalley, Andrea C. and Dietmar Zaefferer. 2007. Ontolinguistics—An outline. In Andrea C. Schalley and Dietmar Zaefferer (eds.), *Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts.* Berlin/New York: Mouton de Gruyter. 3-22.

Schroeter, Ronald and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In Linda Barwick and Nicholas Thieberger (eds.), *Sustainable Data from Digital Fieldwork.* Sydney: Sydney University Press. 99-124.

SIL International. 2010. *Fieldworks Language Explorer (FLEx)*. Available online at [http://fieldworks.sil.org/flex/]. Accessed on 2011-10-28.

———— 2011. *Field Linguist's Toolbox.* [http://www.sil.org/computing/ toolbox/]. Accessed on 2011-10-28.

TEI Consortium. (eds.) 2011. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. 1.9.1. 5 March 2011.* [http://www.tei-c.org/Guidelines/P5/]. Accessed on 2011-10-28.

TypeCraft. n.d. [http://typecraft.org/]. Accessed on 2011-10-28.

Windhouwer, Menzo and Alexis Dimitriadis. 2008. Sustainable operability: Keeping complex resources alive. In *Proceedings of the LREC workshop on Sustainability of Language Resources and Tools for Natural Language Processing (SustainableNLP08)*, 9-18.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alexander Klassmann and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Zaefferer, Dietmar. 2006. Realizing Humboldt's dream: Cross-linguistic grammatography as data-base creation. In Felix Ameka, Alan Dench and Nicholas Evans (eds.), *Catching Language: The Standing Challenge of Grammar-Writing.* Berlin: Mouton de Gruyter. 113-136.