TYTO – A Collaborative Research Tool for Linked Linguistic Data

Andrea C. Schalley

Abstract In this paper, I introduce a computational tool, TYTO ("Typology Tool"), that utilises Semantic Web technologies in order to provide novel ways to process, integrate, and query cross-linguistic data. Its data store incorporates a set of ontologies (comprising linguistic examples, annotations, language background information, and metadata) backed by a logic reasoner software. This allows for highly targeted querying, and, with enough data on the relevant interest areas, TYTO can return answers to rather specific typological questions such as 'Which other languages in the North America, in addition to Yuchi, do encode senior kin and in-group (such as belonging to the same ethnic group) in a suffixal case marking system?' TYTO's data store can be extended with additional ontologies and adapted to allow for project-specific analyses of linguistic data. It is further designed to facilitate collaboration and allow multi-user contributions, including automatic integration of data submitted at different stages by different contributors.

1 Introduction

What would, in an ideal world, an electronic resource supporting the division of typological theories in linguistics look like? The wish-list of typologists would presumably include the following:

1. *Cross-linguistic data*
   The resource should comprise data from all of the languages of the world, and this data should be comprehensive in order to allow for the cross-linguistic comparison of all form and meaning based differences found in the languages of the world. The data set should include both the raw data and its annotation or analysis by experts of the corresponding languages.

Andrea C. Schalley
Griffith University, Brisbane, Australia, e-mail: a.schalley@griffith.edu.au

2. *Grounding in actual linguistic examples*
   It should provide access to actual linguistic examples for all analyses in order for theoretical claims to be verifiable. I.e. all data analyses contained in the resource should be documented and traceable back to example data, and via metadata information to the source of the example data.

3. *Data analysis*
   Language data should be open for reanalysis, in order to correct and expand on previous analyses, and keep the overall system responsive to scientific progress. Such reanalyses should be tracked in the system, i.e. a history of sequential analyses of a data item should be accessible. Also, the data should be analysed very fine-grainedly, capturing information on all possible dimensions of typological variation.

4. *Querying and reporting*
   The resource should be able to inform researchers in responding to highly targeted research questions such as 'Which other languages in the North America, in addition to Yuchi, do encode senior kin and in-group (such as belonging to the same ethnic or family group) in a suffixal case marking system?' Even more so, it should provide for extensive flexibility in accessing the data and their analyses. This in particular includes that a user-chosen number of variation dimensions should be dynamically combinable in user-defined queries. Such queries also determine in what way elements of the knowledge base should be presented, i.e. whether for instance just the number of languages that meet the query, a list of those languages, or a list of all examples for all of the languages should be reported. It should be self-explanatory to anyone how to formulate queries, and the user should be able to determine in which format the results of the query are presented.

5. *Scope*
   In line with the aim of being able to compare linguistic form and meaning, the language data in the knowledge base should have been consistently analysed both with regard to their form and their meaning components. Going back to the above example, for a given data item the knowledge base should comprise whether this data item constitutes a suffix and/or a case marker, and whether it encodes the concepts KIN SENIOR and/or ETHNIC SAME. In other words, both a semasiological view and an onomasiological view on the language data should be facilitated by the knowledge base.

6. *Multi-user contributions*
   The resource should allow multiple contributors and multiple 'consumers' (querying the knowledge base) to collaborate through it and use it at the same time. It should be able to handle overlapping contributions on the same and different areas of the knowledge base at the same and different times. Newly committed data should be automatically integrated into the knowledge base, and data should be made accessible for querying immediately after their submission.

7. *Fieldwork compatibility*
   The resource should be deployable in the field, i.e. (i) it should be possible to store the whole resource locally on a computer and to run the system independently, without access to the Internet, (ii) researchers should be able to enter data collected in the field while in the field, and (iii) they should be able to query the system (including their own newly entered data) and obtain reports on the basis of the data while in the field, in order to inform their hypothesis forming and theory building processes. On their return, researchers should be able to easily and fast feed back their new data into the overall data store, where it should be integrated automatically.
8. *Data entry*
   The entry of language data and their analyses should be organised as user-friendly as possible. It should meet the needs of typologists (e.g. in terms of automatically parsed interlinear glossing), be fast and efficient. The data entry should be flexible enough to provide interfaces for the submission of non-anticipated data be be entered.
9. *Expandability*
   The knowledge base should be expandable. In particular, it should be possible to add new analytical concepts into a structured interconnected 'vocabulary' of descriptive and semantic concepts. Terminological controversies should be catered for, by either resolving them successfully or by integrating all options into the 'vocabulary'. Also, the knowledge base should ideally be able to hold information on both what is possible and what is not possible in a particular language or in the languages across the world more generally (i.e. it should contain positive and negative evidence).

Of course, we do not live in an ideal world, and there are obvious roadblocks to implementing the wish-list fully, and the most obvious ones amongst them are: (i) We do not have enough information gathered about all languages in the world, and hence comprehensiveness cannot be achieved. (ii) Related to this, for an analysis of each language with regard to all typological variation dimensions, we would have to have a complete list of those variation dimensions. This is not the case to date, and much debate is ensuing about those dimensions and the question of language universals more generally (cf., e.g. the recent discussion about the 'myth of language universals', Evans and Levinson, 2009; Rooryck et al, 2010). (iii) Even if (i) and (ii) were solvable, the amount of expert analysis that would have to go into the knowledge base would be immeasurable, and a never-ending task if one takes language change into account. (iv) Terminological controversies are unlikely to be successfully resolved, and it has to be acknowledged that there is a terminological tension between the description of a single language and the comparison of several or all languages (cf., e.g., the discussion in Corbett, 2007; Haspelmath, 2010).

Despite thus entertaining an unrealistic vision, we can try to come as closely as possible to this ideal – with the current computational resources we have available. In this paper, I introduce a typology tool, TYTO, that has been designed to implement as many of these desiderata as possible. While not fully completed to date, I

will give an overview of what the tool can and cannot do as well as outline some of the design decisions and technologies it is based upon in Sect. 3. Before going into detail about TYTO, I will very briefly look at predecessing and related current systems in Sect. 2.[1] I will conclude with an outlook on the further development of TYTO in Sect. 4.

2 Related Projects

One of the first projects that aimed at creating an online resource for typological information was the *Cross-linguistic Reference Grammar (CRG)*, which was initiated jointly by Bernard Comrie, William Croft, Christian Lehmann, and Dietmar Zaefferer about two decades ago (Comrie et al, 1993; Zaefferer, 2006). It aimed "to create some kind of revised electronic version of the famous Lingua descriptive studies [LDS, AS] questionnaire" (Zaefferer, 2006, 113), as published in Comrie and Smith (1977), and hence at integrated comprehensive grammatical descriptions of languages. CRG's knowledge base consists of a predefined AND-OR-tree.

In terms of current systems, one of the prominent ones surely is *The World Atlas of Language Structures (WALS)*, which "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). WALS, in its current second version, holds an impressive amount of cross-linguistic information. It comprises 192 features such as 'Voicing in Plosives and Fricatives', 'Sex-based and Non-sex-based Gender Systems', 'Position of Case Affixes', 'SOVNeg Order', or 'Tea', and hence contains both semasiologically-oriented as well as onomasiologically-oriented features. Contributions are submitted by experts on the features (and not by experts on the languages). It allows to cross-search for two properties from its public interface.

The *Database of Syntactic Structures of the World's Languages (SSWL)* also allows a cross-search for two properties from its public interface. However, in contrast to WALS, contributors are not experts on the properties but experts but on the respective language on which they contribute. "SSWL is a searchable database that allows users to discover which properties (morphological, syntactic, and semantic) characterize a language, as well as how these properties relate across languages. This system is designed to be free to the public and open-ended. Anyone can use the database to perform queries." (SSWL, n.d.) It encompasses data on 130 languages and 54 binary properties, and has received language analyses from 202 contributors to date (SSWL, n.d., as of 27 Nov 2011).

While WALS and SSWL are typological databases in themselves, the from a technical viewpoint closest system to TYTO is the *Typological Database System (TDS)*, which "is a web-based service that provides integrated access to a collection

[1] For a still brief, but more explicit comparison of TYTO with some of the approaches listed in Sect. 2 (also regarding some technical features), cf. Borkowski and Schalley (2011).

of independently created typological databases." (Dimitriadis et al, 2009, 155) It thus provides an interface to the data contained in other typological databases. It supports unified querying across these typological resources with the help of an integrated ontology (network of cross-connected concepts relevant for the domain), and uses a bottom-up approach to the development of this ontology.

The last related resource to be listed here is the grammar-authoring tool *Galoes* (Galoes, n.d.; Nordhoff, 2008). Currently, there is one full grammatical description available, and about another six partial descriptions exist (although some do not allow read access yet). "GALOES supports rendering of linguistic examples, embedded audio files, cross-references, collaboration tools and more. There is an online version on www.galoes.org/grammars and an offline client to be used on your laptop." (Galoes, n.d.) The database can be searched, either by choosing the free search function or pre-defined question from, e.g., the LDS questionnaire or WALS.

3 TYTO

Yet, typologists will continue to encounter specific questions for which answers or at least leads to answers cannot be readily obtained from the resources outlined above – the array of conceivable questions seems limitless. The resource introduced in the following, TYTO,[2] is intended to provide more flexibility in the kind of questions that can be posed to the system, and this of course has implications on how the underlying knowledge base is set up, and what technologies are used to achieve this responsiveness to the needs of users.

In the following, I will take up the wish-list from Sect. 1, addressing in each case how TYTO fares with regard to these ideals. Given the space limitations in this chapter, I will neither be able to provide very detailed accounts of the plethora of topics touched upon in the wish-list nor explain TYTO's technical solutions in detail. This chapter is merely intended as an introduction to the TYTO tool and its underlying ideas. For additional information on some of the approaches taken and design decisions made, cf. Borkowski and Schalley (2011) and Schalley (in press).

3.1 Cross-Linguistic Data

As indicated above, this ideal will remain wishful thinking, given the scattered knowledge we have about the languages of the world, the incomplete list of vari-

[2] TYTO is being developed as part of the Australian Research Council Discovery project DP0878126 ("Social cognition and language"). The aims of this project include the building of a detailed and cross-linguistically valid model of how social cognition is grammaticalized across the world's languages. This is approached through the systematization and synthesis of already-recorded material ('library sample') and newly-gathered data for a small number of languages ('fieldwork data'). TYTO is the tool that will eventually provide the infrastructure to move this model into an electronic, dynamically queryable, and extensible format.

ation dimensions across languages, and our limited resources to describe more languages – or already described languages in more detail (i.e. gather more language data, and analyze it fine-grainedly). Nonetheless, whatever is collected can in principle be integrated into TYTO. However, to attain the aim of a high comparability of cross-linguistic data and hence flexible searchability across the underlying knowledge base, the information has to be semantically tagged in a consistent way. This is achieved using Semantic Web concepts. An underlying ontology forms the backbone of TYTO's knowledge base; it consists of a set of interrelated sub-ontologies, including: (a) the primary semantic domain (in our project: social cognition); (b) linguistic examples (both in their original form and potentially revisions of this original); (c) linguistic annotations (cross-linguistic description with respect to both form and function); (d) language background information (family, size, vitality, geographic region [linguistic and political], society [economy, religion, tradition, etc.]); and (e) metadata (example source information [fieldwork, literature]; general metadata [contributor information etc.]).

The ontology is being developed using the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2004) and the ontology editor Protégé (BMIR, 2011). Due to the modular organization of the overall ontology into these sub-ontologies, reusability of each of the overall ontology's parts is ensured and encouraged. It is envisaged that some of the sub-ontologies will in the future be linked to corresponding resources such as Ethnologue (Lewis, 2009).

3.2 Grounding in Actual Linguistic Examples

TYTO's development is data-driven. The ontology and hence the knowledge base will be incrementally built up, through the analyses that are entered for linguistic examples. The linguistic examples themselves will also be stored in the knowledge base, "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." (Borkowski and Schalley, 2011) All data analyses are thus verifiable and can be traced back to specific linguistic examples. We believe that only a data-driven approach will generate the fine-grainedness of ontological concepts, relations, and constraints needed for typological work. This approach stands in sharp contrast to the approach that has been adopted for the development of GOLD, the *Generalized Ontology for Linguistic Description* (GOLD, 2010).[3]

3.3 Data Analysis

TYTO offers the option to revise example analyses and keeps a history of these revisions in the running system, i.e. this history can be directly accessed through

[3] For critical comments on GOLD, cf. Cysouw et al (2005) and Munro and Nathan (2005).

queries. In principle, it is possible to analyse each linguistic example with regard to all possible variation dimensions (i.e. those that are relevant to it). This is unlikely to happen in any one analysis step, so allowing for revisions is a way of also allowing the addition of analyses with regard to additional variation dimensions (the same linguistic example can hence function as the evidence for several or all variation dimensions).

3.4 Querying and Reporting

Users are able to pose targeted questions as required by their research foci and needs, via tailored queries. The system is based on a logic reasoner software and Semantic Web technology (SPARQL, Prud'hommeaux and Seaborne, 2008) and hence can be queried in a flexible way that allows for combining a chosen number of variation dimensions comprised in the ontology within a query. Provided with the right query, TYTO can answer highly targeted questions such as our example from Sect. 1. What and how information is presented is specified through the reporting engine (JasperReports, JasperForge, 2000-2010), which also allows for a wide range of output formats (including, e.g., PDF). The only current drawback is that the user needs to somewhat familiarise themselves with the structure of the ontology, the query language, and the report designer. However, designed reports can be shared and reused easily in the user community.

3.5 Scope

As indicated previously, an analysis both with regard to form and meaning components is possible using TYTO, i.e. both a semasiological view and an onomasiological view on the language is facilitated (and made explicit in the corresponding substructuring of the ontology). It is extremely unlikely that analysis comprehensiveness will be achievable (given knowledge and resource limitations), but this is catered for by the open-world assumption underlying TYTO's ontological approach: if the reasoner cannot prove a statement to be true, such a statement being considered as unknown, and the system can deal with this lack of knowledge.[4] The open-world assumption thus appropriately reflects the fact that neither single contributors or users of the knowledge base nor the overall knowledge base will ever have complete knowledge.

[4] Even though this might sound very natural, it is in contrast to most computational approaches that use the closed-world assumption and hence assume that their knowledge about the application domain is complete.

3.6 Multi-User Contributions

The tool is specifically designed as to support collaborative effort in linguistic typology, offering the potential of flexible contributions spanning across different times, locations, and groups of contributors and 'consumers'. This is achieved through an elaborate version control system, which will run automatic consistency checks on the ontology using the reasoner, for automatic integration of data into the knowledge base (or human postprocessing in the case of conflicts). For a more detailed description of this process, including how TYTO deals with conflicting analyses by contributors and how it is envisaged to persuade a significant number of potential contributors to participate (via creating a critical mass of data at the outset), cf. Borkowski and Schalley (2011) and Schalley (in press).

3.7 Fieldwork Compatibility

The software is designed to be installed locally on users' computers. "This permits use while on fieldwork – disconnected from the Internet – [...], so interested parties can use the complete TYTO system independently." (Borkowski and Schalley, 2011) Fieldwork results can then be easily fed back into the central TYTO knowledge base; this will be handled by the version control system, as this is a special case of employing the multi-contributor capabilities.

3.8 Data Entry

An initial input system has been developed which takes interlinear glossed data, parses it on the basis of the Leipzig Glossing Rules (Bickel et al, 2008) (with some additions as necessary for computational processing). Each component is then available for selection and can be linked to categories in the ontology, e.g. a suffix indicating that the denoted entity is a member of the ethnic in-group can be linked to the ontology category ETHNIC SAME. In addition, form fields are available, for entering metadata and other relevant information, such as the language (e.g. Yuchi), for which the data is an example. These form fields also provide an option for entering non-anticipated data. For a more detailed description of the workflow, cf. Borkowski and Schalley (2011) and Schalley (in press).

3.9 Expandability

TYTO's ontological approach, with its open-world assumption, comes with the advantages of ontologies: ease of knowledge base extension and maintainability. It is possible to integrate new concepts into the ontology (the knowledge base's struc-

tured interconnected 'vocabulary' of descriptive and semantic concepts), change the ontological class hierarchy (including multiple inheritance), add additional linguistic examples, and provide further analyses to already existing linguistic examples. TYTO, given its data-driven nature, can, however, only provide positive evidence at this stage in its development, as an analysis of linguistic examples that are well-formed can only provide information on what is possible.

4 Conclusion

This chapter could only provide a rather cursory introduction to the collaborative typology tool TYTO. TYTO will remain a major enterprise requiring further concerted development to bring all its features to full fruition. The design of the tool is very modular, it consists of a number of separable components that other projects can select and adjust to their needs by mixing and matching as required. The incremental development of the ontology will continue for years to come, and roadblocks are expected in particular for the ontology development due to terminological controversies and some inadequate descriptive devices for a number of linguistic phenomena. Nonetheless, it is worthwhile pursuing this path, as it will bring issues out into the open as well as collate information. We are planning on linking this information to resources holding related information about, e.g., language, culture, and geography, as well as making our own data available in linkable form.

Acknowledgements I gratefully acknowledge the support I received from the Australian Research Council (Grant *Social Cognition and Language*, DP0878126). In addition, I would in particular like to thank Alexander Borkowski and Nicholas Evans for their collaboration on this project. Without our stimulating discussions, their insight, and their invaluable contributions on the conceptualization, design, and implementation of TYTO, the tool would not be where it is now.

References

Bickel B, Comrie B, Haspelmath M (2008) Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Available online at http://www.eva.mpg.de/lingua/resources/glossing-rules.php. Accessed on 2011-11-27

BMIR (2011) The Protégé Ontology Editor and Knowledge Acquisition System. Stanford Center for Biomedical Informatics Research. Available online at http://protege.stanford.edu/. Accessed on 2011-11-27

Borkowski A, Schalley A (2011) Going beyond archiving - a collaborative tool for typological research. In: Thieberger N, Barwick L, Billington R, Vaughan J (eds) Sustainable data from digital research: Humanities perspectives on digital scholarship, Custom Book Centre, University of Melbourne, Melbourne

Bow C, Huges B, Bird S (2003) Towards a general model of interlinear text. In: Proceedings of EMELD 2003, available online at http://emeld.org/workshop/2003/bowbadenbird-paper.pdf. Accessed on 2011-11-27

Comrie B, Smith N (1977) Lingua descriptive studies: questionnaire. Lingua 42:1–72

Comrie B, Croft W, Lehmann C, Zaefferer D (1993) A framework for descriptive grammars. In: Crochetière A, Boulanger JC, Ouellon C (eds) Actes du XVe Congrès International des Linguistes/Proceedings of the XVth International Congress of Linguists, Les Presses de l'Université Laval, Sainte-Foy, pp 159–170

Corbett G (2007) Canonical typology, suppletion, and possible words. Language 83(1):8–42

Cysouw M, Good J, Albu M, Bibiko HJ (2005) Can gold 'cope' with wals? retrofitting an ontology onto the world atlas of languages structures. In: Proceedings of EMELD 2005, available online at http://emeld.org/workshop/2005/proceeding.html. Accessed on 2011-11-27

Dimitriadis A, Windhouwer M, Saulwick A, Goedemans R, Bíró T (2009) How to integrate databases without starting a typology war: the typological database system. In: Everaert M, Musgrave S, Dimitriadis A (eds) The Use of Databases in Cross-Linguistic Studies, Mouton de Gruyter, Berlin, pp 155–207

Dryer M, Haspelmath M (eds) (2011) The World Atlas of Language Structures Online. Max Planck Digital Library, Munich, available online at http://wals.info/. Accessed on 2011-11-27.

Evans N, Levinson S (2009) The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences 32:429–492

Galoes (n.d.) Available online at http://www.galoes.org/. Accessed on 2011-11-27.

GOLD (2010) Generalised Ontology for Linguistic Description. Available online at http://www.linguistics-ontology.org/gold.html. Accessed on 2011-11-27.

Haspelmath M (2010) Comparative concepts and descriptive categories in crosslinguistic studies. Language 86(3):663–687

JasperForge (2000-2010) Jasperreports: Open Source Java Reporting Library. Available online at http://jasperforge.org/projects/jasperreports. Accessed on 2011-11-27.

Lewis M (ed) (2009) Ethnologue: Languages of the World, Sixteenth edition. SIL International, Dallas, online version available at http://www.ethnologue.com/. Accessed on 2011-11-27.

McGuinness D, van Harmelen F (2004) Owl web ontology language. overview. W3C Recommendation 10 February, available online at http://www.w3.org/TR/owl-features/. Accessed on 2011-11-27

Munro R, Nathan D (2005) Towards portability and interoperability for linguistic annotation and language-specific ontologies. In: Proceedings of EMELD 2005, available online at http://emeld.org/workshop/2005/proceeding.html. Accessed on 2011-11-27

Nordhoff S (2008) Electronic reference grammars for typology: challenges & solutions. Language Documentation and Conservation 2(2):296–324

Prud'hommeaux E, Seaborne A (2008) SPARQL Query Language for RDF. W3C Recommendation 15 January, available online at http://www.w3.org/TR/rdf-sparql-query/. Accessed on 2011-11-27.

Rooryck J, Smith N, Liptak A, Blakemore editors D (2010) Special issue on Evans & Levinson's "The myth of language universals". Lingua 120(12):2651–2758

Schalley A (in press) Many languages, one knowledge base: Introducing a collaborative ontolinguistic research tool. In: Schalley A (ed) Practical Theories and Empirical Practice, John Benjamins, Amsterdam/Philadelphia

SSWL (n.d.) Database of Syntactic Structures of the World's Languages. Available online at http://sswl.railsplayground.net/. Accessed on 2011-11-27.

Zaefferer D (2006) Realizing Humboldt's dream: Cross-linguistic grammatography as data-base creation. In: Ameka F, Dench A, Evans N (eds) Catching Language: The Standing Challenge of Grammar-Writing, Mouton de Gruyter, Berlin, pp 113–136