A social layer for typological databases

Michael Cysouw

1. Social considerations in database design

Thinking about social considerations in the design of a database only becomes an issue when the database is considered to be more than just a personal tool for one particular researcher. In still the majority of cases, personal usage is the setting in which typological databases are constructed and used: one researcher (or a close-knit group of researchers in regular contact) is dealing with a large body of information that has to be structured, and the database is used as a tool to guide the organisation of the information. In this situation, many of the day-to-day decisions that are taken in the actual coding of characteristics for the database are not documented. This is not necessarily a problem, because the consistency of the coding is normally ensured by explicit and implicit assumptions of the researcher (or by the regular team discussions in case of a group of researchers).

However, the limits of this approach are reached when the resulting database is made publicly available (e.g. after the project has finished, or in futuristic ‘wiki’-like settings of typological databases with unmoderated cooperation). This is an important consideration, as it is to be expected that the publication and sharing of databases will in the future become more and more important. First, and foremost, many a researcher’s wish to see the long-term effort of building a database to have any impact on the scientific discussion already results currently in the online availability of various typological databases (cf. Baerman et al. 2002; Gast et al. 2007). Also, publicly available databases are beginning to become an accepted entry on a curriculum vitae, honouring the time and effort spent on the structured collection of information. And finally, and probably most pressingly, various funding agencies have started to ask for database publication when the funding finishes. If these trends continue, more and more databases will be made available for other researchers to use. In this situation, the future usage of the databases in the social realm of scientific practice should be a consideration in the design of the database.

In: Andrea Sansò (ed.), *Language Resources and Linguistic Theory*, 59-66. [Materiali Linguistici 59.] Milano: Franco Angeli, 2007.

2. Interoperability problems

There are various problems that have to be addressed once a database is made public. The most obvious problem is that other researchers will have to understand what kind of information is to be found in the database. This is normally dealt with by describing in general terms what the different codings in the database are supposed to represent. To take one prominent publicly available typological database, the *World Atlas of Language Structures* (henceforth WALS, Haspelmath et al. 2005) has a short text chapter with every typology defining and explaining the different possibilities that are distinguished in each typology displayed (every typology is presented in the form of a world map). In these text chapters, often a few problematic cases are discussed, but most of the specific assignments of characteristics to the languages is left unexplained. A user will simply have to trust the integrity of the author. For most of these assignments a literature reference is indicated in the *Interactive Reference Tool* accompanying the WALS, and in many (though far from all) there is even a page number accompanying the reference. However, often the user is still left wondering what the author had in mind when assigning a particular typological characteristic to a particular language.

This does not at all mean that the coding in typological databases is regularly erroneous. Far from that, I believe that most typologies currently available are consistent and largely accurate. Most apparent errors are caused by misunderstandings between the author of the typology and a user of the data. The author of a typology might implicitly have a particular exemplary language in mind as the prototype for a particular typology, without spelling out all the details of this implicit definition in the accompanying text. A user, reading only the explicit definitions, might have a slightly different prototypical case in mind, thereby unwittingly interpreting the definition differently from the author. Both these interpretations will normally have their value – it is not a question of right or wrong – but there is no way for the author and the user to actually spell out their differences of opinion and thereby resolving the issue. Unfortunately, in many situations occurring in today’s typological practice, such unresolved misunderstandings take on a negative dialectic. Typically, a specialist on a particular language taking a different stance on the details of the distinctions between the various typological possibilities will accuse the typologist of misrepresenting the language of his or her speciality. Vice versa, the typologist will accuse the specialist for narrow-mindedness, not understanding the wide range of diversity among the world’s languages that have to be classified, which forces the establishment of strict boundaries often not along the lines of traditional interpretations of linguistic phenomena. Instead of helping each other, much too often the language specialist and the domain specific typologists are working against each other.

On a different level there is also a big problem with the interoperability between different typological databases. Not only language specialist and typologist might have different opinions on the precise details of a cross-linguistic viable definition of a linguistic concept; also among typologists there is no consensus about the meaning of linguistic terminology. Mismatches between two typological databases cannot automatically be interpreted as one having errors, but will often have to be interpreted as different opinions on a particular aspect of linguistic structure. This implies that it will turn out to be impossible in most cases for different typological databases to be interoperable in a direct sense. Even if we were to link two databases together in a technical sense, it would not be possible to merge any two typologies from the two databases (which would actually be very valuable, for example, to extend the size of samples, or to cross-check validity).

For example, there is no consensus among linguists about what it means for a language to have case. In WALS, there are seven typologies related to the marking of case which do not even agree on the question whether a language has case or not. Just to spell out one example (but this exercise can be repeated for every cross-section of two case-related typologies by different authors in WALS): there are eight languages that are both classified as having a particular kind of alignment in their nominal case marking (and thus implying the presence of some kind of case, Comrie 2005 = WALS 98), and as lacking any form of either symmetrical or asymmetrical case marking (and thus implying the absence of any kind of case, Iggesen 2005a = WALS 50). Relative to the 187 languages shared by these two typologies, these eight contradictory languages represent more then four percent of languages coded (and for those interested, the eight languages in question are Drehu, Guaraní, Hebrew, Igbo, Khasi, Maori, Tukang Besi, and Urubú-Kaapor). Again, this mismatch between different typologies is probably not due to erroneous classifications, but exists because of different interpretations of what counts as case.

The exemplary inspection of Drehu, the first mismatch in the above example, illustrates the usefulness of an extra "social" layer of documentation in typological databases. Checking in the *Interactive Reference Tool* of WALS, it turns out that for Drehu both Iggesen (2005a) and Comrie (2005) refer to the same source, with even an overlap of the relevant page numbers. Iggesen says that Drehu does not have case with reference to Moyse-Faurie (1983: 76, 146-152). In contrast, Comrie classifies Drehu as having active/inactive case marking with reference to Moyse-Faurie (1983: 147). This illustrates nicely that adding a literature reference does not necessarily help to resolve conflicting characterisations. However, Comrie happens to have added example sentences with a note to this classification in the *Interactive Reference Tool*:

> "In the non-Past, illustrated here [in example (1), MC], the agent marker is used obligatorily with transitive subjects and optionally with agentive intransitive subjects; in the Past, it is used with all subjects (though with some exceptions for inanimates). Pronouns apparently follow the same patterns, though few possibilities are illustrated." (Comrie 2005, quoted from the *Interactive Reference Tool*)

(1) Drehu (Moyse-Faurie 1983: 147)

kola	*huliwa*	*hnei*	*wamo*
DUR	work	AG	Wamo

'Wamo is working.'

This note makes clear that Comrie considers the agentive particle *hnei* to be case marking. That might set one thinking, and indeed: rereading the text from Iggesen (2005b = WALS 49), he explicitly states that non-bound forms are excluded from his typology:

> "In the languages lacking morphological case (e.g. Vietnamese), grammatical relations are expressed by word order and/or *morphologically and prosodically independent function words* (in general, prepositions and postpositions), and partly also by morphological devices on the verb." (Iggesen 2005b: 202, italics added, MC)

This example illustrates that a short note explaining the reasons for a particular decision in the typological coding of a language is extremely helpful to understand the reasons for discrepancies between two typologies (or between a typology and the specialist's knowledge on a language). It is not even necessary that the crucial information has to be put in the note (and of course nobody will know in advance what the crucial information will be for a user of the database). In Comrie's note above, there is no mention of the morphological status of the marker (which was the crucial information pinning down the difference between the two typologies of case). He only made some comment on an "agent marker" and the example shows this agent marker as an independent word. A linguist reading such information will infer the difference in classification between Comrie and Iggesen from this information alone (but of course a computer will not easily grasp such implicit information).

3. Two layers in typological databases

The whole point of a social layer in typological databases is to add information to the database in a form that is intended to be read by linguists, and that is not necessarily be processable automatically. In many databases this is already being put to practice in the form of "note"-fields or "comments"-fields, as exemplified by Comrie's note above. However, contrary to current practice, these notes should not be seen as just a subsidiary piece of information, nor interpreted as a minor nuisance caused by detailed information on the structure of a particular language that is too fine-grained or too idiosyncratic to be included in the current structure of the database. In contrast, I propose that these notes should be considered as a semi-independent part of typological databases that allow for a proper discussion of the details of the structure of individual languages.

From this perspective, a typological database becomes a two-layered entity. The primary layer is a collection of "language-structure annotations". Such annotations always deal only with one language, and are related to a specific research question. Returning to the example about case marking, Comrie's (2005) research question was something like "What kind of case alignment does the language have with full noun phrases?". For the language Drehu he then provides the note discussed previously, consisting of a literature reference, example sentences, and a few lines with interpretation of these sources of information. Minimally, such a note would be a collection of relevant information to answer the question at hand (i.e. literature references with page numbers). However, ideally it also includes a few sentences about the interpretation of the author, clarifying what is considered to be the crux of the matter, and maybe even a few example sentences.

Only in a secondary layer of the database will the typological classification be fixed, using the controlled vocabulary as decided upon within the project. One immediate pay-off of a strict adherence to this division of labour will be that any later change in the controlled vocabulary (e.g. the addition of a structural type, or a complete reshuffling of the structural analysis) will be much easier to run through the languages already entered in the database, because the relevant information will be clearly documented. However, the real benefit of such a two-layered database structure only surfaces once someone else wants to make use of the data collected.

The primary layer, consisting of language-structure annotations, is a *social layer* meant for cooperation and discussion. Basically, the annotations are collections of relevant information for a particular question. So, the immediately obvious usage of this layer is that other people interested in the same (or a closely related) question can easily find out about relevant information and form their own opinion on that basis. However, users will also be able to comment on annotations by making a new annotation in their own database. For example, such annotations could be used for adding newly found relevant information, for pointing out mistakes, or for drawing different conclusions based on the same information.

This whole layer of language-structure annotations with typologically relevant information exists independently of the layer with actual typological decisions of the form “language X is of type T”. Such typological classifications will probably always remain biased by personal opinions of the individual author (or research group). Any such classification or parameter is therefore best considered a *personal layer* of typological databases, strictly speaking only relevant for the author of the parameter. It is of course possible for a user to accept the classification wholesale, and use it for whatever purpose one has in mind (e.g. crossing it with other parameters). However, when the details of the parameter itself is of interest to a user, then it might be better to sit down and read all the language-structure annotations provided by the database and make a new personalised classification on that basis. Only in this way will it be possible to control directly for the many decisions necessary, like which types to distinguish and which characteristics to consider relevant for each type.

4. A vision of the future

So what would the typological practice look like if databases consistently added a social layer? To sketch such a picture, I will assume that (electronic) database publishing in the (near) future will use RDF (*Resource Description Framework*, which is the format of the Semantic Web), or a format in like spirit. In such a format, each individual piece of information is assigned a unique identifier (in RDF parlance these identifiers are called URI, *Uniform Resource Identifier*, of which the well-known URLs are a subclass). Through this identifier each piece of information can be identified, accessed and referred to individually, and relations between each two pieces of information can be completely specified. In this framework, a database is simply a large collection of URIs with relations between these URIs (in a way, this is a relational database taken to its extreme). The real beauty of this approach is that now every individual piece of information can also be referred to separately by other people using the information. Thus each individual language-structure annotation will be available as a separate entity, a sort of micro-publication, linked both to the typological research question to which it is related (and thus indirectly to the original database) and to the language it is about.

The first implication of this scenario is that language-structure annotations have to be readable in isolation. Every reference to facts described elsewhere in another note has to be made explicit. This might in the short run result in slightly more work, but this will pay off in the long run, as the information collected is much better reusable.

Another important implication of publishing databases is that a clear distinction is necessary between electronically "published" and "unpublished" information. At some point, the author of a publicly available database should declare the database, or a part of the database, as "published". After that point the published annotations should be unchangeable. This is crucial to allow other people to use the database and refer to it. If the database is still in constant flux, than any conclusions taken on the basis of that database by somebody else can easily become null and void. If the original author later changes his or her mind, or finds out about new relevant information, this can simply be added as a new annotation referring to the original annotation.

Now, the experience of making a classification of a particular domain of linguistic structure tells us that it is very difficult to make a classification permanent (here I refer to a database using a controlled vocabulary, called the "secondary" or "personal" layer in the previous section). The problem is that until the very end the criteria for classifying languages will still be open for change. Only when a project is really finished will it be possible to "publish" a classification. Although there will normally be a classification for the day-to-day usage of the researcher, this classification cannot be made available to others because of its instability.

In contrast, the collection of the relevant information will normally be finished much earlier. This information (in the social layer) might then even be "published" before the final verdict has been reached concerning the classification (in the personal layer). The language-structure annotations might thus be published long before the classification can be published (depending of course on the personal approach to the sharing of information). Any additional relevant information found after the publication of such a annotation will have to be entered in a new annotation linked to the old annotation. One interesting consequence of this approach is that cooperation is possible with other researchers working on a related problem by exchanging such annotations. It would for example be possible for your own database to warn you when new potentially relevant information has been published by colleagues, which might make you reconsider your classification.

It is of course not necessary to publicly make annotation available before the whole typological project has been finished. However, I would propose that such (apparently selfless) sharing of information actually would have a real function in the field of linguistic typology, namely in the determination of primacy of noting something interesting. One of the important functions of linguistic typology is to draw attention to phenomena in particular languages that are noteworthy from a cross-linguistic point of view. Often, a specialist for a particular language will not notice that a certain construction or characteristic of this language is unusual or important from a world-wide perspective. It is the role of the typologist to point that out (note that the ‘language specialist’ and the ‘typologist’ in this situation might of course very well be the same biological person). The recognition that a particular structure in a language is noteworthy is an important scientific achievement, and is also treated as such in the field of linguistic typology: many typologists are proud of having found a language that clearly illustrates something of interest (of course they have not ‘found’ the language in the literal sense, as mostly they only ‘found’ the description). Now, publishing such a noteworthy characteristic in the form of a language-structure annotation will fix scientific primacy for this discovery. Others that want to use the same insightful case will now have to cite this annotation, and thereby promoting the status of the author of the annotation.

A final important implication of the structure of typological databases proposed here would be that it would be much easier to start off any new typological project by using the annotations already published in the social layer. As soon as a researcher has formulated a typological question, a complete search through all published language-structure annotations can give a quick impression of the previous work on that question, on the amount of available information, or maybe even on the rough outline of the worldwide diversity. The important point here is that language-structure annotations are humanly readable text, with “tags” in the form of the research questions to which they belong.

This kind of information is relatively easily searchable by google-like text mining. At least, is is much easier to make searchable than a collection of databases full of semi-cryptically encoded formal fields. Such databases can only be linked by referring to fully formulated ontologies of linguistic categories, something that I consider to be impossible considering the worldwide variation in human language structure (or at least very difficult and time consuming, with probably only a limited pay-off). In searching through the social layer of human-readable notes, authors of databases do not have to decide on any difficult (or even undecidable) terminological definitions, because here we can leave it to the (linguistically proficient) user to decide on what makes sense or not.

5. Conclusion

The structure of a typological database as proposed in this paper is not very different from the current practice. Actually, it only makes a little change: instead of coding a language type and adding a note or reference to this coding (current practice), I propose to consider the notes and references the primary kind of information, and see the coding in a restricted vocabulary (i.e. the database proper) as an addition on top of these annotations. In a foreseeable future of widespread internet publication, such annotation might even become a kind of micro-publication that will be individually citable. The collection of all these notes will then form a social layer for discussion of linguistic categorisation, making various new kinds of collaboration, searching, and scientific interaction possible.

References

Baerman, M., D. Brown, and G. Corbett. 2002. *The Surrey Syncretisms Database.* Available online at http://www.smg.surrey.ac.uk/syncretism/index.aspx.

Comrie, B. 2005. Alignment of case marking of full noun phrases. In: Haspelmath et al., Chapter 98.

Gast, V., D. Hole, E. König, P. Siemund, and S. Töpper. 2007. *Typological Database of Intensifiers and Reflexives.* Available online at http://www.philologie.fu-berlin.de/~gast/tdir/.

Haspelmath, M., M. S. Dryer, D. Gil, and B. Comrie. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Iggesen, O. A. 2005a. Asymmetrical case marking. In: Haspelmath et al., Chapter 50.

Iggesen, O. A. 2005b. Number of cases. In: Haspelmath et al., Chapter 49.

Moyse-Faurie, C. 1983. *Le Drehu, langue de Lifou (Iles Loyauté).* Paris: Selaf.