

Working Group 6

From EMeld

Working Group 6 – Standards and Data Models

Members: Thieberger (Co-Chair), Hinrichs (Co-Chair), Cysouw, Sloetjes, Yi, Vaselina, Langendoen, Beck, Anderson

Issues in standards for the linguistic community

The community is very small and can't create and adopt standards in the way that ISO does. In this context, standards emerge from practice so we have focused both on formal standards where they exist, but also on the most commonly used formats, usually arising from the use of particular tools, and these are the formats that should be addressed by any interchange format.

We noted that adherence to standards where they do exist is critical, for example ISO 639 on language names, has meant that large amounts of data can now be retrieved in a way that was not possible before the linguistic community agreed to use ISO 639.

One initiative that holds promise for standardisation of terminology and shared understanding of analytical approaches is the [Glottopedia](#), a wiki encyclopedia of linguistics.

However, it is clear that the ISO standards route, for example, is too expensive and labor intensive for us to pursue. In the absence of formal standards we need to recognise that some common practices are defacto standards and may be *good enough* in the spirit of Voltaire's "the perfect is the enemy of the good".

Data - Type - Description

Summary of standards (as per [Rumble, John Jr., Bonnie Carroll, Gail Hodge, and Laura Bartolo. 2005. Developing and Using Standards for Data and Information in Science and Technology](#))

Lexicography

- [MDF](#) in toolbox - informal
- [TEI](#) - formal (ISO)
- [LMF \(lexical mark-up framework\)](#)- formal/in development
- [WordNet](#) - implicit
- DATR - informal
- [LIFT \(Lexicon Interchange FormaT\)](#)- interchange standard for lexica - informal (could include reference to GOLD and TMF).

Terminology

- [TMF](#), how to describe terms rather than specifying particular terms,e.g., GOLD could conform to this. - formal (ISO 16642)
- [Feature Structures - ISO TC37/SC3](#) - formal
- [Leipzig Glossing Rules](#) - informal
- [TBX \(TermBase eXchange\)](#) -

Multi-media annotation

(no real standards but multiple annotation tools and corpora, each with their own formats, see the [annotation list here](#))

Language Identification

- [ISO639...](#) -formal
- [WALS](#) -informal
- [MultiTree](#) - informal

Metadata

- [OLAC](#) - formal
- [IMDI](#) - informal(Why would this be considered "informal"?)

- DDI - formal
- [OLIF](#) - proposed standard

Word Lists

- Swadesh - informal
- [IDS](#) more info: [\[1\]](#)- informal
- Russell Gray's lexical database coding, as used for Austronesian, Bantu & Mayan. More info: [\[2\]](#) - informal
- Starostin's Tower of Babel database. more info: [\[3\]](#) - informal

Semantic Fields

- [DPP](#), Dictionary Development Process (Ron Moe) -

Cognate Sets

- [False Friends](#)
- [Starling](#)

Typology

- [Typological Database System \(TDS\) Project](#) is providing a portal to diverse databases (not so much a standard as means of accessing variously encoded information)

Character Sets

- [Unicode](#) - Formal (rendering problems, fonts)

Standards/Models we need

Then we moved on to discuss needed standards/models and identified the following topics:

Grammar writing and grammar models

There are two ways of understanding this topic:

1) computationally functioning grammar formalism: e.g. Finite state morphology (e.g. Xerox XFST)

The area of grammar formalism seems ripe for standardization. The draft document for encoding typed feature structures (chapter authors: Langendoen and Simons) - ISO TC37/SC3 could form a good basis for a standard encoding of augmented phrase structure. XFST and closely related formalisms for encoding finite-state automata and finite-state transducers are equally suitable starting points for standardization.

2) grammatical description: templates (Routledge, FLEX), questionnaires (Lingua, Dahl-Tense Aspect..)

There is no tool to assist in grammar writing, despite the appearance of a grammar authoring system written by Rand Valentine in HyperCard in the early 1990s ([Rook](#)). It seems that [Fieldworks / FLEX](#) is going to provide a grammar authoring environment in the near future and that may provide a standard.

See Jeff Good's paper [on developing a grammar authoring system](#). A practical implementation of such ideas is Sebastian Nordhoff's *Galoes* system [\[4\]](#).

A grammatical description can have various approaches, for example, working from a language-specific form (onomasiological), and this might be conceived as a generalised lexicon entry, or working from a cross-linguistic concept or a questionnaire (semasiological), for example as collected at [\[5\]](#).

Interlinear Text

There is no standard model for the construction of interlinear glossed text (IGT), despite the but it is a very commonly used form of data. Currently several initiatives are addressing IGT and ideally will coordinate their activities.

Annotation of a corpus

Standards for a corpus resulting from fieldwork (not necessarily the same as the corpora being developed for NLP) - none identified.

Typological databases

Do they need standards? Typically typological databases are in the form of a table, but they need to be mappable onto each other. This requires an ontology-like system to ensure terms relate and a language-mapping, for example through ISO 639. However, ISO-639 may be too high level and our recommendation is that references should be to a (written) source rather than to a language, as there may be different varieties and different analyses. So, a recommendation is that any generalisation should refer to sources first,

and these sources should then be related to language identification (perhaps linking to them in a bibliography which identifies language for each entry).

Citation of data

Need to establish a citation form for primary data. Archiving the data provides a persistent identifier for the data, so we need to agree on the form for the citation which includes who the participants are. Work has been done on this by Heidi Johnson at [AILLA](#) and is being followed up by [Delaman](#).

More problematic still is the citation of secondary data, for example in the form of large database compiled by multiple authors. Exactly how the various contribution should be cited normally differs from project to project, and there is yet no clear standart for database citation emerging.

Orthography

Need for an orthographic mapping statement standard to allow comparison of texts in various orthographies (including character sets). Unicode Consortium would like to have a listing of the characters that are used for various languages, and this could be sent in to a Common Locale Data Repository. (It is possible that such a listing could include information on various character sets, encodings, multigraphs, etc.)

Competing Standards

While not really competing standards there are competing practices, for example, using MS Word or Toolbox for dictionary creation, and this is where we need advice for new practitioners, so that they can understand the costs/benefits of each approach. Sources of such advice include the [EMELD pages](#), [SIL](#), [Resource Network for Linguistic Diversity](#)

Recommendations - How to go forward from here

Despite the lack of standards we feel that there are data models that have currency due to their ubiquity and efficacy in achieving immediate goals. These models need to be addressed in order to build interoperability.

As a community we have been working towards interoperable data models via involvement in various activities, e.g., EMELD, OLAC and we feel that these could be extended in future. We have found that sharing expertise and experience at venues like

EMELD, DTSL and TILR is crucial to building both a sense of community and the models for improving individual work practices.

If standards are to be developed they must be easy to read, especially for the target group.