

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Position Paper

Cambell Prince
SIL

As part of linguistics research, field linguists collect and analyze a range of data – one example being the lexicon. Currently the majority of this data is in SFM, which is an impoverished format creating a barrier to data interaction and migration. While it may be considered desirable to move to a more highly specified format or to other tools such as FLEX, in practice this is difficult. Some of the difficulties are:

- It may not be known that the data is under-specified, or ambiguous.
- It is currently difficult to test data against a schema.
- The rules required to infer structure differ from time to time between datasets, and also within the dataset.

Two projects that aim to promote improved interaction between software systems with respect to the lexicon are 'Solid' and 'LIFT'.

Solid – Reducing Technical Barriers to Data Interaction

Solid aims to address the quality issue surrounding SFM based lexicons. Solid improves SFM data by validating the data against a schema, while allowing the data to remain in SFM. It is hoped that this tool will be used to improve the quality of SFM data in the field.

Solid also provides tools that allow data to be exchanged with other personnel and software systems. Data can be edited and enhanced to become compliant with a more explicit schema MDF for example. Further the data can be exported to LIFT, or another format suitable for import into FLEX.

Questions to be explored:

- What quality issues exist in SFM datasets that constitute a barrier to data interaction.
- How does Solid promote data interaction by promoting conformance to both arbitrary, and standard schemas.



LIFT – A Standard Format for Data Interchange

In the wider context of computing standards for data interchange are evident, and by many measures successful. Examples include HTML, and MIME. These standards allow many disparate software systems to interact successfully with the data that they represent.

In the realm of Lexicography, one standard currently being promoted is LIFT. I am currently exploring the use of LIFT in the context of incorporating support for this standard into 'Solid'.

Questions to be explored:

- How do standards such as LIFT benefit data exchange?
- How can these standards be implemented and used?
- What barriers to adoption exist?

Further information on the projects mentioned above can be found at the following websites.

- Solid <http://projects.mseag.org/solid/>
- LIFT <http://code.google.com/p/lift-standard/>
- WeSay <http://www.wesay.org/>

