



GLOBAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY
SOFTWARE & DATA ENGINEERING

Volume 13 Issue 5 Version 1.0 Year 2013

Type: Double Blind Peer Reviewed International Research Journal

Publisher: Global Journals Inc. (USA)

Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Review of Multilingual Glossing and Related Formats

By Preeti Kamboj & Punita Meelu

Doon Valley Institute of Engineering & Technology, Karnal

Abstract - This review paper gives a Historical review of language documentation formats and various glossing techniques available. There is a need of a Multilingual glossing format that is capable of providing features described in this paper. The paper introduces advanced glossing and tools available for the same.

Keywords : *linguistics, annotations, glossing, glossing tables, at, documentation formats.*

GJCST-C Classification : *D.0*



Strictly as per the compliance and regulations of:



A Review of Multilingual Glossing and Related Formats

Preeti Kamboj^α & Punita Meelu^σ

Abstract - This review paper gives a Historical review of language documentation formats and various glossing techniques available. There is a need of a Multilingual glossing format that is capable of providing features described in this paper. The paper introduces advanced glossing and tools available for the same.

Keywords : *linguistics, annotations, glossing, glossing tables, at, documentation formats.*

I. INTRODUCTION

In recent years, the documentation of languages, especially of endangered languages, has received a growing interest within the linguistic community. Most researchers agree that the core of language documentation should consist primarily of recorded and transcribed texts which should not only be translated but also annotated or glossed to be of use for a wide range of purposes.

However, although Lehmann proposes as the principal aim of that format “to make the grammatical structure [of a text in an unknown language] transparent”, all it provides is a rendering of the “meaning or function” of the individual morphemes.

A type of format widely used, especially in the context of grammar writing within a functional framework, is interlinear morphemic translations, or interlinear glossing (IG) for short, first systematized by C. Lehmann.

This has proven useful when exemplifying established facts or illustrating a discussion, but it is by no means sufficient for the aims of language documentation. These aims include that it should be possible to write a grammar of the language or variety being documented, given a sufficient large number of texts that are completely documented^[1].

a) Types of Glossing Information

1. **Syntactical:** Syntactical information includes Syntax related information of the language.
2. **Morphological:** Morphological information includes various lines that contain the instances, phonological, semantic, categorical, structural and relational information.

Author α : Student, M.Tech(C.S.E), Doon Valley Institute of Engineering & Technology, Karnal.

Author σ : Assistant Professor C.S.E Department, Doon Valley Institute of Engineering & Technology, Karnal.

AG (Advanced Glossing) is a general format that does not stipulate details of a possible technical implementation. However, it is obvious that in the digital age any such format should be applicable by means of appropriate computer software.

b) Applications of Advanced Glossing

1. To provide transparent structure to any documentation regardless of languages.
2. Multilingual Environment.
3. Feature Extraction
4. Interactive teaching and E-Learning
5. Application in computer aided machine translation system.

II. REQUIREMENTS FOR A LANGUAGE DOCUMENTATION FORMAT

The following conditions (which may not be independent) as minimal requirements that any language documentation format must meet if the documentation is to be suitable for the purposes of linguistics:

- i. It must be possible to write a grammar of the language or variety being documented given a sufficiently large number of texts completely documented within that format.
- ii. The language used as the documentation language must be clearly interpretable; in particular, it must be possible to clearly distinguish between phonetic, phonological (phonemic), morphological, syntactic, and semantic information in the documented text.
- iii. The documentation format must allow both for partial documentation of a text and for the gradual, systematic filling in of gaps during the documentation process.
- iv. The documentation format must be such that information gathering for the complete documentation of a text (which may not be possible in all cases) can, in principle, be achieved under field conditions.

The character of these conditions: they are (i) minimal and (ii) conditions for the purposes of linguistics. Even for these purposes, additional conditions easily come to mind, and the above requirements do not yet cover the conditions imposed on language documentation for purposes outside linguistics.^[2]

These requirements must be met, if one wants a language to be documented. The Typological Glossing doesn't meet these requirements. Thus the need for something more is felt. Then came the Advanced Glossing. Advanced Glossing meets these all requirements and along with that provide the structure in the form of Advanced Glossing Tables. (AGTs)

III. RELATED WORK

Sebastian Drude and et. al. in their paper "Advanced Glossing: A Language Documentation Format and its implementation with Shoebox"^[1] presents a proposal for a general glossing format designed for language documentation, and a specific setup for the Shoebox-program that implements Advanced Glossing to a large extent. AG provides specific lines for different kinds of annotation – phonetic, phonological, orthographical, prosodic, categorical, structural, relational, and semantic, and it allows for gradual and successive, incomplete, and partial filling in case that some information may be irrelevant, unknown or uncertain. The implementation of AG in Shoebox sets up several databases. Each documented text is represented as a file of syntactic glossings. The morphological glossings are kept in a separate database.

Hans-Heinrich Lieb, Sebastin Drude presents the requirements for a language documentation format in "Advanced Glossing: A language Documentation Format (Working Paper, November 2000)"^[2]. The language document format must be suitable for the language linguistics format. It shows that the Typological Glossing doesn't meet these requirements. They also emphasizes on the need of the presentation of the Advances Glossing.

J D Riding presents their work in "Statistical Glossing - Language Independent Analysis in Bible Translation"^[3]. Bible translation sets a number of particular challenges for machine translation and analysis. The nature of the work is such that translators are often working with local vernacular languages for which there are little in the way of lexis and other linguistic databases. Commercial text processing and translation systems are rarely able to contribute. Translation Consultants (TCs) charged with advising translation teams often have few computer based aids to assist them in their work. This paper describes the development of the Statistical Glossing Tool (SGT) which ships with the Paratext translation editor. SGT is a language independent method for the analysis of bible translations. It provides an objective assessment which TCs can use to help them identify key issues in a new translation where further work may be needed. SGT requires no information about target languages other than the text itself.

Puneet Kamboj, Jatin Aneja, Ajay Malhotra provides their work in "Web and Mobile based Dynamic Glossing tool for Foreign Language Text".^[4] In that they says that Annotation is add-on information attached at a particular point in a document. Much software exists for attaching annotations to document text. This thesis uses a concept of hot words similar to annotations for attaching information to keywords or phrases in published books. Hot words provides a new way of describing keywords or phrases in books by attaching text, image, audio and video based information to them. To ensure global reach this software is presented on web and mobile based platforms supporting multiple languages with an ability to add more languages.

The software developed serves as a great tool for e-learning by providing more information on keywords or phrases. Students can also contribute by suggesting additional annotations. This keeps the information ever expanding.

The Gloss Tool is based on three tier architecture with web service used to develop business layer to expose complete API over the web.

IV. CONCLUSION

There exist various methods and tools to incorporate multilingual information along with annotations. One of such tools is SIL Toolbox.

This paper describes techniques and Applications related to incorporating Glossing Information into a Language independent format. The format must support Advanced glossing Tables, hence ability for future extension.

The paper also depicts work carried out in the field of Advanced Glossing, majorly by "Sebastian Drude".

V. FUTURE SCOPE

Although there exists a lot of work on Language Glossing and Advanced Glossing Tables. I was looking for a method to actually extract information from a Multimedia Glossed Table, regardless of Language boundaries.

A feature Extraction tool or Method would help many endangered Languages in following ways:

- It can be used to Extract Dictionaries.
- It can be used to Extract Morphological Information.
- It can be used to extract the syntactic information.
- It can be used to translate most important works in that Language.
- Can be applied in computer aided Statistical machine translation systems.

REFERENCES RÉFÉRENCES REFERENCIAS

1. Sebastian Drude "Advanced Glossing - a language documentation format and its implementation with Shoebox.
2. Hans-Heinrich Lieb, Sebastin Drude, "Advanced Glossing: A language Documentation Format"(Working paper, November 2000).
3. J D Riding "Statistical Glossing - Language Independent Analysis in Bible Translation".
4. Puneet Kamboj, Jatin Aneja, Ajay Malhotra "Web and Mobile based Dynamic Glossing tool for Foreign Language Text".
5. Integrated Data Management and Analysis for the Field Linguistic, The Linguist's Shoebox: Tutorial and User's Guide.





This page is intentionally left blank