Using GrAF for Advanced Convertibility of IGT data

Dorothee Beermann
Norwegian University of Science and Technology (NTNU)

Peter Bouda
Centro Interdisciplinar de Documentação Linguística e Social

FARO AS Et selskap for sosialt entreprenørskap og utvikling
– a company for social entrepreneurship and development

NTNU – Trondheim
Norwegian University of Science and Technology

CCURL at LREC 2014, May, Reykjavik, Iceland
Overview

1. Interlinear Glossed Text (IGT)
2. The Whats and Whys
3. Implementation (GrAF: POIO API)
4. A case study
5. Software for Advanced Convertability
6. Future work
IGT is the standard data format in the classical fields of linguistics. They are not only part of the electronic records that linguists have archived for endangered languages, linguists have also posted a large quantity of IGT as part of their publications online. Even more IGT are part of the printed media. The largest resource of structured text from less-resourced languages is IGT. IGT is omnipresent, yet it is hard to process automatically.

Not that people haven't tried. An early attempt to make this data accessible is Lewis, W. D. (2003)
Me-ka-a sɛ Akosua re-didi.
I-say-COMPL COMP Akosua PROG-eat
‘I said that Akosua was eating.’

COMP=Complementizer;
COMPL=Completive aspect;
PROG=Progressive aspect

(Amfo 2010)

Bow et. al (2003)
Interlinear Glossed Text

**IGT Editors**

Toolbox: TXT, XML, WAV  
(IGT and lexicon editor; single-user desktop system)  
Elan: EAF, MPEG, WAV  
(Multi-media editor; single-user desktop system)  
Typecraft: XML  
(IGT editor and IGT bank; linguistic service)

**The IGT manifold**

Hard to tell how many, but perhaps the majority of IGT are created using MS WORD. The state of affairs is still that valuable linguistic data exist only as part of a linguistic publication.
2. our point of departure

Life cycle of Interlinear Glossed Text (IGT)

**workflow fieldwork**
- transcribe
- translate
- annotate

capture → submission → access rights

- type-set in an editor (MS WORD)
- data from introspection
- use of secondary sources

publication

? delivery

second life cycle
2. our point of departure

From de dicto to de facto standards
our approach

Since the late 90ies we aim for

**De dicto standards**

through

'Best practice' guidelines

Leipzig Glossing Rules

**De facto standards**

Work off the data types conventional in those linguistic communities engaged in the creation of IGT

* Language Documentation
  → Toolbox based IGT
* data-driven classical fields of linguistics
  → MS Word based IGT

Facilitate and promote IGT mobility by converting in and out of GrAF using: Poio API

Have a linguistic service as front-end (TypeCraft) and GrAF-based advanced converter as backend (POIO API)
2. Implementation

Graph Annotation Framework (ISO 24612)
It allows language resource management

GrAF consists of:

- An abstract data model (we come to that).
- An API for manipulating the data model.
  
  so we use an API and representations as data structures, not a file format

- An XML serialization of the data model is available.
3. Implementation

GrAF Data Model

[Diagram of GrAF Data Model]

Generated by UModel

www.ntnu.edu
3. Implementation

GrAF Data Model: Example

![Diagram]

- **Node**
  - `+xml:id` phrase id="43631"

- **Link**
  - `+target = utterance/W-Spch/r8`

- **Region**
  - `+id = utterance/W-Spch/r8`
  - `+anchors = 780 4090`

- **Annotation**
  - `+xml:id =43631`
  - `+label = phrase`
  - `+ref = phrase id="43631"`

- **FeatureStructure**
  - `+annotation_value`: bitibane bubiyae naka Turuxhiyae
  - `+time_slot1 = ts4`
  - `+time_slot2 = ts23`
Think of GrAF as an assembly language for linguistic annotation; then Poio API is a library to map from and to higher-level languages.

 Subset of GrAF to represent tier based annotation
  - Interlinear glossed text (IGT)

 Filters and filter chains for regular expression searches on tiers

 Plugin mechanism for file formats
  - Mapping semantics: tiers and annotations to nodes and edges

 Meta-data for additional information (Toolbox, Elan; TypeCraft tier types etc.)
2. Implementation

**TypeCraft** - the target's data model

Association is bidirectional
Numerical relationship indicated

![Diagram of Conceptual Classes](image-url)
4. a case study

Reduced set of tiers

tier_map = {
    poioapi.data.TIER_UTTERANCE: ['phrase', 'utterance_gen'],
    poioapi.data.TIER_WORD: ['word', 't'],
    poioapi.data.TIER_MORPH: ['morph', 'm'],
    poioapi.data.TIER_POS: ['pos', 'p'],
    poioapi.data.TIER_GLOSS: ['gloss', 'g'],
    poioapi.data.TIER_TRANSLATION: ['translation', 'f'],
    poioapi.data.TIER_DESCRIPTION: ['description', 'nt']
}
4. A case study

**Advanced Convertability for Toolbox and MS Word based IGT to TypeCraft**

**Luguru (Kagulu) Language:** [kki]
Tanzania (Bantu)

240 000 speakers

Archived Toolbox project (txt)
http://www.elar-archive.org/index.php

Malin Petzell, Africanist
Göteborg University

**Language Documentation**

**Mandinka:** [mnk]
Mali, Senegal, the Gambia, Guinea, Ivory Coast (Mande)

1.300 000 speakers

IGT based appendix to a book publication

Denis Creissels, Université Lumière (Lyon 2)

Member of the research team Dynamique Du Langage.
Specialised in languages of Africa

**Language Typology**
A long time ago, the hyena and rabbit were friends, then rabbit told the hyena 'let us have a journey'.
Mandinka

I táa-tá wô le ŋáama fô ... súw-o kóno,
3PL aller-ACPP DEM FOC à_la_façon jusqu’à maison-D dans
Elles sont allées comme ça jusqu’à la maison,

i bée ye i la lóo-sít-óo boyi-ndi.
3PL tous ACPP 3PL GEN bois-attacher-D tomber-CAUS
et toutes les deux ont déposé leur fagot de bois.

Saa míŋ be maañóo la lóo-sít-óo kaŋ,
serpent.DREL COPLLOC jeune_épouse.DGEN bois-attacher-D sur
Alors le serpent qui était sur le fagot de la jeune co-épouse

a murum-murun-tá naŋ, a yé musu-keebaa-máa kiŋ, câpät,
3SG tourner-tourner-ACPP CTRP 3SG ACPP femme-âgé-SELECT.D mordre
ADVCL
s’est retourné, il a piqué la vieille,
### Linguistic Analysis of the data

Incoming data recognises mostly the same linguistic Categories which however are not bound to a designated tier (different tiers, several tiers). Categories might be distinguished by typographical means. Inter- and intra-annotator disagreement is frequent.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>l táatá wŏ le ŋáama fŏ ... súwo kóno</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>táatá</td>
<td></td>
</tr>
<tr>
<td>morph</td>
<td>táa-tá</td>
<td>Type1: stem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type2: affix</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type3: clitic</td>
</tr>
<tr>
<td>trans.</td>
<td>aller</td>
<td>Type1: word-word</td>
</tr>
<tr>
<td>gloss</td>
<td>à_là_façon</td>
<td>Type2: word-phrase</td>
</tr>
<tr>
<td>symb.</td>
<td>ACPP</td>
<td>Type1: word gloss</td>
</tr>
<tr>
<td>gloss</td>
<td></td>
<td>Type2: morph gloss</td>
</tr>
<tr>
<td>POS</td>
<td>DEM</td>
<td></td>
</tr>
<tr>
<td>free</td>
<td>Elles sont allées comme ça jusqu’à la maison</td>
<td></td>
</tr>
<tr>
<td>trans.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. A case study
5. Software for Advanced Convertability

TypeCraft Importer

Welcome to Typecraft Importer

Select the input file format:
Toolbox

Please select a language:
Luguru

Select Files:
Browse... KaguruCorpus.txt

- File 1: KaguruCorpus.txt

Welcome to Typecraft Importer

Please match these tags with Typecraft’s ones:

<table>
<thead>
<tr>
<th>Gloss</th>
<th>TC Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV7</td>
<td>NONE</td>
</tr>
<tr>
<td>1PAST</td>
<td>NONE</td>
</tr>
<tr>
<td>PRN</td>
<td>NONE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>POS</th>
<th>TC POS</th>
</tr>
</thead>
<tbody>
<tr>
<td>int</td>
<td>NONE</td>
</tr>
</tbody>
</table>

Back  Import
Luguru seen in the TypeCraft Editor

5. Software for Advanced Convertability


NTNU – Trondheim
Norwegian University of Science and Technology

www.ntnu.edu

CCURL at LREC 2014, May, Reykjavik, Iceland
Future work

ELAN - TypeCraft - bilateral exchange of data

Integration of TypeCraft's tier and annotation inventory into ISO CAT

Converter: TODO
- make converter more stable
- improve user dialog
**Links**

Poio:  
http://media.cidles.eu/poio/

GrAF:  
http://www.xces.org/ns/GrAF/1.0/

TypeCraft  
http://www.typecraft.org
References


Beermann, Dorothee; Mihaylov, Pavel. (2013) TypeCraft collaborative databasing and resource sharing for linguists. *Language Resources and Evaluation*. Springer. The Netherlands


Thank you