



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

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



NTNU – Trondheim
Norwegian University of
Science and Technology

Interlinear Glossed Text (IGT)

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)



Interlinear Glossed Text from Akan [aka]

Text Phrase

Text *25. Amponsa kaa sɛ ɔawie.

Save

Phrase: : Amponsa kaa sɛ ɔawie.

Free translation: Amponsa said that he has finished.

Construction parameters: Change

Word:	Amponsa	kaa	sɛ	ɔawie			
Morph:	amponsa	ka	a	sɛ	ɔ	a	wie
Baseform:	amponsa	ka	a	sɛ	ɔ	a	wie
Meaning:	A.	say					finish
Gloss tags:	SBJ		PAST	COMPL	3SG.SBJ	PRF	
POS:	N	V	PRT	V			

Construction description:
v_past-tr-obDECLcomp

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

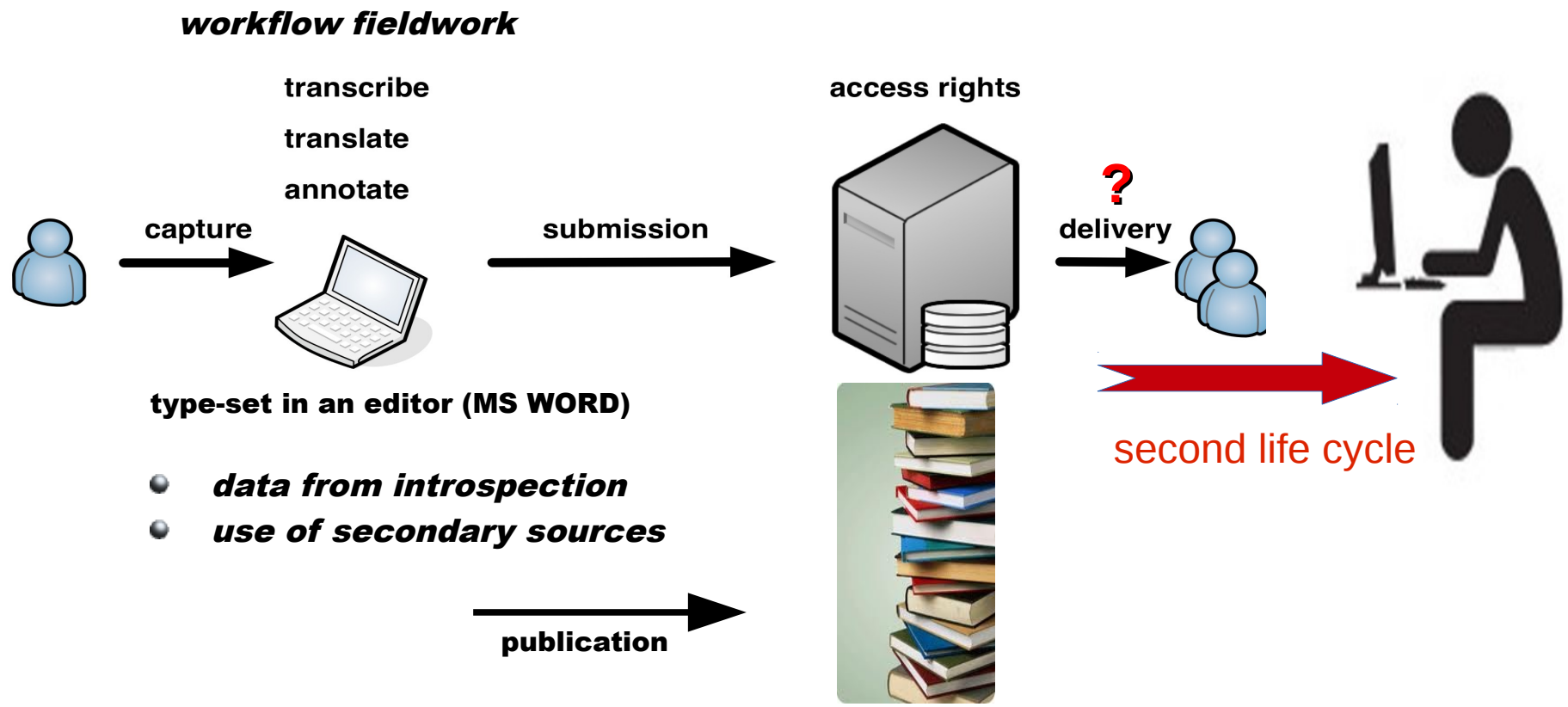
Hard to tell how many, but perhaps the majority of IGT are created using MS WORD. The state of affairs is still that valuable linguistic data exist only as part of a linguistic publication.



NTNU – Trondheim
Norwegian University of
Science and Technology

2. our point of departure

Life cycle of Interlinear Glossed Text (IGT)



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



IGT-DATA MOBILITY

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)



NTNU – Trondheim
Norwegian University of
Science and Technology

GrAF

Development of Linguistic annotation framework (LAF)

Graph Annotation Framework (ISO 24612)

It allows language resource management

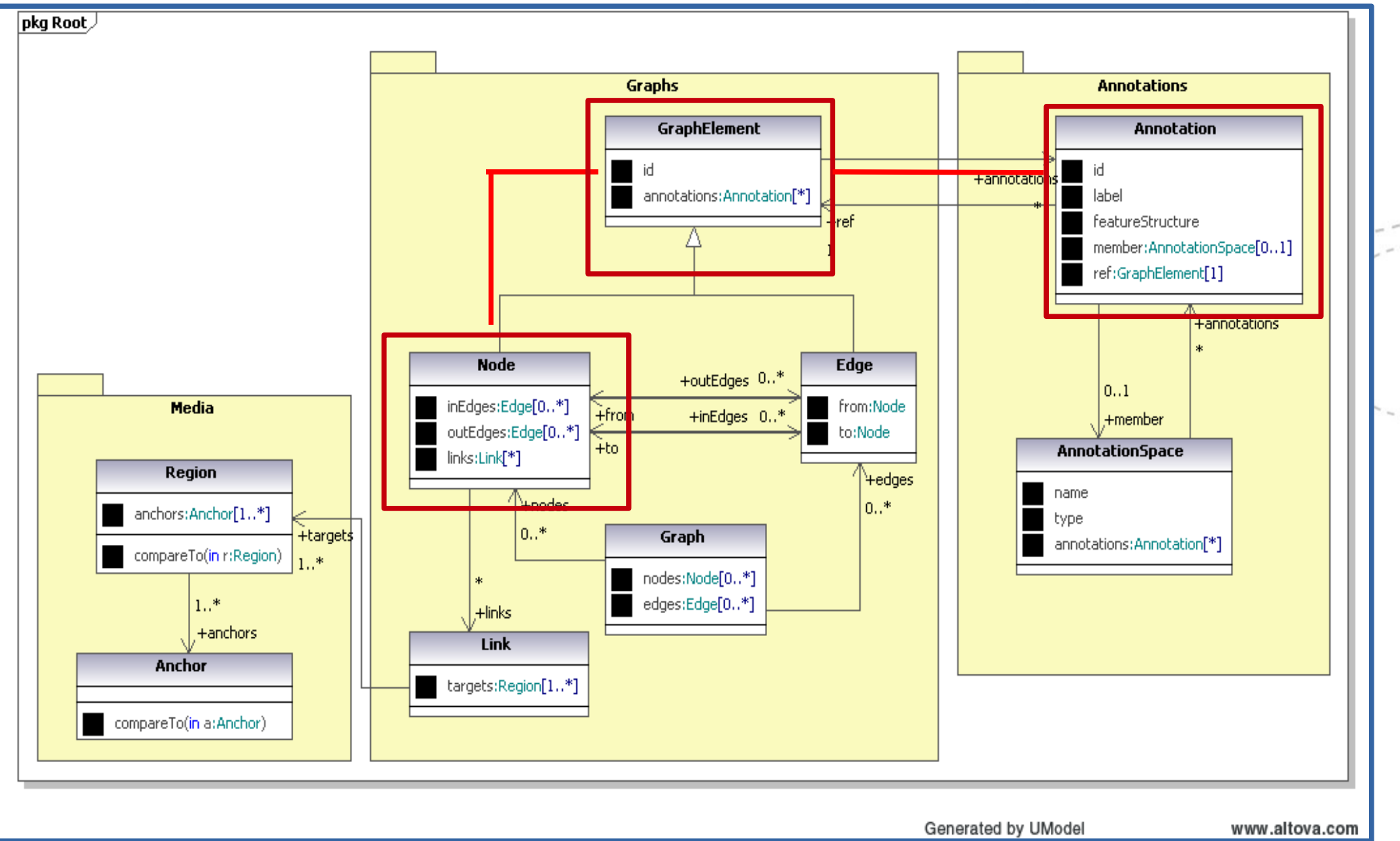
GrAF consists of:

- 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.



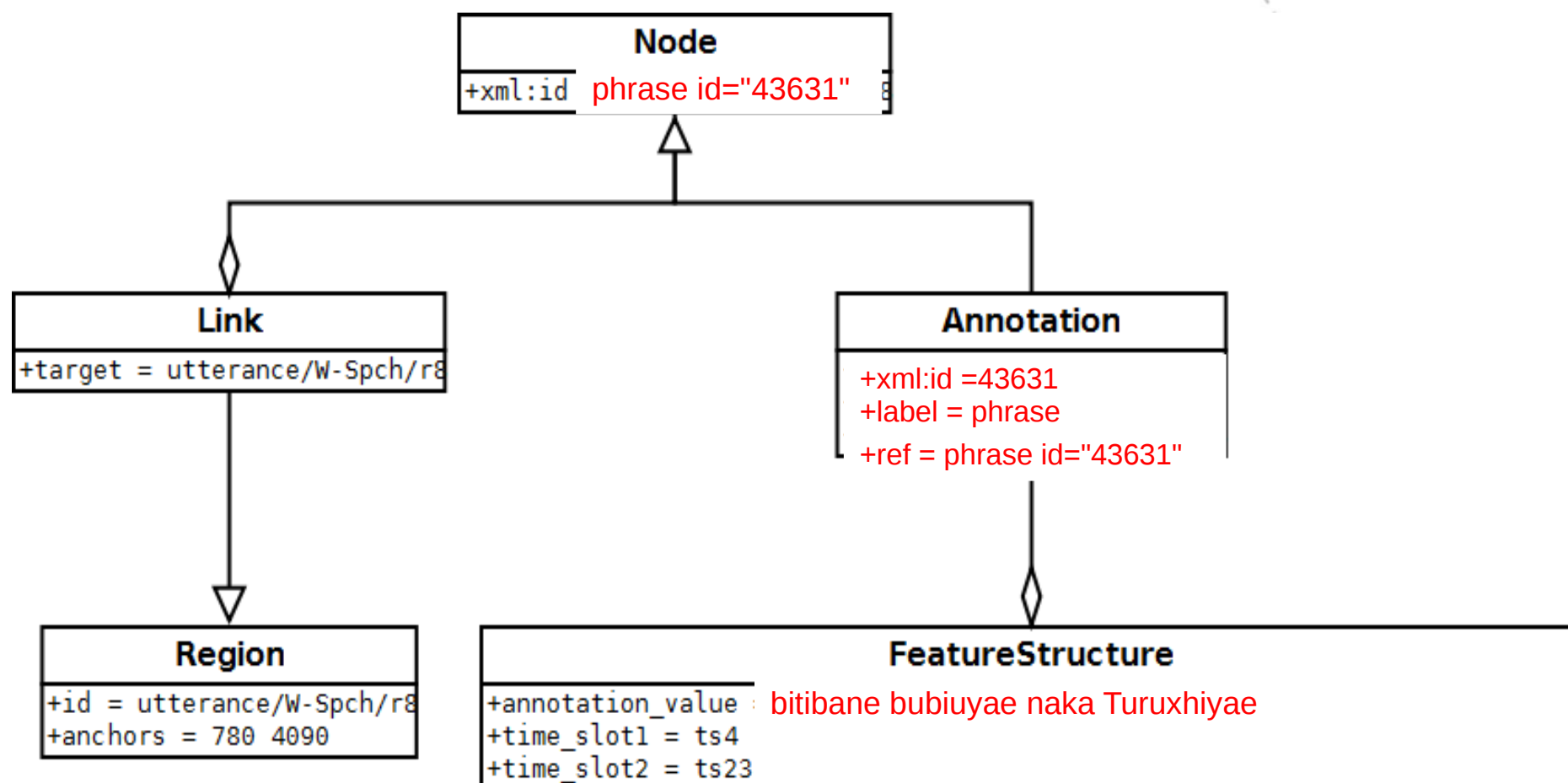
NTNU – Trondheim
Norwegian University of
Science and Technology

GrAF Data Model



3. Implementation

GrAF Data Model: Example





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

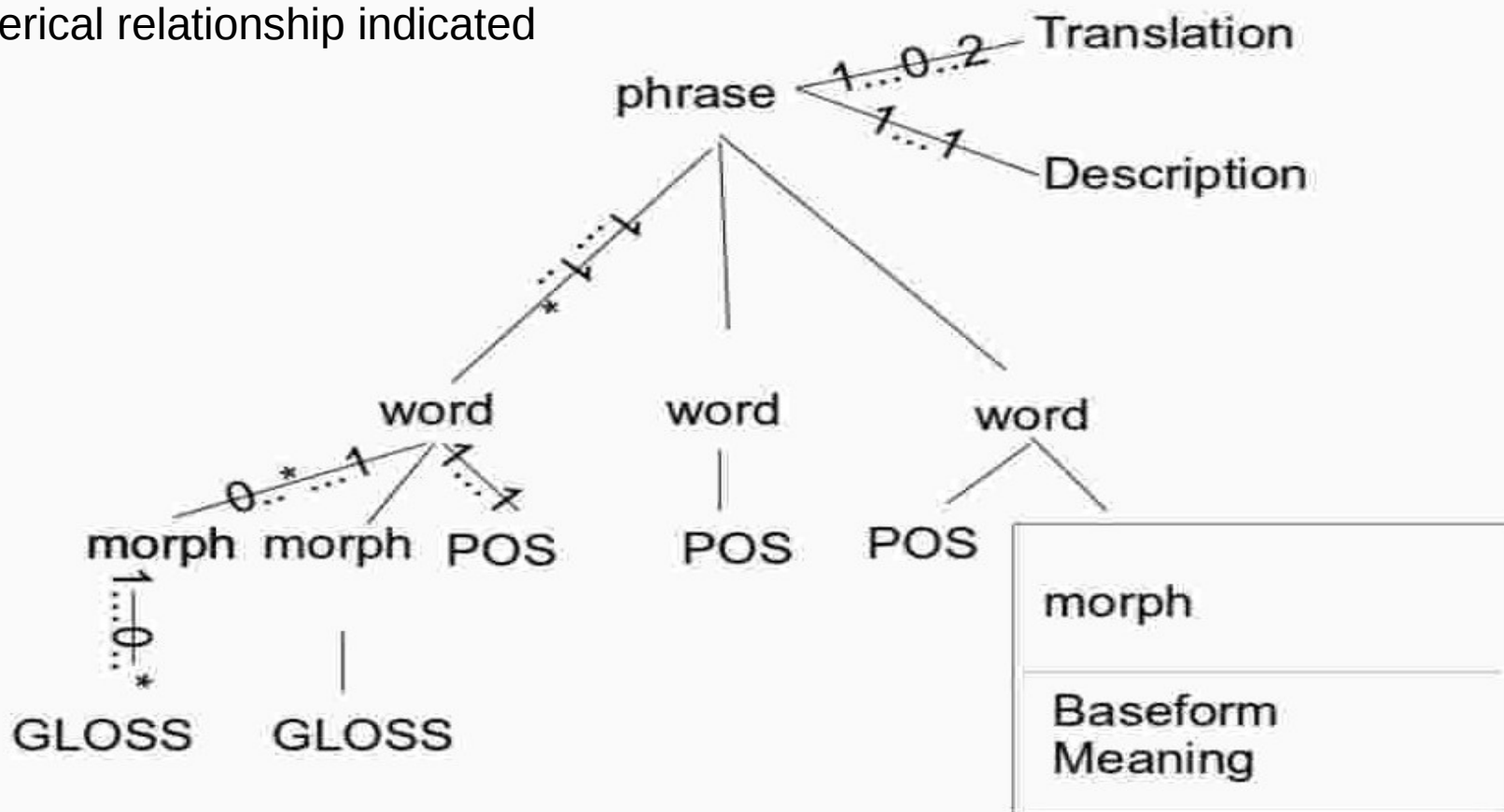
- Mapping semantics: tiers and annotations to nodes and edges

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

TypeCraft - the target's data model

CONCEPTUAL CLASSES

Association is bidirectional
Numerical relationship indicated



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']}
```



Advanced Convertability for Toolbox and MS Word based IGT to TypeCraft

Field
work

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

influential work
in language
typology

Language Typology



NTNU – Trondheim
Norwegian University of
Science and Technology

Luguru

\name mjs1 The hyena and the rabbit

\ref mjs3001revised

\t Baho katali difisi na

\m baho katali di- fisi na

\g DEM long time ago 5- hyena:5/6 CONJ

\p prn adv ncp- n conj

\t sungula hawowa mbuya kamei

\m sungula ha- wa- uw -a mbuya kamei

\g hare:9/10 PAST- 2- be -FV friend:1a,9/2,10 then

\p n tm- ncp- v -fv n adv

\t sungula kamgamba, "chigende nhambo."

\m sungula ka- m- gamb -a chi- gend -e N- tambo

\g hare:9/10 1.PAST- 1- speak -FV7- go -FV 9/10- journey:9/10

\p n sm- ncp- v -fvacp- v -fv ncp- n

\f A long time ago, the hyena and rabbit were friends, then rabbit told
the hyena 'let us have a journey'.



Mandinka

I táa-tá wǒ le ñáama fǒ ... 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 mín be maañóo la lóo-sít-óo kaŋ,

serpent.DREL COPLOC 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, cápǎt,

3SG tourner-tourner-ACPP CTRP 3SG ACPP femme-âgé-SELECT.D mordre

ADVCL

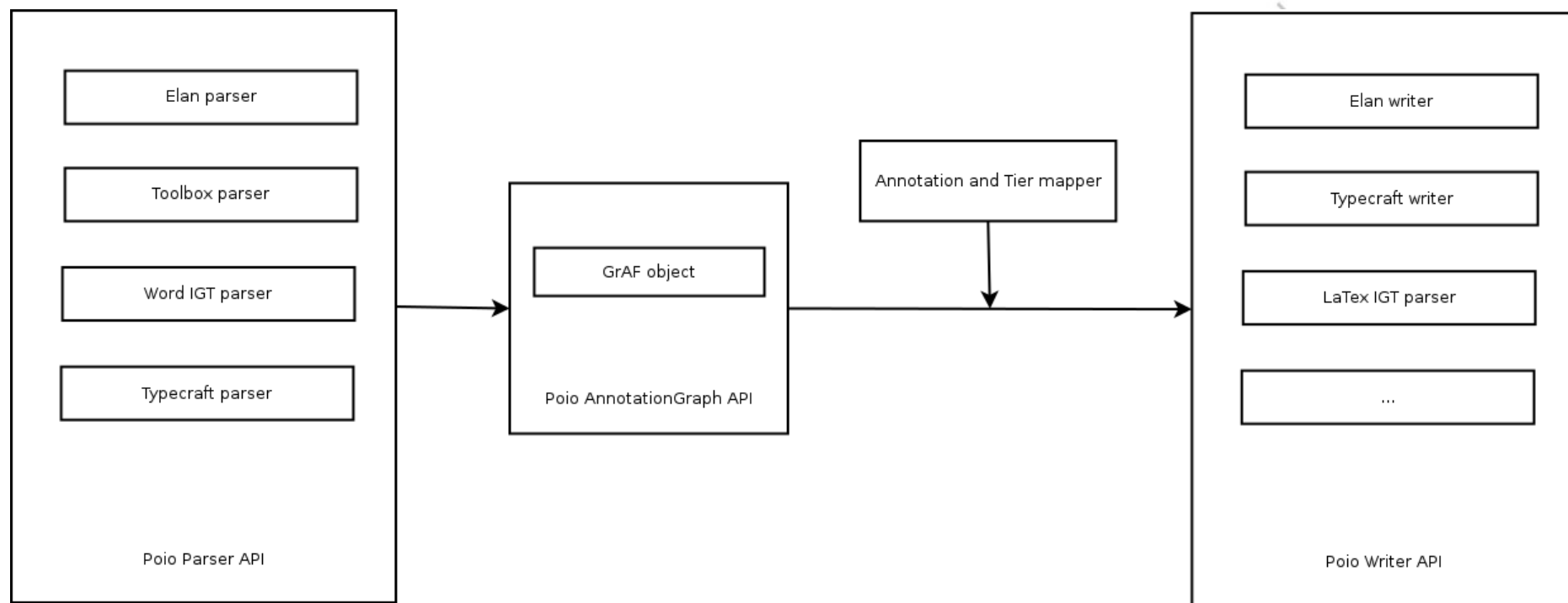
s'est retourné, il a piqué la vieille,



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

Category	Example	Specification
text	<i>I táatá wǒ le ñáama fǒ ... súwo kóno</i>	
word	<i>táatá</i>	
morph	<i>táa-tá</i>	Type1: stem Type2: affix Type3: clitic
trans. gloss	<i>aller</i> <i>à_la_façon</i>	Type1: word-word Type2: word-phrase
symb. gloss	<i>ACPP</i>	Type1: word gloss Type2: morph gloss
POS	<i>DEM</i>	
free trans.	<i>Elles sont allées comme ça jusqu'à la maison</i>	



TypeCraft Importer

lctest.typecraft.org/tx2/jsp/converter.jsp

Most Visited Getting Started Joan Bresnan - Predic... http://nltk.org/book/... SketchGram

Welcome to Typecraft Importer

Select the input file format:

Toolbox :

Please select a language:

Luguru

Select Files:

Browse... KaguruCorpus.txt

- File 1: KaguruCorpus.txt

Validate

lctest.typecraft.org/tx2/jsp/converter.jsp

Most Visited Getting Started Joan Bresnan - Predic... http://nltk.org/book/... SketchGram

Welcome to Typecraft Importer

Please match these tags with Typecraft's ones:

Gloss	TC Gloss
FV7	NONE :
1PAST	NONE :
PRN	NONE :

POS	TC POS
int	NONE :

Back **Import**

Luguru seen in the TypeCraft Editor

5. Software for Advanced Convertability

Text » Phrase » Theme »

Save Share: Private Publish Template New phrase Delete phrase View phrase list View senses

B I U AaBbCc Paragraph AaBb Heading 1 AaBb Heading 2 AaBb Heading 3 AaBb Heading 4 AaBbCc Pre Remove formatting

Language: Luguru Change

Title:

Title translation:

Content description:

Baho katali difisi na sungula hawowa mbuya kamei sungula kamgamba, "chigende nhambo."Kamei howoluta kunyumbangwa imwe, wahokeligwa.Kamei howekala bahaya majuwa mengi ljuwa dimwedu sungula hoyomgamba mwinya ikaya "aseye nhosiku chikuluta ukaya, haya!"Nhechilo sungula kadiya mayowe gose, kamei kalonda meji, kasugusa, keja mutwila difisi.Kamei nhosikusiku sungula hoyomtamilu imukaya mwinya nyumba, "ulandise finhu fyako halika fyose fiswamu".Kamei imwinya k sungula kamgamba, "lete yehoki".Kamei sungula kak meji, kotwila mwigoda meji noyadiile.Difisi dikomigwa.

2. Kamei howoluta ku...

Save FTrans 1 FTrans 2 CParam Base Meaning Gloss POS

Phrase: Kamei howoluta kunyumbangwa imwe, wahokeligwa.

Free translation 1: Then they went to one house and they were welcomed.

Free translation 2:

Constr. params: Change

Word:	Kamei	howoluta				kunyumbangwa		imwe,
Morph:	kamei	ha	wa	lut	a	ku	nyumba	ngwa
Baseform:	kamei	ha	wa	lut	a	ku	nyumba	ngwa
Meaning:	then			go			house	somebody's
Gloss tags:		PAST	CL2		FV	CL17		CL5



NTNU – Trondheim
Norwegian University of
Science and Technology

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

- Amfo, Nana Aba Appiah. 2010. Noun Phrase Conjunction in Akan: The Grammaticalization Path. *Pragmatics* 20:1.27-41.
- Beermann, Dorothee; Mihaylov, Pavel. (2013) TypeCraft collaborative databasing and resource sharing for linguists. *Language Resources and Evaluation*. Springer. The Netherlands
- Blumtritt, Jonathan and Bouda, Peter and Rau, Felix. 2013. Poio API and GraF-XML: A radical stand-off approach in language documentation and language typology. In: Proceedings of Balisage: The Markup Conference 2013.
- Bow, C. and B Hughes and S. Bird. 2003. Towards a General Model of Interlinear Text. In Proceedings of the E-Meld Workshop on Digitizing and Annotating Texts and Field Recordings. Lansing: LSA Institute, Michigan State University.
- Lewis, W. D. 2003. Mining and migrating interlinear glossed text, in 'Proceedings of the EMELD Workshop on Digitizing and Annotating Texts and Field Recordings', East Lansing, MI. <http://emeld.org/workshop/2003/Lewis-paper.pdf>



research
IGT-based
Glossed
different
distribution
classical
Interlinear
Facilitating
linguistics
creation
t
use
data
exchange
linguistic
IGT
annotated
fields
research

Thank you

