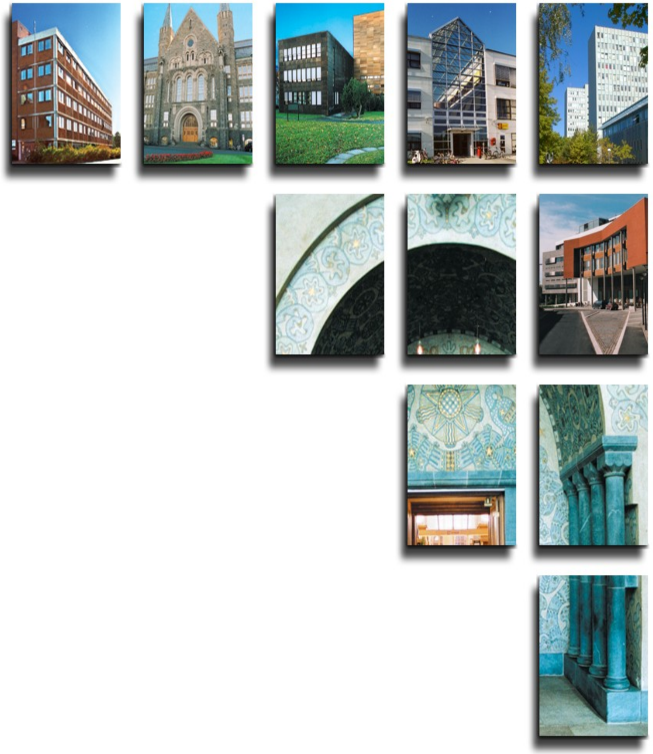
NTNU

Innovation and Creativity



Annotations

Dorothee Beermann

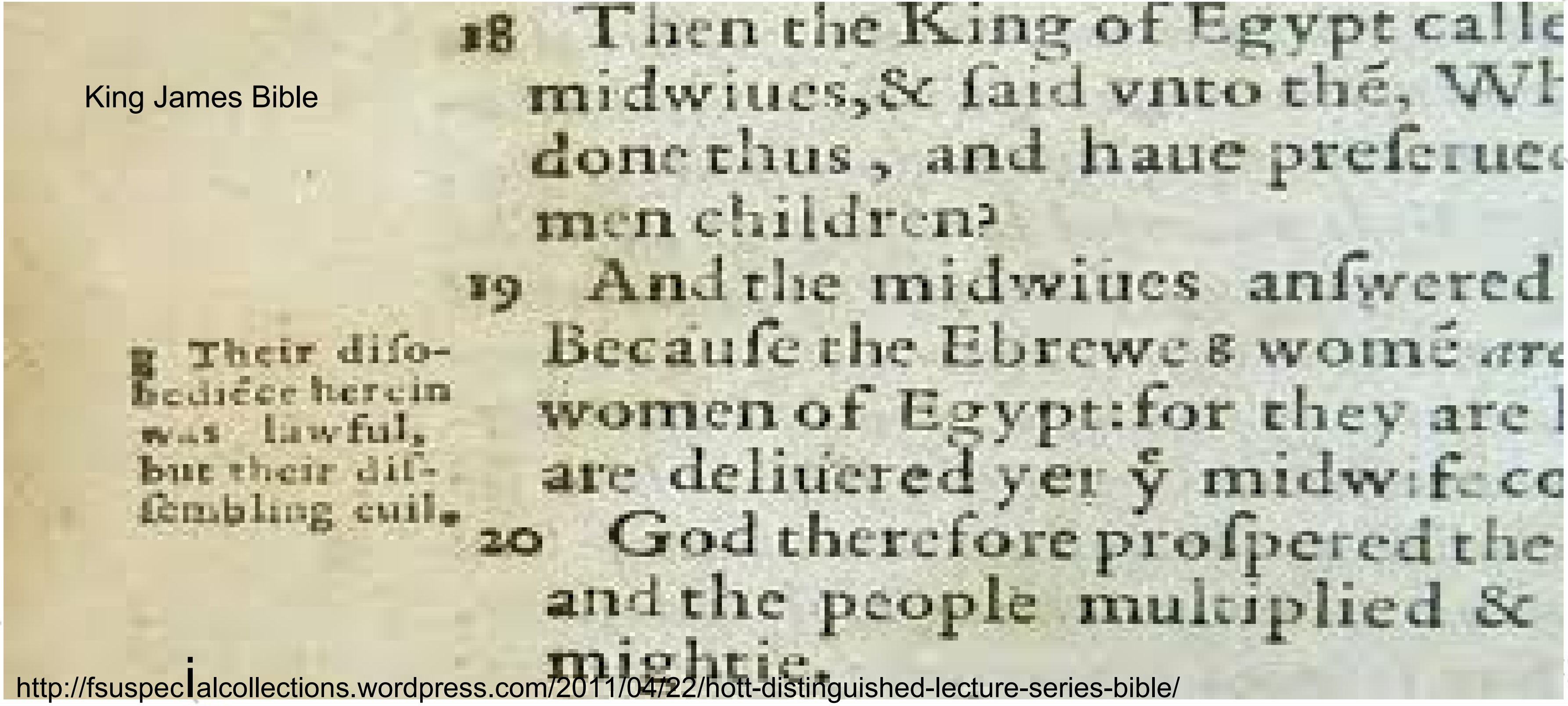
NTNU, Trondheim, Norway

September 2012

Annotations

annotation (n.)
mid-15c., from L. annotationem (nom. annotatio), noun of action from pp. stem of annotare "to add notes to," from ad- "to" (see ad-) + notare "to note, mark" (see note (v.)).

Online etymology dictionary, http://www.etymonline.com/



King James Bible

18 Then the King of Egypt calle
midwiues,& ſaid vnto thé, Wh
done thus, and haue preſerue
men children?
19 And the midwiues anſwered
Becauſe the Ebrewe g womé are
women of Egypt: for they are l
are deliuered yet ẏ midwife co
20 God therefore proſpered the
and the people multiplied &
mightie.

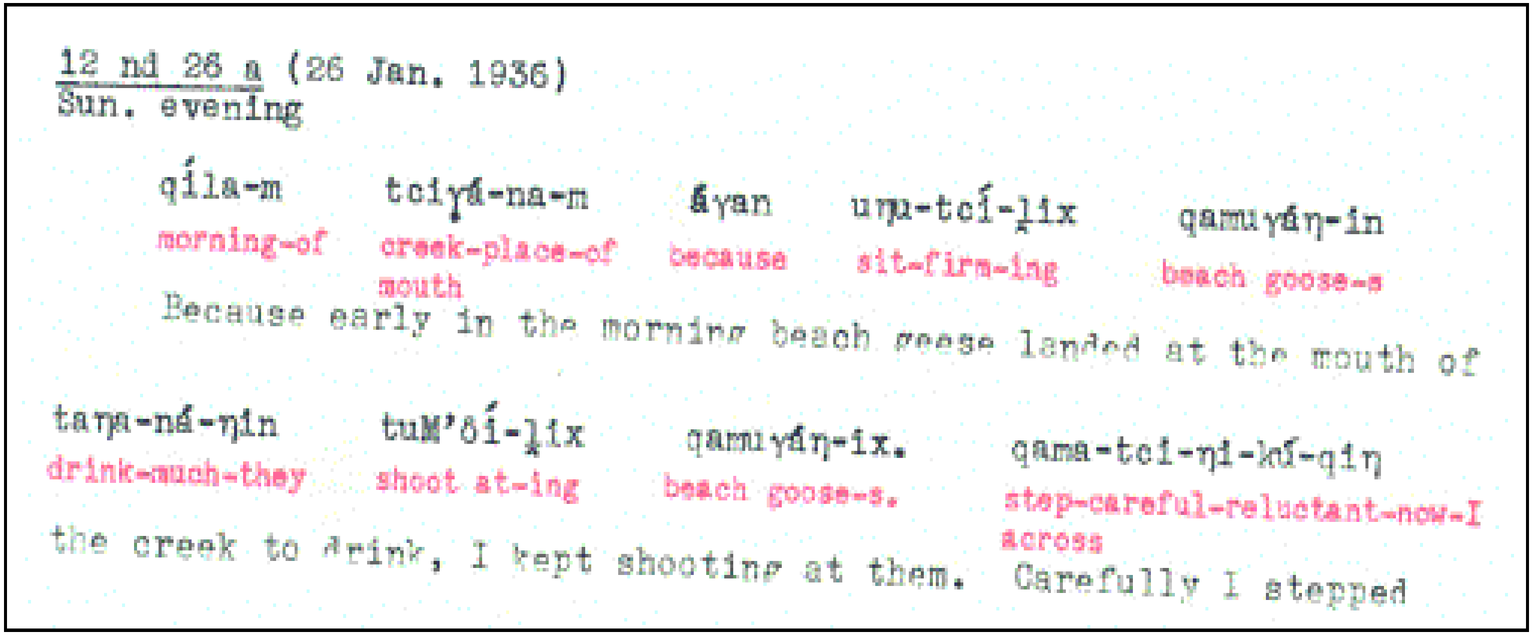
g Their diſobedience herein was lawful, but their diſſembling euil.

http://fsuspecialcollections.wordpress.com/2011/04/22/hott-distinguished-lecture-series-bible/

Linguistic Annotations

Linguistic annotations can be divided into types. Next to glosses annotations may be comments about the source itself, or comments expressing different degrees of certainty. Background information may also appear in annotation. We should not forget editorial annotations. These different types are often mixed on an hoc basis.

Interlinear Glossed Text



12 nd 26 a (26 Jan. 1936)
Sun. evening

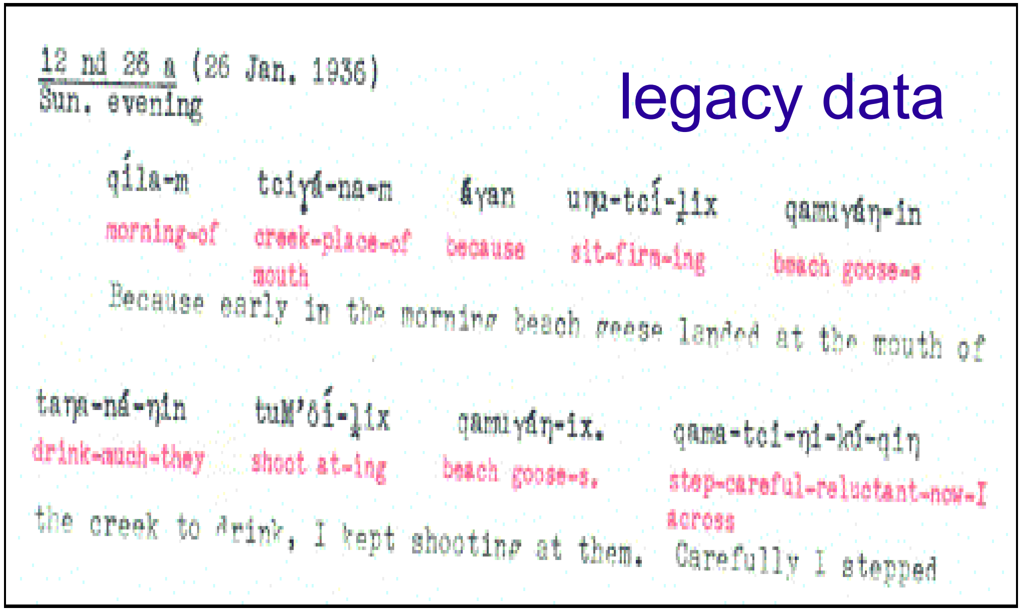
qíla-m	tciyá-na-m	áγan	uŋa-tcí-ḻix	qamuγáŋ-in
morning-of	creek-place-of mouth	because	sit-firm-ing	beach goose-s

Because early in the morning beach geese landed at the mouth of

taŋa-ná-ŋin	tuM'ôí-ḻix	qamuγáŋ-ix.	qama-tci-ŋi-kó-qiŋ
drink-much-they	shoot at-ing	beach goose-s.	step-careful-reluctant-now-I across

the creek to drink, I kept shooting at them. Carefully I stepped

Holten, 2003.



12 nd 28 a (26 Jan. 1936)
Sun. evening

legacy data

qíla-m	tciyá-na-m	áyan	uŋa-tcí-ḷix	qamuyáŋ-in
morning-of	creek-place-of mouth	because	sit-firm-ing	beach goose-s

Because early in the morning beach geese landed at the mouth of

taŋa-ná-ŋin	tuM'ǒí-ḷix	qamuyáŋ-ix.	qama-tci-ŋi-kí-qiŋ
drink-much-they	shoot at-ing	beach goose-s.	step-careful-reluctant-now-I across

the creek to drink, I kept shooting at them. Carefully I stepped

GLOSS

inserted as an explanation, " 1540s (earlier gloze, c.1300). Both glossology1) have been use (1716) and glottology (184d in the sense "science of language."

The first line represents the original text, broken into morphemes using hyphens.
The second line (in red ink) provides an English translational gloss for each morpheme.

Notice that lexemes and functional units receive translations glosses, such as “ing” instead of PROG. The free translation not necessarily reflects the meaning of the glosses. The number of glosses does not always correspond to the number of morphemes which makes it difficult to relate the glosses to the original text.

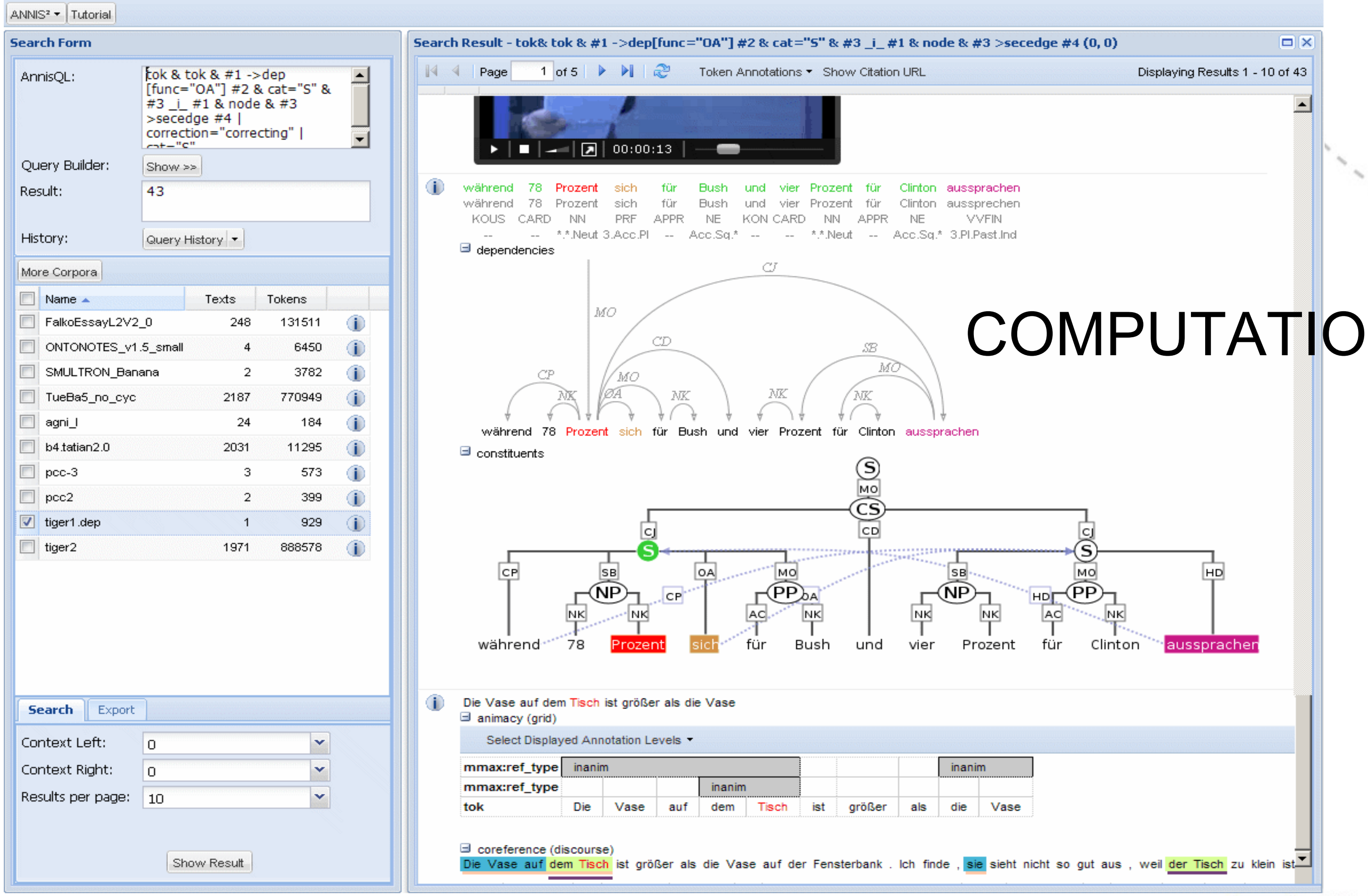
Linguistic annotations

search and visualization architecture for complex multilevel annotated linguistic corpora

ANNIS2

University of Potsdam, Berlin, Germany

COMPUTATIONAL



AnnisQL: tok & tok & #1 ->dep [func="OA"] #2 & cat="S" & #3 _i_ #1 & node & #3 >secedge #4 | correction="correcting" | cat="S"

Result: 43

☐	Name ▴	Texts	Tokens
☐	FalkoEssayL2V2_0	248	131511
☐	ONTONOTES_v1.5_small	4	6450
☐	SMULTRON_Banana	2	3782
☐	TueBa5_no_cyc	2187	770949
☐	agni_l	24	184
☐	b4.tatian2.0	2031	11295
☐	pcc-3	3	573
☐	pcc2	2	399
☑	tiger1.dep	1	929
☐	tiger2	1971	888578

Context Left: 0

Context Right: 0

Results per page: 10

während	78	Prozent	sich	für	Bush	und	vier	Prozent	für	Clinton	aussprachen
während	78	Prozent	sich	für	Bush	und	vier	Prozent	für	Clinton	aussprechen
KOUS	CARD	NN	PRF	APPR	NE	KON	CARD	NN	APPR	NE	VVFIN
--	--	*.*.Neut	3.Acc.Pl	--	Acc.Sg.*	--	--	*.*.Neut	--	Acc.Sg.*	3.Pl.Past.Ind

während 78 Prozent sich für Bush und vier Prozent für Clinton aussprachen

während 78 Prozent sich für Bush und vier Prozent für Clinton aussprachen

Die Vase auf dem Tisch ist größer als die Vase

mmax:ref_type	inanim								inanim	
mmax:ref_type				inanim						
tok	Die	Vase	auf	dem	Tisch	ist	größer	als	die	Vase

Die Vase auf dem Tisch ist größer als die Vase auf der Fensterbank . Ich finde , sie sieht nicht so gut aus , weil der Tisch zu klein ist

NTNU

Innovation and Creativity

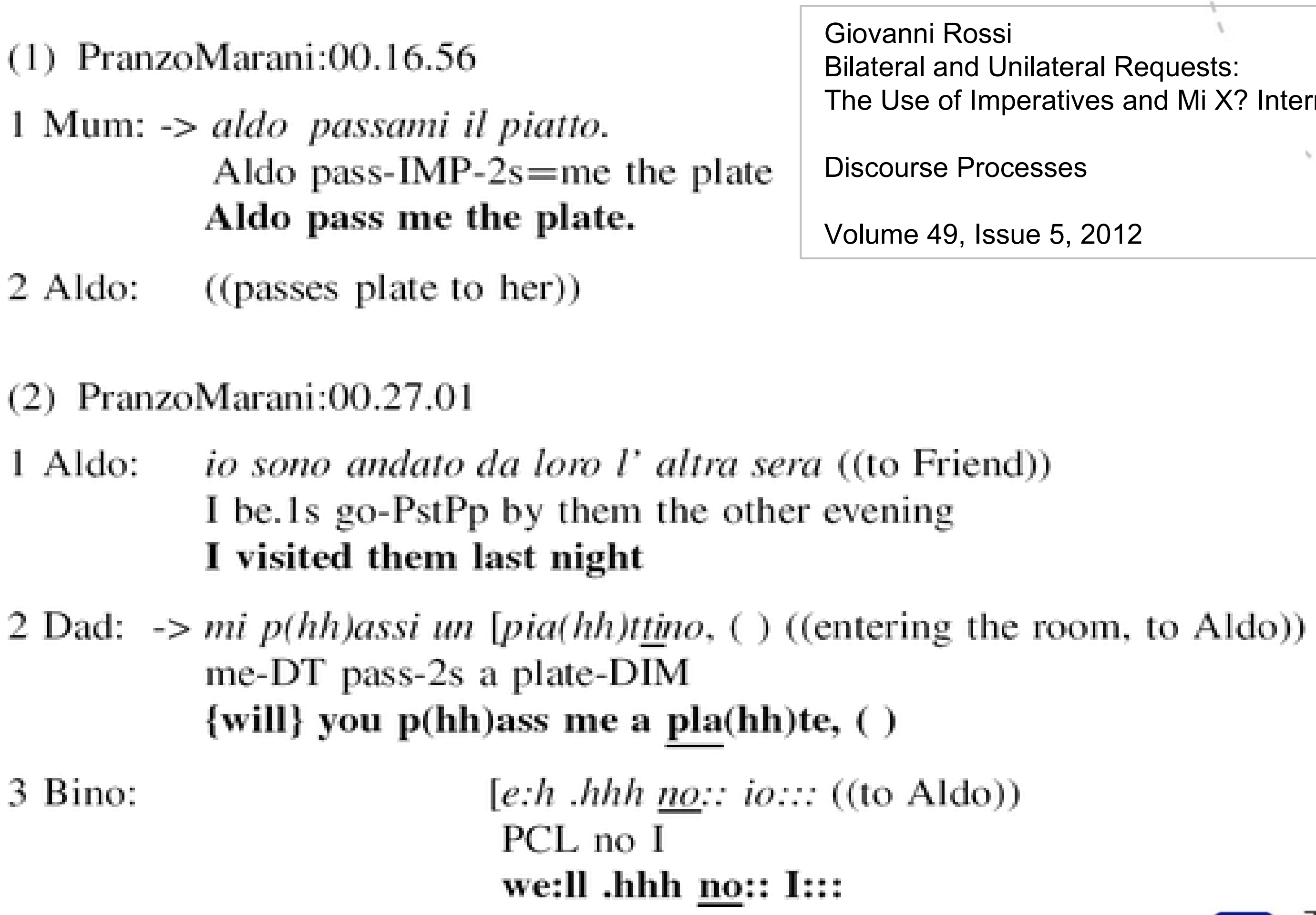
Glosses - Interlinear Glossed Text (IGT)

THE LINGUISTIC DEFAULT

Giovanni Rossi
Bilateral and Unilateral Requests:
The Use of Imperatives and Mi X? Interrogatives in Italian

Discourse Processes

Volume 49, Issue 5, 2012



(1) PranzoMarani:00.16.56

1 Mum: -> *aldo passami il piatto.*
Aldo pass-IMP-2s=me the plate
Aldo pass me the plate.

2 Aldo: ((passes plate to her))

(2) PranzoMarani:00.27.01

1 Aldo: *io sono andato da loro l' altra sera* ((to Friend))
I be.1s go-PstPp by them the other evening
I visited them last night

2 Dad: -> *mi p(hh)assi un [pia(hh)ttino,* ( ) ((entering the room, to Aldo))
me-DT pass-2s a plate-DIM
{will} you p(hh)ass me a pla(hh)te, ( )

3 Bino: *[e:h .hhh no:: io:::* ((to Aldo))
PCL no I
we:ll .hhh no:: I:::

Different linguistic frameworks and their "DATA"

Linguists disagree on what "DATA" is

* naturally occurring language

* annotated expressions (not necessarily only text)

* elicitations in the form of sentence collections

* all of the above

* only naturally occurring language

* only structured data

Linguists differ in what they think "DATA" is
they however agree in what they publish as "DATA"

NTNU
Innovation and Creativity

Linguistic Typology

(1) Lavukaleve (Terrill°, ex. 9)
nga-bakala nga-uia tula
1SG.POSS-paddle(M) 1SG.POSS-knife(F) small.SG.F
'my paddle and my small knife

(6) prepositive: Lenakel (Moyse-Faurie & Lynch°, ex. 28a)
I-em-va m-m-angn.
1SG-PAST-come and-PAST-eat
'I came and ate.'

Coordinating constructions to appear in: Coordinating constructions. (Typological Studies in Language, 58.) Haspelmath, Martin (ed.) 2004. Amsterdam: Benjamins.

NTNU
Innovation and Creativity

Generative Grammars

(1) a. Ú-hlál' é-dolóbh-e:ni.

1SM-live LOC-5.city-LOC

'S/he lives in the city centre.'

b. Ú-nge:n' é-ndl-i:ni.

1SM-enter LOC-9.house-LOC

'S/he entered the house.'

(2) a. Aba-ntu aba-dala ba-hlala ku-lezi zi-ndlu.

2-people 2-old 2SM-stay LOC-10.these 10-houses

'Old people live in these houses.'

Cheng & Downing Locative Relatives in Durban Zulu,

ZAS Papers in Linguistics 53, 2010:33-51

Interlinear Glossed text is an integral part of linguistic publications independent of the linguist's theoretical affiliation. An IGT normally lacks any index to where and when it occurred or any other information that would identify it as a particular instance of a language.
This is not necessarily a problem since the function of IGT is not uniform.

IGT

IGT

In the **logical tradition,** where linguists follow in the footsteps of the philosophical and mathematical sciences, an IGT is an idealised representation of the linguistic reality that the theory describes. Work of Louis Hjelmslev is an example of this approach, and of course Noam Chomsky's work stands in this tradition. This use of the IGT leads to a characteristic style of exposition where IGTs serve as threads of the discussion. Lyons (1977) calls the corresponding type of linguistic data **system sentences**.

IGT

IGT

IGT

IGT

The function of IGT within the empirically-oriented fields of linguistics is different.
Here an interlinear glossed phrase serves as a *Data Sample*.
It might have been gathered through linguistic interviews or other elicitation methods. However, considering the format of the data alone, there are no principled differences between IGTs across linguistic traditions.

What is different is the emphasis that is put on the representativeness and authenticity of the data;
this is where the real difference between the two main schools of linguistics seems to lie.

NTNU
Innovation and Creativity

IGT – problematic data

IGTs are the most common form of annotated data in linguistics.
Yet, it is exactly this type of data that has recently come under scrutiny.
Researchers from different linguistic fields have questioned its validity,
and the integrity of theories that 'are built' on this kind of data.

From the psycholinguistic side it has been claimed that linguists are not
(sufficiently) concerned with methods that regulate data collections.
It has been pointed out that IGT are mostly based on binary grammaticality judgments.

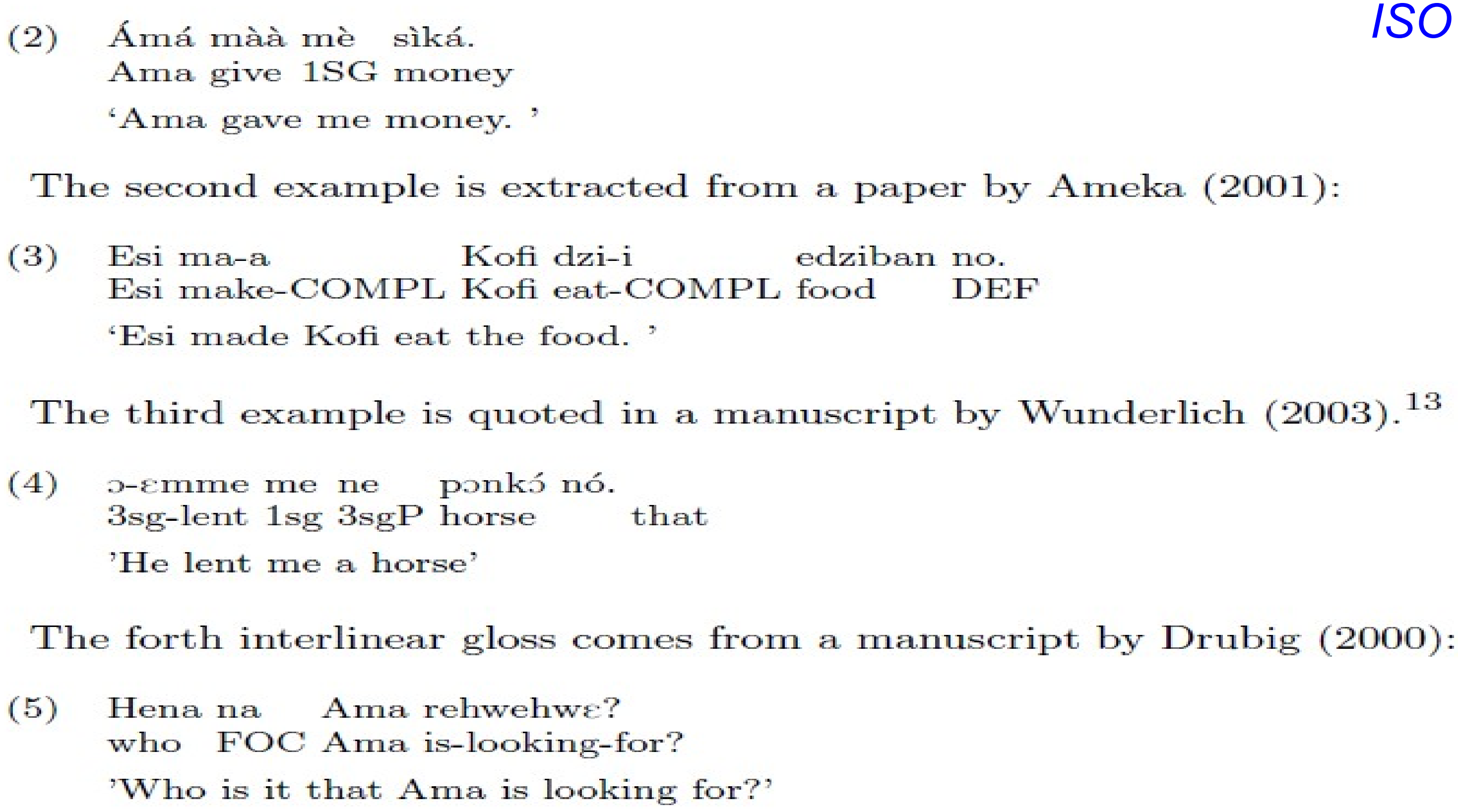
Moreover also IGT, like all other data based on human judgment,
should be exposed to empirical control in order to assure a reliable mode of
data elicitation.
Also from the side of functional linguistics, in particular from the ranks of linguists working with LDD, methodological issues have been raised, calling for the standardisation of IGT and improved methods of data management.

In the classical - functional as well as generative - fields of linguistics,
the lack of glossing standards is still one of the main hindrances
for IGTs to be a prime linguistic resource.

NTNU
Innovation and Creativity

4 IGTs extracted from ODIN*

AKAN,
Kwa language,
Ghana,
ISO 639-1 ak



(2) Ámá màà mè sìká.
Ama give 1SG money
‘Ama gave me money. ’

The second example is extracted from a paper by Ameka (2001):

(3) Esi ma-a Kofi dzi-i edziban no.
Esi make-COMPL Kofi eat-COMPL food DEF
‘Esi made Kofi eat the food. ’

The third example is quoted in a manuscript by Wunderlich (2003).[13]

(4) ɔ-ɛmme me ne pɔnkɔ́ nó.
3sg-lent 1sg 3sgP horse that
’He lent me a horse’

The forth interlinear gloss comes from a manuscript by Drubig (2000):

(5) Hena na Ama rehwehwɛ?
who FOC Ama is-looking-for?
’Who is it that Ama is looking for?’

* ODIN - The Online Database of Interlinear Text

http://odin.linguistlist.org/

Misunderstandings:

Comparing (3) with(4) *nó* is glossed as 'DEF' in (3) and 'that' in (4).
According to most records *nó* is a definite marker, and only given in the right context may be interpreted as a distal marker. *Nó* needs to be distinguished from *nò* which is a 3sg personal pronoun.
The verb ma meaning 'give', must also in example (3) carry low tone on both vowels to indicate the non-present tense form of the verb.
In (4) the free translation of the object as *'a horse*' is inconsistent with the word-level annotation for the same sentence.

Insufficient morphological analysis:

The verb *màà* receives no morphological analysis in (2).
Although tone plays an important role in the expression of verbal inflection in Akan, no attempt is made, except in (2), to render tone in the glossing.
Due to lack of word internal analysis, we miss the fact that the verb initial *re*- in (5) is the progressive marker.
The general lack of part of speech information makes it impossible to determine the grammatical category of the word *na* in example (5).
In (4) the gloss 3sgP is ambiguous between 3 singular personal pronoun and 3 singular possessive Pronoun.
In this case the gloss refers to the latter and denotes the pre-nominal possessive pronoun which is co-referential with the subject.
The meaning of the phrase is close to: “ *that one of his horses*” due to the noun phrase final *nó*

NTNU
Innovation and Creativity

These few examples illustrate typical ways in which interlinear glosses can fall short of being informative or even valid.

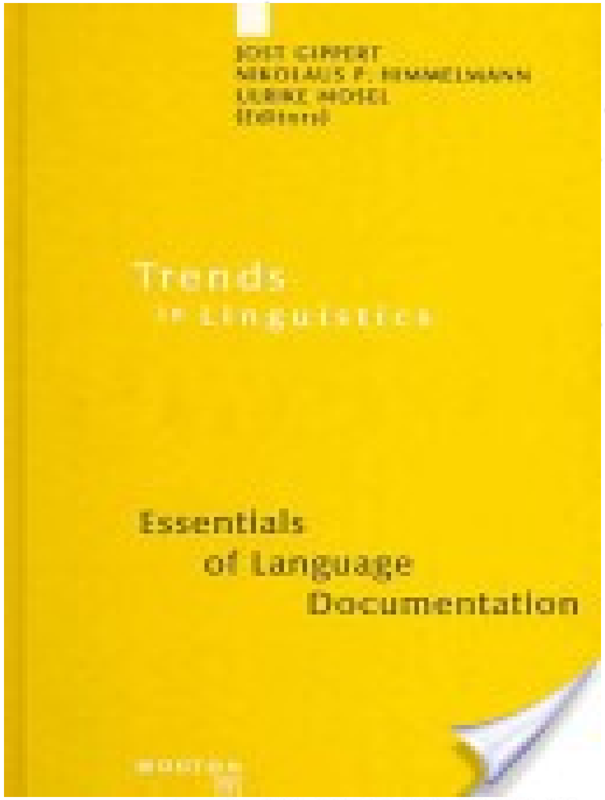
Yet, published IGTs, in particular in the literature about less-resourced languages, are often the only structured data available for that language or that phenomenon.

As we already pointed out, to speak about 'linguistic data' is an abstraction given the role that it plays for different linguistic approaches. Yet, no matter what their function is within linguistic research, they must be informative and accurate

What does it really mean for annotated material to be accurate, and how much accuracy can we expect?

Gippert, Himmelmann and Mosel's book:

Essentials of Language Documentation



contains several articles on this topic

Incremental annotation

Mosel and Schultze-Berndt in particular address questions relating to the creation of annotations as part of a linguistic discovery process.

Mosel points out that IGT is as much the result of linguistic research as it serves as its input, annotated data only reflects a current stage of knowledge and therefore might be more underspecified than one wishes, or even might be ambiguous and incomplete.

Annotating - a discovery process

"That there is a trade-off between the amount of information and the time spent on annotation" is pointed out by Schultze-Bernd who also states that annotations can be improved by subsequent research given that the raw data is equally available as the IGTs themselves.

Linguists are rediscovering their methodology, in the process they explore new media and new routines for data management and begin to set standards for linguistic tools and resources alike.

Data-oriented linguistics

What we do and what we need

Incremental annotation

Exploratory research

Annotation as an integral part of linguistic research

Linguistic tools for experts

but not necessary for

tool-experts

distributed and linked resources

NTNU
Innovation and Creativity

Linguistic Annotations and how linguists use them

I would like to thank my colleagues:

Aimée Lahaussois from the Lacito at the Sorbonne and

Matt Coler from INCAS3,

as well as their co-workers for

allowing me to present parts of their work.

In this presentation, I can only show a small aspect of Lahaussois' and Coler's work for a more representative view see
Lahaussois: http://lacito.vjf.cnrs.fr/membres/lahaussois.htm.
Coler:http://www.incas3.eu/people/mattcoler

First case:

A comparable corpus for Kiranti

Aimée Lahaussois (Lacito, CNRS, Paris) and
Séverine Guillaume (CNRS)

Comparable corpus, "which selects similar texts in more than one language or variety, [with] as yet no agreement on the nature of the similarity. [...] The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus."

(Sinclair, 1996—EAGLES: " Preliminary recommendations on Corpus Typology")

Source data for Kiranti (spoken in East Nepal)

"Kakcilip story"
Thulung (12 minutes)
Khaling (13 minutes)
Koyi (63 minutes—contains more than just the Kakcilip story)
Data is interlinearised: transcription,
gloss and translation tiers, with sound synchronization;

This is work by Aimée Lahaussois - Séverine Guillaume

NTNU
Innovation and Creativity

shared construction(s)

[THU]

uʦi-walwak-**ka**	ʔe	ʥau-nuŋ	kʰleu-nuŋ-**ka**	ʦʰəhi	səlla	bet-**ʦi**
3DU.POSS-sibling-ERG	HS	Jau-COM	Khleu-COM-ERG	CONTR	advice	do-3DU>3SG.PST

'Jau and Khleu came to a decision.'

[KHA]

tunɵ̂l	didi	bahini	grɵ̂mmɛ	lasmɛ-su-**ʔɛ**	mɵ̂l	mʉ-**ssu**
one.day	older.sister	younger.sister	Gromme	Lasme-DU-ERG	counsel	do-3DU>3SG.PST

'One day, Gromme and Lasme had a discussion.'

"advice/counsel + to do"=to come to a decision

ergative marking on agent, agreement marker 3DU>3SG.PST

This is work by Aimée Lahaussois - Séverine Guillaume

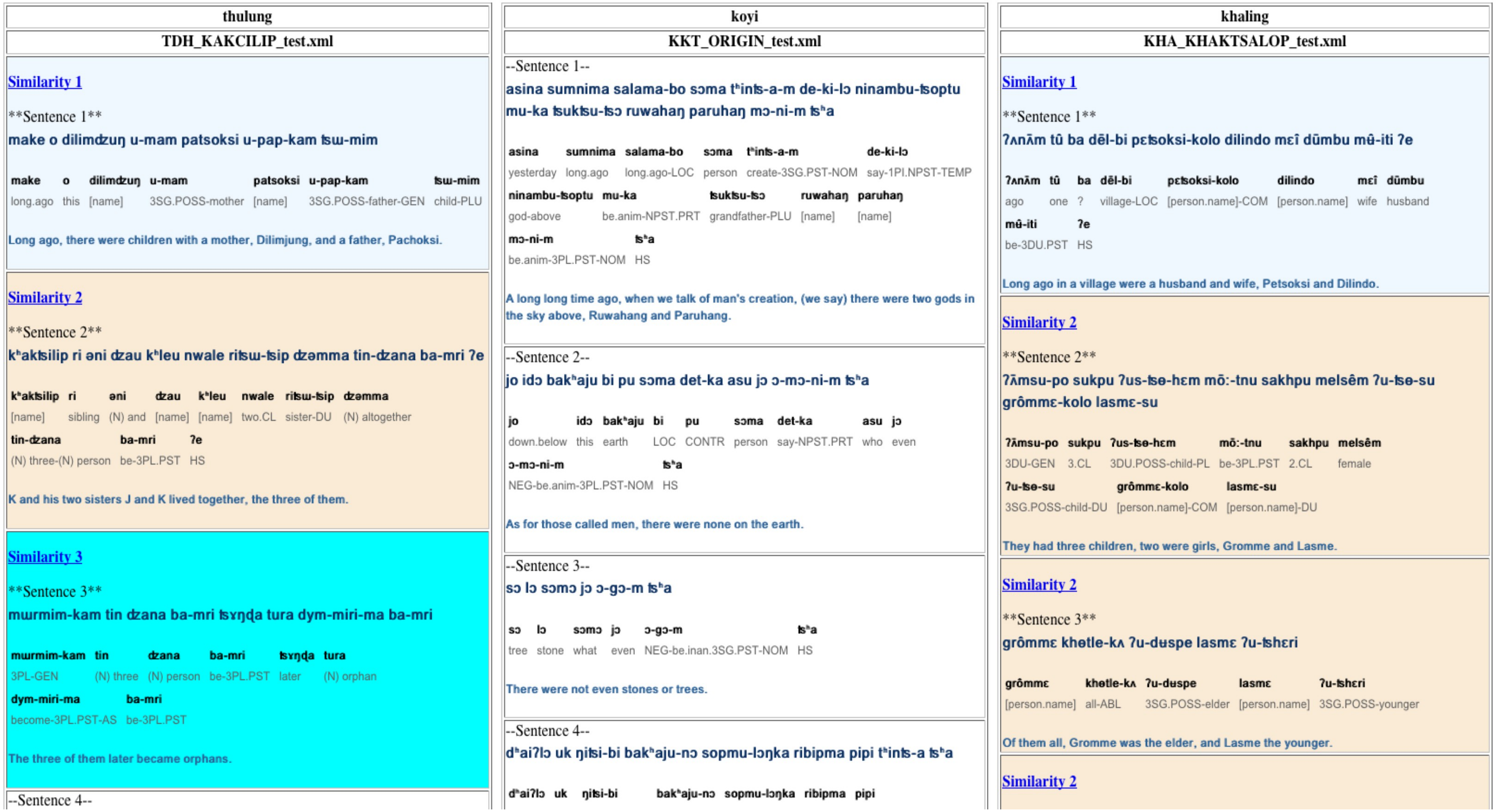
Example of piece of alignment file:

This is work by Aimée Lahaussois - Séverine Guillaume

```
<similarities>
        <files>
             <file xml="TDH_KAKCILIP_test.xml" lang="thulung" sound="../audio/Kakcilip.wav"/>
             <file xml="KKT_ORIGIN_test.xml" lang="koyi" sound="../audio/Origin.wav"/>
             <file xml="KHA_KHAKTSALOP_test.xml" lang="khaling" sound="../audio/Khaktsalop.wav"/>
        </files>
        <similarity id="1">
             <color>aliceblue</color>
             <file id="TDH_KAKCILIP_test.xml">
                  <sentence id="s1"/>
             </file>
             <file id="KHA_KHAKTSALOP_test.xml">
                  <sentence id="s1"/>
             </file>
        </similarity>
        <similarity id="2">
             <color>antiquewhite</color>
             <file id="TDH_KAKCILIP_test.xml">
                  <sentence id="s2"/>
             </file>
             <file id="KKT_ORIGIN_test.xml">
                  <sentence id="s191"/>
             </file>
             <file id="KHA_KHAKTSALOP_test.xml">
                  <sentence id="s2"/>
                  <sentence id="s3"/>
                  <sentence id="s4"/>
             </file>
        </similarity>
 </similarities>
```

This is work by Aimée Lahaussois - Séverine Guillaume

Integral text view



thulung

TDH_KAKCILIP_test.xml

Similarity 1

Sentence 1

make o dilimʤuŋ u-mam patsoksi u-pap-kam ʦɯ-mim

make	o	dilimʤuŋ	u-mam	patsoksi	u-pap-kam	ʦɯ-mim
long.ago	this	[name]	3SG.POSS-mother	[name]	3SG.POSS-father-GEN	child-PLU

Long ago, there were children with a mother, Dilimjung, and a father, Pachoksi.

Similarity 2

Sentence 2

kʰakʦilip ri əni ʤau kʰleu nwale riʦɯ-ʦip ʤəmma tin-ʤana ba-mri ʔe

kʰakʦilip	ri	əni	ʤau	kʰleu	nwale	riʦɯ-ʦip	ʤəmma	tin-ʤana	ba-mri	ʔe
[name]	sibling	(N) and	[name]	[name]	two.CL	sister-DU	(N) altogether	(N) three-(N) person	be-3PL.PST	HS

K and his two sisters J and K lived together, the three of them.

Similarity 3

Sentence 3

mɯrmim-kam tin ʤana ba-mri ʦɤŋɖa tura dym-miri-ma ba-mri

mɯrmim-kam	tin	ʤana	ba-mri	ʦɤŋɖa	tura	dym-miri-ma	ba-mri
3PL-GEN	(N) three	(N) person	be-3PL.PST	later	(N) orphan	become-3PL.PST-AS	be-3PL.PST

The three of them later became orphans.

--Sentence 4--

koyi

KKT_ORIGIN_test.xml

--Sentence 1--

asina sumnima salama-bo sɔma tʰinʦ-a-m de-ki-lɔ ninambu-ʦoptu mu-ka ʦukʦu-ʦɔ ruwahaŋ paruhaŋ mɔ-ni-m ʦʰa

asina	sumnima	salama-bo	sɔma	tʰinʦ-a-m	de-ki-lɔ	ninambu-ʦoptu	mu-ka	ʦukʦu-ʦɔ	ruwahaŋ	paruhaŋ	mɔ-ni-m	ʦʰa
yesterday	long.ago	long.ago-LOC	person	create-3SG.PST-NOM	say-1PI.NPST-TEMP	god-above	be.anim-NPST.PRT	grandfather-PLU	[name]	[name]	be.anim-3PL.PST-NOM	HS

A long long time ago, when we talk of man's creation, (we say) there were two gods in the sky above, Ruwahang and Paruhang.

--Sentence 2--

jo idɔ bakʰaju bi pu sɔma det-ka asu jɔ ɔ-mɔ-ni-m ʦʰa

jo	idɔ	bakʰaju	bi	pu	sɔma	det-ka	asu	jɔ	ɔ-mɔ-ni-m	ʦʰa
down.below	this	earth	LOC	CONTR	person	say-NPST.PRT	who	even	NEG-be.anim-3PL.PST-NOM	HS

As for those called men, there were none on the earth.

--Sentence 3--

sɔ lɔ sɔmɔ jɔ ɔ-gɔ-m ʦʰa

sɔ	lɔ	sɔmɔ	jɔ	ɔ-gɔ-m	ʦʰa
tree	stone	what	even	NEG-be.inan.3SG.PST-NOM	HS

There were not even stones or trees.

--Sentence 4--

dʰaiʔlɔ uk ŋiʦi-bi bakʰaju-nɔ sopmu-lɔŋka ribipma pipi tʰinʦ-a ʦʰa

dʰaiʔlɔ uk ŋiʦi-bi bakʰaju-nɔ sopmu-lɔŋka ribipma pipi

khaling

KHA_KHAKTSALOP_test.xml

Similarity 1

Sentence 1

ʔʌnʌ̄m tû ba dēl-bi pɛʦoksi-kolo dilindo mɛî dūmbu mɵ̂-iti ʔe

ʔʌnʌ̄m	tû	ba	dēl-bi	pɛʦoksi-kolo	dilindo	mɛî	dūmbu	mɵ̂-iti	ʔe
ago	one	?	village-LOC	[person.name]-COM	[person.name]	wife	husband	be-3DU.PST	HS

Long ago in a village were a husband and wife, Petsoksi and Dilindo.

Similarity 2

Sentence 2

ʔʌ̄msu-po sukpu ʔus-ʦɵ-hɛm mō:-tnu sakhpu melsêm ʔu-ʦɵ-su grômmɛ-kolo lasmɛ-su

ʔʌ̄msu-po	sukpu	ʔus-ʦɵ-hɛm	mō:-tnu	sakhpu	melsêm	ʔu-ʦɵ-su	grômmɛ-kolo	lasmɛ-su
3DU-GEN	3.CL	3DU.POSS-child-PL	be-3PL.PST	2.CL	female	3SG.POSS-child-DU	[person.name]-COM	[person.name]-DU

They had three children, two were girls, Gromme and Lasme.

Similarity 2

Sentence 3

grômmɛ khɵtle-kʌ ʔu-dʉspe lasmɛ ʔu-ʦhɛri

grômmɛ	khɵtle-kʌ	ʔu-dʉspe	lasmɛ	ʔu-ʦhɛri
[person.name]	all-ABL	3SG.POSS-elder	[person.name]	3SG.POSS-younger

Of them all, Gromme was the elder, and Lasme the younger.

Similarity 2

Lahaussois starts from primary data which she aquires through
field work in Nepal.
She works from audio sources which she transcribes and glosses (Her lanuages are oral languages only.).

On top of the morpheme level annotations, she adds another layer of annotations for comparing similarities between her languages.

Construction level as well as narrative and lexical similarities can be compared.

She and he colleague developed a representation of these simlarities which allows linguists to search for different types of similarities and to compare them easily using a graphical user interface.

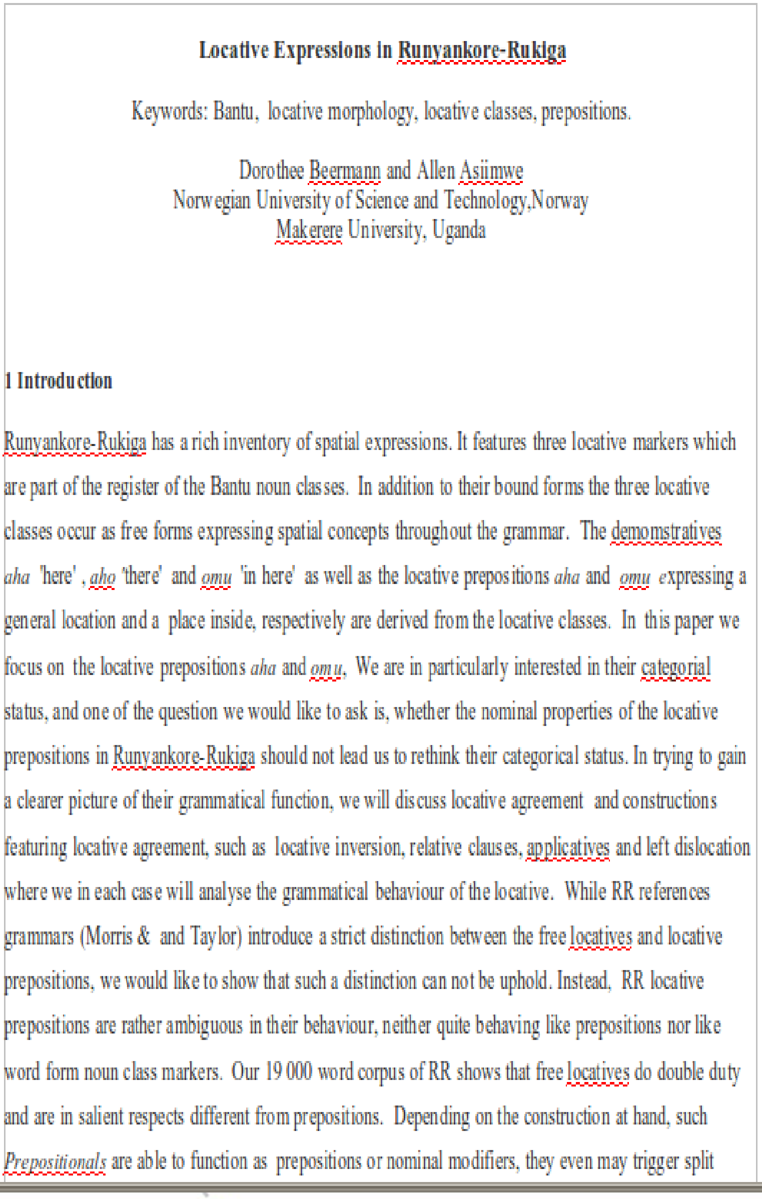
Second Case

Locative Expressions in Runyankore-Rukiga (RR)

Keywords: Bantu, locative morphology, locative classes, prepositions.

Dorothee Beermann and Allen Asiimwe
Norwegian University of Science and Technology,Norway
Makerere University, Uganda

Publishing research results + create reusable research data



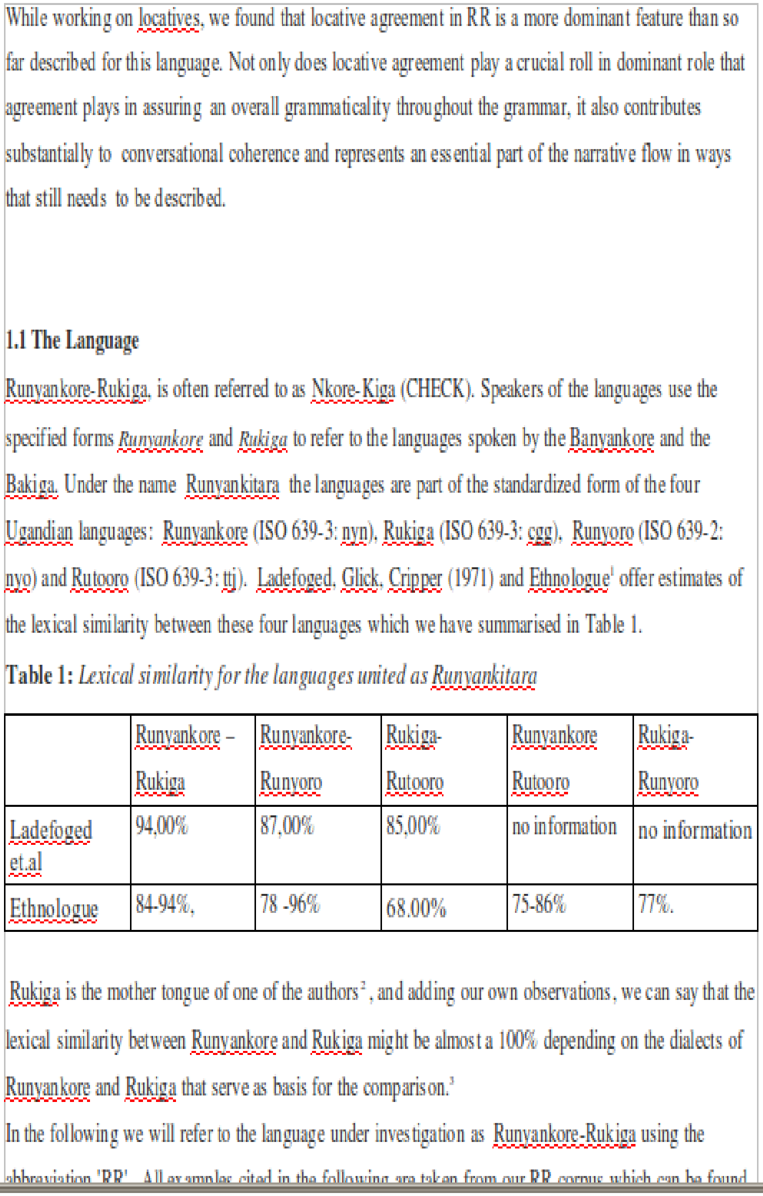
Locative Expressions in Runyankore-Rukiga

Keywords: Bantu, locative morphology, locative classes, prepositions.

Dorothee Beermann and Allen Asiimwe
Norwegian University of Science and Technology,Norway
Makerere University, Uganda

1 Introduction

Runyankore-Rukiga has a rich inventory of spatial expressions. It features three locative markers which are part of the register of the Bantu noun classes. In addition to their bound forms the three locative classes occur as free forms expressing spatial concepts throughout the grammar. The demomstratives *aha* 'here' , *aho* 'there' and *omu* 'in here' as well as the locative prepositions *aha* and *omu* expressing a general location and a place inside, respectively are derived from the locative classes. In this paper we focus on the locative prepositions *aha* and *omu*. We are in particularly interested in their categorial status, and one of the question we would like to ask is, whether the nominal properties of the locative prepositions in Runyankore-Rukiga should not lead us to rethink their categorical status. In trying to gain a clearer picture of their grammatical function, we will discuss locative agreement and constructions featuring locative agreement, such as locative inversion, relative clauses, applicatives and left dislocation where we in each case will analyse the grammatical behaviour of the locative. While RR references grammars (Morris & and Taylor) introduce a strict distinction between the free locatives and locative prepositions, we would like to show that such a distinction can not be uphold. Instead, RR locative prepositions are rather ambiguous in their behaviour, neither quite behaving like prepositions nor like word form noun class markers. Our 19 000 word corpus of RR shows that free locatives do double duty and are in salient respects different from prepositions. Depending on the construction at hand, such *Prepositionals* are able to function as prepositions or nominal modifiers, they even may trigger split



While working on locatives, we found that locative agreement in RR is a more dominant feature than so far described for this language. Not only does locative agreement play a crucial roll in dominant role that agreement plays in assuring an overall grammaticality throughout the grammar, it also contributes substantially to conversational coherence and represents an essential part of the narrative flow in ways that still needs to be described.

1.1 The Language

Runyankore-Rukiga, is often referred to as Nkore-Kiga (CHECK). Speakers of the languages use the specified forms *Runyankore* and *Rukiga* to refer to the languages spoken by the Banyankore and the Bakiga. Under the name Runyankitara the languages are part of the standardized form of the four Ugandian languages: Runyankore (ISO 639-3: nyn), Rukiga (ISO 639-3: cgg), Runyoro (ISO 639-2: nyo) and Rutooro (ISO 639-3: ttj). Ladefoged, Glick, Cripper (1971) and Ethnologue[1] offer estimates of the lexical similarity between these four languages which we have summarised in Table 1.

Table 1: *Lexical similarity for the languages united as Runyankitara*

	Runyankore – Rukiga	Runyankore-Runyoro	Rukiga-Rutooro	Runyankore Rutooro	Rukiga-Runyoro
Ladefoged et.al	94,00%	87,00%	85,00%	no information	no information
Ethnologue	84-94%,	78 -96%	68.00%	75-86%	77%.

Rukiga is the mother tongue of one of the authors[2], and adding our own observations, we can say that the lexical similarity between Runyankore and Rukiga might be almost a 100% depending on the dialects of Runyankore and Rukiga that serve as basis for the comparison.[3]

In the following we will refer to the language under investigation as Runyankore-Rukiga using the abbreviation 'RR'. All examples cited in the following are taken from our RR corpus which can be found

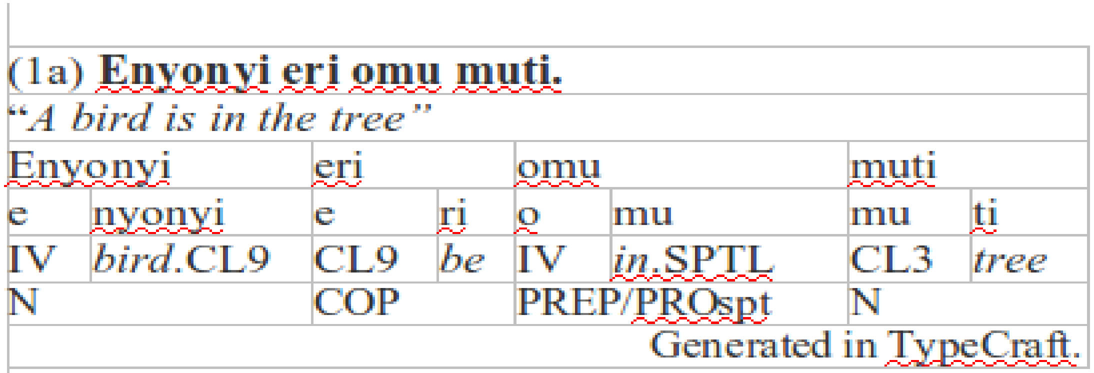
Publishing research results + create reusable research data

All examples cited in the following are taken from
our RR corpus which can be found
in the Open Access online multilingual database TypeCraft.
The TypeCraft (TC) database is augmented by an online
linguistic editor which we used for the annotation of our
data. For the purpose of this publication the Interlinear Glossed Text is
reduced to a four tier format.
For a more in-depth view of example sentences , the article refers the reader
to the TC database.

Using the TypeCraft database we have built a corpus of 19 602
words, corresponding to 54 574 annotated morphemes. An
article of standard length might include 20 may be 30 examples.
TC contains 4260 examples from RR which can be inspected when
evaluating the work presented in the article.

NTNU
Innovation and Creativity

Novel is that research and research results are presented as linked resources, using the possibilities that Open Access databasing offers.
The reader can inspect the complete dataset
(4260 tokens instead of 30) that our publication is based on.
The data is free and can be used for further research.



(1a) **Enyonyi eri omu muti.**							
"A bird is in the tree"							
Enyonyi		eri		omu		muti	
e	nyonyi	e	ri	o	mu	mu	ti
IV	*bird*.CL9	CL9	*be*	IV	*in*.SPTL	CL3	*tree*
N		COP		PREP/PROspt		N	
Generated in TypeCraft.							

We started from primary data which we generated through collaborative work between native-speaker linguists and linguists at the supporting University, NTNU.

We work with audio and text material which we transcribed and annotated.

Departing from our analysis of the primary material, we confront our findings with those reported in prior research.

One of our goals is to publish linked ressources so that our research and the data that supported it become available to the general public. This not only makes our work easily accessible for peer-review, it also helps to create re-usable linguistic ressources in the form of IGTs.

Third Case

Machine Translation for Aymara

Matt Coler (INCAS3)
Peter Homola (Codesign)

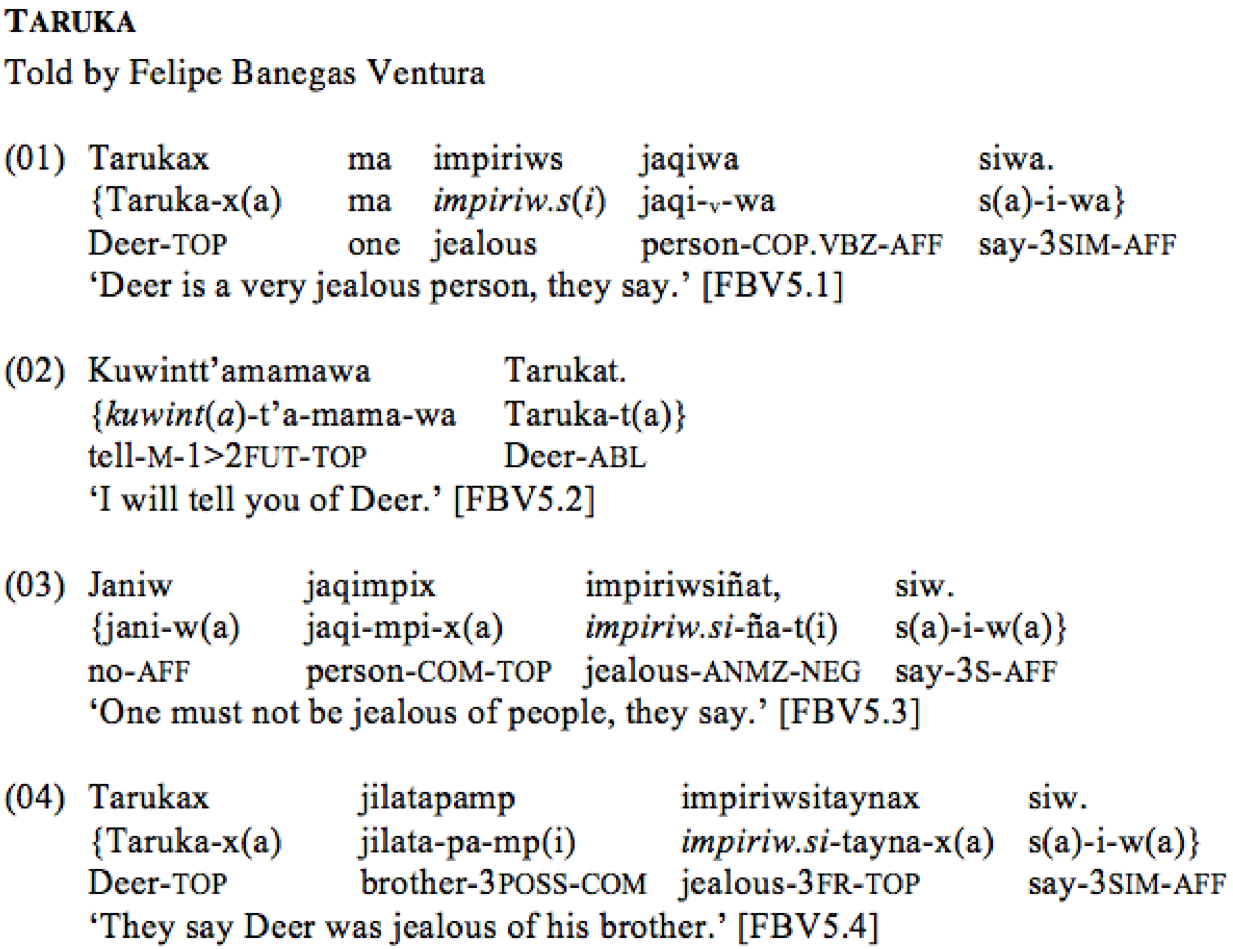
Aymara, language of Bolivia

ISO 639-3: ayr

Population: 1,790,000 in Bolivia (1987).
Population total all countries: 2,262,900.

This is work by Matt Coler (INCAS3)

and Peter Homola (Codesign)



TARUKA

Told by Felipe Banegas Ventura

(01)

Tarukax	ma	impiriws	jaqiwa	siwa.
{Taruka-x(a)	ma	*impiriw.s*(*i*)	jaqi-$_{V}$-wa	s(a)-i-wa}
Deer-TOP	one	jealous	person-COP.VBZ-AFF	say-3SIM-AFF

'Deer is a very jealous person, they say.' [FBV5.1]

(02)

Kuwintt'amamawa	Tarukat.
{*kuwint*(*a*)-t'a-mama-wa	Taruka-t(a)}
tell-M-1>2FUT-TOP	Deer-ABL

'I will tell you of Deer.' [FBV5.2]

(03)

Janiw	jaqimpix	impiriwsiñat,	siw.
{jani-w(a)	jaqi-mpi-x(a)	*impiriw.si*-ña-t(i)	s(a)-i-w(a)}
no-AFF	person-COM-TOP	jealous-ANMZ-NEG	say-3S-AFF

'One must not be jealous of people, they say.' [FBV5.3]

(04)

Tarukax	jilatapamp	impiriwsitaynax	siw.
{Taruka-x(a)	jilata-pa-mp(i)	*impiriw.si*-tayna-x(a)	s(a)-i-w(a)}
Deer-TOP	brother-3POSS-COM	jealous-3FR-TOP	say-3SIM-AFF

'They say Deer was jealous of his brother.' [FBV5.4]

Agglutinative, suffix only, rich morphology"

Aside from unmarked subj all syntactic relations are case marked typically on NP head

SOV; mod-head

NTNU

Innovation and Creativity

Juma-n-x	*jiw-i-w*	*kimsa*	*ch'iyar*	*phisi-ma-xa*
you-GEN-TOP	die-PAST3-FOC	three	black	cat-POSS2-TOP

"Your three black cats died."

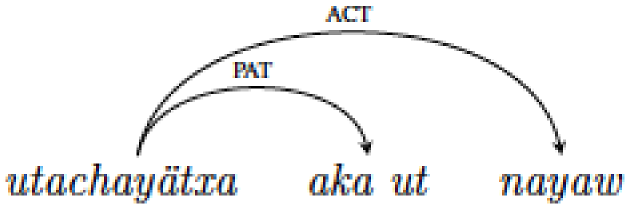
$$\begin{bmatrix} \text{PRED} & \text{`jiwa}\langle(\uparrow\text{SUBJ})\rangle\text{'} \\ \text{TENSE} & \textsc{past} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`phisi'} \\ \text{POSS} & \begin{bmatrix}\text{``jumanx''}\end{bmatrix} \\ \text{ADJ} & \left\{ \begin{bmatrix}\text{``kimsa''}\end{bmatrix}, \begin{bmatrix}\text{``ch'iyar''}\end{bmatrix} \right\} \end{bmatrix} \end{bmatrix}$$

This is work by Matt Coler (INCAS3)
and Peter Homola (Codesign)

Naya-w	*aka*	*ut*	*utacha-yä-t-xa.*
I-FOC	this	house	build-PAST-1→3-TOP

"This house was built by me. (It is me who built this house.)"

$$\begin{bmatrix} \text{PRED} & \text{`build}\langle(\uparrow\text{SUBJ})(\uparrow\text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PERF} \\ \text{SUBJ} & \begin{bmatrix}\text{PRED} & \text{`I'}\end{bmatrix} \\ \text{OBJ} & \begin{bmatrix}\text{PRED} & \text{`house'} \\ \text{SPEC} & \begin{bmatrix}\text{PRED} & \text{`this'}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$



This is work by Matt Coler (INCAS3) and Peter Homola (Codesign)

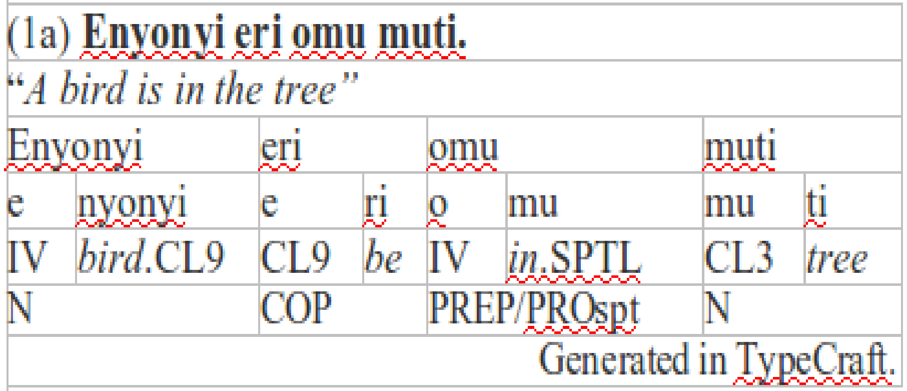
Coler worked from primary data which he acquired
during field work in Bolivian.
He annotated his data creating Interlinear Glossed Text.

Using language technology developed by the
Lexical-Functional Grammar community of linguists,
he created on top of interlinear glossed primary
date new layers of syntactic and dependency annotations.
This allows him to parse his language.
His goal his to allow machine translation also for languages
that so far are under-resourced.

NTNU
Innovation and Creativity

Interlinear Glosses

google pictures

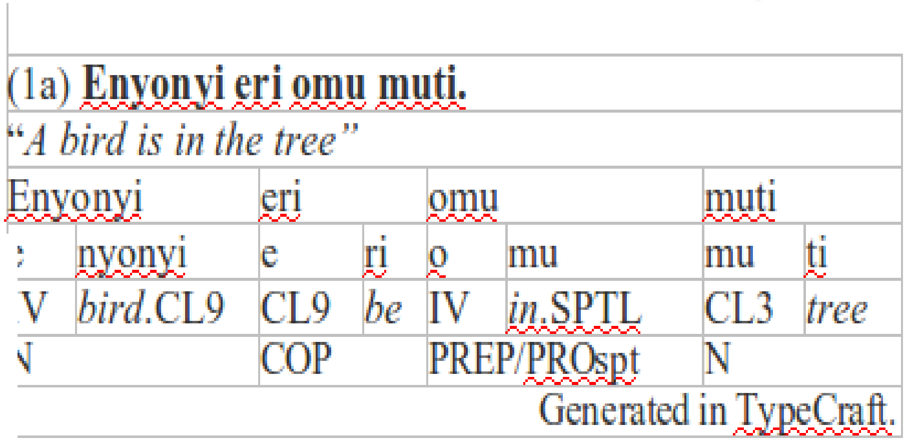


(1a) **Enyonyi eri omu muti.**							
"A bird is in the tree"							
Enyonyi		eri		omu		muti	
e	nyonyi	e	ri	o	mu	mu	ti
IV	*bird*.CL9	CL9	*be*	IV	*in*.SPTL	CL3	*tree*
N		COP		PREP/PROspt		N	
Generated in TypeCraft.							

...describe something we otherwise would not see

NTNU
Innovation and Creativity

Interlinear Glosses

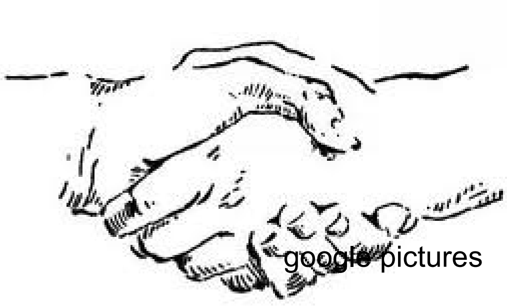


(1a) Enyonyi eri omu muti.

"A bird is in the tree"

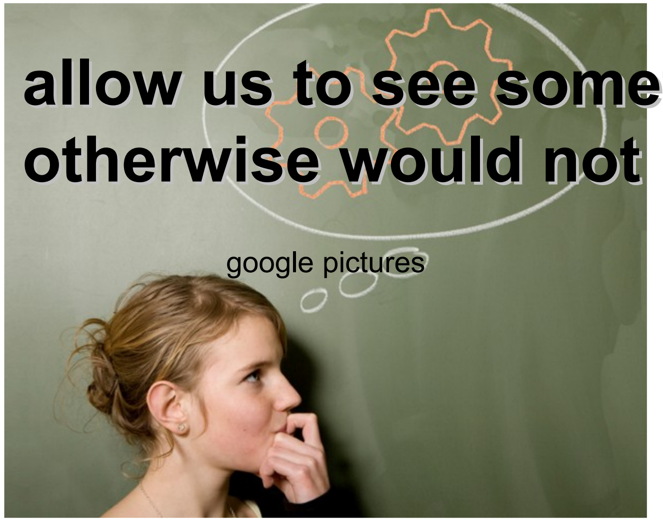
Enyonyi		eri		omu		muti	
e	nyonyi	e	ri	o	mu	mu	ti
V	*bird*.CL9	CL9	*be*	IV	*in*.SPTL	CL3	*tree*
N		COP		PREP/PROspt		N	

Generated in TypeCraft.



google pictures

...and share it



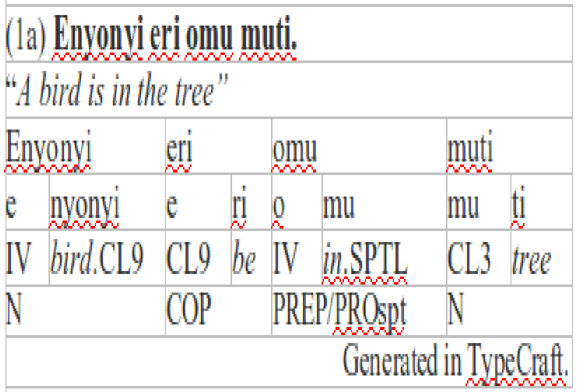
allow us to see something we otherwise would not have recognised

google pictures

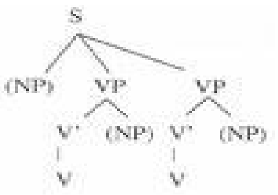
Annotations allow us to

```
<phrase valid="VALID" id="28">
    <original>Enyonyi eri omu muti</original>
    <translation>A bird is in the tree</translation>
    <description>Locative deixis</description>
    <globaltags tagset="Default" id="1"/>
    <word head="false" text="ènyònyì">
        <pos>N</pos>
        <morpheme baseform="" text="e">
            <gloss>IV</gloss>
        </morpheme>
        <morpheme meaning="" baseform="" text="n">
            <gloss>CL9</gloss>
        </morpheme>
    <word head="false" text="eri">
        <pos>V</pos>
        <morpheme baseform="" text="e">
            <gloss>CL9</gloss>
            <gloss>SM</gloss>
        </morpheme>
        <morpheme meaning="be" baseform="okuba" text="ri"/>
    </word>
    <word head="false" text="òmù">
        <pos>PREP</pos>
        <morpheme meaning="" baseform="" text="o">
            <gloss>IV</gloss>

                </morpheme>
        <morpheme meaning="in" baseform="omu" text="mu"/>
    </word>
    <word head="false" text="mùtì">
        <pos>N</pos>
        <morpheme baseform="" text="mu">
            <gloss>CL3</gloss>
        </morpheme>
        <morpheme meaning="tree" baseform="omuti" text="ti"/>
    </word>
  </phrase>
</typecraft>
```



(1a) Enyonyi eri omu muti.							
"A bird is in the tree"							
Enyonyi		eri		omu		muti	
e	nyonyi	e	ri	o	mu	mu	ti
IV	bird.CL9	CL9	be	IV	in.SPTL	CL3	tree
N		COP		PREP/PROspt		N	
Generated in TypeCraft.							



add efficiency

$$\left[\begin{array}{ll} \text{PRED} & \text{'jiwa}\langle(\uparrow\text{SUBJ})\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{'phisi'} \\ \text{POSS} & \left[\text{"jumanx"}\right] \\ \text{ADJ} & \left\{\left[\text{"kimsa"}\right], \left[\text{"ch'iyar"}\right]\right\} \end{array}\right] \end{array}\right]$$

References

Beermann, Dorothee 2012. ‘Collaborative online resource building for less-resourced languages. Language Endangerment: Methodologies and New Challenges. CRASSH, Cambridge University.

Beermann, Dorothee and Pavel Mihaylov. 2012.TypeCraft Collaborative databasing and Resource sharing for Linguists. In Proceedings of the 9th Extended Semantic Web Conference, May 27th - 31st, 2012. Workshop, Interacting with Linked Data. (to appear in CEUR Workshop Proceedings).

Beermann, Dorothee and Allen Asiimwe. Locative Expressions in Runyankore-Rukiga. (to appear).

Cheng, Lisa and Laura J. Downing. 2010. Locative Relatives in Durban Zulu, ZAS Papers in Linguistics 53, pp 33-5.

Coler, Matt and Peter Homola. 2012. Aymara - English machine translation using dependency representation. Language Endangerment Methodologies and New Challenges. CRASSH, Cambridge University.

Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel 2006. *Essentials of Language Documentation. Mouton de Gruyter*

Lahaussois, Aimée. 2012. The Kiranti comparable corpus. Language Endangerment: Methodologies and New Challenges. CRASSH, Cambridge University.

Haspelmath, Martin. (ed) 2004. Coordinating constructions. (Typological Studies in Language, 58.) Benjamins, Amsterdam

Holten, Gary. 2003. Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive. manuscript

Rossi, Giovanni 2012. Bilateral and Unilateral Requests: The Use of Imperatives and Mi X? Interrogatives in Italian. Discourse Processes, Volume 49, Issue 5.

Schultze-Berndt, Eva. "Linguistic Annotation." In Essentials of Language Documentation, ed. Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel, 213-251. Berlin: Mouton de Gruyter, 2006.

Online resources

Online Etymology Dictionary, 2001-2012 Douglas Harper: http://www.etymonline.com/

ANNIS2, Search and Visualization in Multilevel Linguistic Corpora:http://www.sfb632.uni-potsdam.de/d1/annis/

ODIN - The Online Database of Interlinear Text: http://odin.linguistlist.org/