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.)).


King James 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 ad hoc basis.

Interlinear Glossed Text

12 nd 26 a (26 Jan. 1936)
Sun. evening
qíla-m tciyá-na-m ñyan umu-tci-ilix qamuyän-in
morning-of creek-place-of because sit-firm-ing beach geese-s
Because early in the morning beach geese landed at the mouth of
tana-ná-nin tum'öi-ilix qamuyän-ix. qama-tci-ni-ki-qin
drink-much-they shoot at-ing beach geese-s. step-careful-reluctant-now-I
across the creek to drink, I kept shooting at them. Carefully I stepped

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
(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”
Linguistic Typology

(1) Lavukaleve (Terrill°, ex. 9)
nga-bakala nga-ua 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.
Amsterdam: Benjamins.
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.

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.
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.
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.
4 IGTS extracted from ODIN*

(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).\textsuperscript{13}

(4) c-emme me ne pønkó 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 rehwehwe?
    who FOC Ama is-looking-for?
    ‘Who is it that Ama is looking for?’

* ODIN - The Online Database of Interlinear Text

\(\text{http://odin.linguistlist.org/}\)
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.
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
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
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
shared construction(s)

[THU]

utsi-walwak-ka  dzau-nuŋ  kʰleu-nuŋ-ka  tshəhi  səlla  bet-tsi
ʔe
3DU.POSS-sibling-ERG  Jau-COM  Khleu-COM-ERG  CONTR  advice  do-3DU>3SG.PST
HS

'Jau and Khleu came to a decision.'

[KHA]

tunɵ̂l  didi  bahini  grômme  lasmɛ-su-ʔɛ  mêl  mu-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:

```
<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>
```

This is work by Aimée Lahaussois - Séverine Guillaume
The Kiranti comparable corpus
Aimée Lahaussois - Séverine Guillaume
Lahaussois starts from primary data which she acquires through field work in Nepal. She works from audio sources which she transcribes and glosses (Her languages 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 her colleague developed a representation of these similarities 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
Locative Expressions in Runyankore-Rukiga

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

Dorothee Bjoernbye 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 demonstratives aba ‘bora’, asa ‘there’ and anya ‘there in’ as well as the locative prepositions aba and anya expressing a general location and a place inside, respectively, are derived from the locative classes. In this paper, we focus on the locative prepositions aba and anya. We are in particular interested in their categorial status and one of the questions we would like to ask is, whether the nominal properties of the locative prepositions in Runyankore-Rukiga should not lead us to rethink their categorial 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 & Taylor) introduce a strict distinction between the free locatives and locative prepositions, we would like to show that such a distinction cannot be upheld. Instead, RR locative prepositions are rather ambiguous in their behaviour, neither quite behaving like prepositions nor like word form noun class markers. Our 10,000 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 prepositions 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 role in dominating 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 need to be described.

1.1 The Language

Runyankore-Rukiga is often referred to as Nkoro-Kiga (CHECK). Speakers of the languages use the specified forms runyankore and rukiga to refer to the languages spoken by the Runyankore and the Rukiga. Under the name Runyankore, the languages are part of the standardized form of the four Ugandan languages: Runyankore (ISO 639-3: run), Rukiga (ISO 639-3: rku), Runoro (ISO 639-2: run) and Rutooro (ISO 639-3: rto). Ladebeg and Geijer (1971) and Ethnologue 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 Runyankikiga

<table>
<thead>
<tr>
<th></th>
<th>Runyankore - Rukiga</th>
<th>Runyankore - Runoro</th>
<th>Rukiga - Rutooro</th>
<th>Runyankore - Rutooro</th>
<th>Rukiga - Runoro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ladebeg et al.</td>
<td>94.0%</td>
<td>87.0%</td>
<td>85.0%</td>
<td>no information</td>
<td>no information</td>
</tr>
<tr>
<td>Ethnologue</td>
<td>84.94%</td>
<td>78.86%</td>
<td>68.00%</td>
<td>75.86%</td>
<td>77.0%</td>
</tr>
</tbody>
</table>

Rukiga is the mother tongue of one of the authors¹, 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.²

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 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.
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.
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 resources 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 resources 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 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 impiiriws jaqiwa siwa.
{Taruca-x(a) ma *impiiriw*.s(i) jaqi-va 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 impiiriwsiñat, siw.
{jan-i-w(a) jaqim-pi-x(a) *impiiriw*.si-ñat(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 impiiriwsitaynax siw.
{Taruca-x(a) jilat-a-pa-mp(i) *impiiriw*.si-taynax(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]

**SOV; mod-head**

Agglutinative, suffix only, rich morphology"

Aside from unmarked subj all syntactic relations are case marked typically on NP head
Juma-n-x jiw-i-w kimsa ch’iyar phisi-ma-xa
you-GEN-TOP die-PAST3-FOC three black cat-POSS2-TOP
“You your three black cats died.”

This is work by Matt Coler (INCAS3)
and Peter Homola (Codesign)
"This house was built by me. (It is me who built this house.)"

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 is to allow machine translation also for languages that so far are under-resourced.
Interlinear Glosses

...describe something we otherwise would not see

<table>
<thead>
<tr>
<th>Enyonyi</th>
<th>eri</th>
<th>omu</th>
<th>muti</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>nyonyi</td>
<td>eri</td>
<td>omu</td>
</tr>
<tr>
<td>IV</td>
<td>bird.CL9</td>
<td>CL9</td>
<td>be</td>
</tr>
<tr>
<td>N</td>
<td>COP</td>
<td>PREP/PROspt</td>
<td>N</td>
</tr>
</tbody>
</table>

Generated in TypeCraft.
Interlinear Glosses

...and share it

allow us to see something we otherwise would not have recognised
Annotations allow us to add efficiency.
References


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


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


Online resources


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/