A Four-Level Model for Interlinear Text

Catherine Bow, Baden Hughes and Steven Bird

Department of Computer Science and Software Engineering

University of Melbourne, Victoria 3010, Australia

{cbow,badenh,sb}@cs.mu.oz.au

December 22, 2003

Abstract

Interlinear text has long been a valuable device in language documentation and linguistic description. However, the task of creating, editing and publishing interlinear text is an onerous one. Interlinear text is governed by simple rules, yet laborious manual formatting in a word processor is the norm. A handful of specialized software tools facilitate the creation of interlinear text, permitting customizable views and alignment to audio and video. However, word processors and specialized software alike fail to deliver on a key promise of digitization, namely reusability. In order to facilitate reusability, we have developed a general-purpose conceptual model of interlinear text consisting of four levels: text, phrase, word and morph. The details of the model are informed by our analysis of a representative sample of current practice. We have implemented the model using standard XML technologies.

1 Introduction

Interlinear texts serve a variety of purposes in linguistic research, illustrating linguistic features ranging from morphological structure to discourse structure, and documenting oral literature. They feature in descriptive grammars, theoretical analyses, and in critical editions of classical and religious texts. The formatting of these texts follows various high-level conventions: glosses are usually displayed below the words they translate; grammatical glosses are usually written as all-capitals; line-wrapping applies to blocks consisting of words and their glosses; line-wrapping occurs at word and not morpheme boundaries; and so on. Despite these commonalities, it is difficult to find two texts that are exactly alike in their structure and presentation. This variety of interpretation presents a problem for the development of general purpose tools.

Interlinear text typically consists of a row of source text annotated with one or more rows of morphosyntactic analysis and translation. Vertical alignment represents the correspondence between the elements of each row. An example of a line of interlinear text is given in (1).

(1)

Comment	trouvez-vous	mes	lunettes	de	soleil?
how	*find.2PL-you.PL*	*my.PL*	*glass.PL*	*of*	*sun*

What do you think of my sunglasses?

The rows of an interlinear text can represent many different information types: orthographic, phonemic, phonetic and prosodic forms; morphemes and morphosyntactic categories; word-level and phrase-level translations into one or more languages; and any kind of analytical coding or commentary. Commentary is especially diverse, ranging from cultural notes and cross references to other sources of information that aid the interpretation of the text, to queries about a detail of the transcription and notes on what should be checked in the next meeting with a language consultant. The alignment between the various rows may be morph-by-morph, word-by-word, or even phrase-by-phrase. The survey will cover many of these practices, and will carefully probe the distinction between structure and presentation.

This paper proposes a general-purpose model for interlinear text, and shows that it is sufficiently expressive and extensible to represent a broad range of practice. Such a model will permit widespread re-use of interlinear text, and we envisage scenarios in which a user searches a large interlinear text for examples of word usage, a grammatical construction, or a counter-example to a theory, and then easily exports the selected text and adapts its formatting for inclusion in another document such as a grammar, teaching materials, or a research paper.

This paper is structured as follows. Section 2 contains a survey of interlinear text formats. In section 3, we identify the key architectural requirements identified by the survey, and we propose a general-purpose model in section 4. Finally, in section 5, we discuss our conclusions and identify topics for further research.

For clarity in the ensuing discussion we give precise interpretation to some key terminology. An *interlinear text* consists of a sequence of *lines*, where each line consists of two or more *rows*. Each row has a determinate *type* (e.g. IPA transcription, morphological analysis, French gloss). One or more rows are identifiable as *source text*, while the remaining rows are *annotations* of the source text, or annotations of those annotations, etc. If a line of interlinear text is too long to fit on the page or screen, it may be *wrapped* by simultaneously moving material from multiple rows down to the beginning of the next interlinear text line.

2 Survey of Interlinear Text

Several dozen sources were surveyed to collect a broad sample of interlinear text. Some texts were supplied by colleagues, some taken from the internet and some from printed grammars, and full texts were preferred over isolated examples. The majority were text-based, with some linked to audio files, and an attempt was made to cover the major geographic and typological areas. The broad range of samples surveyed attempts to adequately reflect the types of information contained in interlinear texts, as well as to examine how such information is structured.

A typical example of a basic three-line interlinear text is shown in (2). In this sample, one row of phonetic transcription is aligned word-by-word with a gloss, and an unaligned phrasal translation is given on the third line.

(2) Nepali Text (Genetti, 1994:166)

ek	**coTi**	**yoTaa**	**kukur**	**khaanaa-ko**	**khoji-maa**	**baahira**	**nisk-yo.**
one	instance	one	dog	food-GEN	search-LOC	outside	come.out-3smL.PST

Once, one dog went out in search of food.

The words are segmented into morphs using hyphens, the gloss is aligned to the whole word, rather than to individual morphs, and complex glosses are separated by full stops (e.g. *nisk-yo* = come.out-3smL.PST). There are examples of portmanteau morphemes, for example *-yo* is glossed as -3smL.PST, a third person singular masculine low-grade honorific past tense marker.

The phrases are numbered, and a footnote identifies the people who recorded, transcribed and glossed the text.

The Ainu text in (3) is similar to the Nepali text in that each word of the text is aligned to a gloss, the phrases are numbered, though the display is different in that a phrasal translation of the entire text is given below the interlinear text.

(3) Ainu Text (Shibatani, 1990:85)

3.

Tapan	*inuma*	*ran-pes*	*kunne*	*cirikinka*	*enkasike*	*nispa-mut-pe*
such	treasure	cliff	like	rise high	over there	master-wear-thing

otu-santuka	*o-uka-uyru*	*otu-pusa-kur*	*suypa*	*kane*	*asso-kotor*	*mike*
many-hilt	APPL-REC-exist	many-knot-shadow	sway	gold	wall	glitter

kane	*anramasu*	*auwesuye.*
gold	pleasing	interesting

Translation:
(1) My foster brother and foster sister raising me, we lived then. (2) The god-built mountain castle, inside the mountain castle, I was raised. (3) The pile of treasure was heaped like a cliff ...

Unlike the Nepali example, the phrasal translation of the Ainu text is divided into numbered phrases with a one-to-one correspondence to the source text. Thus the structure of the two is the same, while the presentation is different. Like the Nepali example, the words are segmented into morphs using hyphens, though multi-word English glosses are not linked (e.g. *cirikinka* = rise high). Observe that polymorphemic words are sometimes given a monomorphemic gloss (e.g. *ran-pes* = cliff, *asso-kotor* = wall), breaking the simple one-to-one correspondence between morphs and glosses.

The Nivkh example in (4) contains a row of phonetically transcribed text and a row of glosses, then some metadata on the source of the text, a section of notes, followed by a free translation. The alignment between word and gloss is unlike the cases considered above: each word is segmented into morphs using both whitespace and hyphens, and each morph is vertically aligned. Observe that hyphens are also used in the representation of complex glosses (e.g. grow-up). A line break occurs word-internally (for the gloss *brother*).

(4) Nivkh Text (Comrie, 1981:276f)

pʻ	*-at'ik*	*-xe*	*pʻ*	*-nanak*	*-xe*	*pań*	*-d'.*	*atík*
REFL	younger-brother	AND	REFL	elder-sister	AND	grow-up	FIN	younger-

	matka	*-d'.*	*kʻu*	*-ɣe*	*puńd*	*-ɣe*	*bo*	*-ror*	*pʻu*	*-r*
brother	be-small	FIN	arrow	AND	bow	AND	take	GER-3SG	go-out	GER-3SG

tevrq	*χa*	*-d'*	*iɣ*	*-ror*	...
bird	shout	FIN	kill	GER-3SG	...

(Extracted from V. Z. Panfilov, *Grammatika nivxskogo jazyka*, vol. 2. Moscow-Leningrad, 1965.)

Notes

... *hadoxq'aud'*: the combination of the negative auxiliary *q'au-* with the infinitive of the main verb is one common way of expressing negation

FREE TRANSLATION

A younger brother and an elder sister grew up. The brother was small. He took his arrows and bow, went out, and shot birds. When he killed them, he brought them and the elder sister plucked the birds' feathers...

In this example, the words are not represented directly; they must be reconstructed by concatenation. The notes section highlights interesting elements of the text, and may link to a morph, word, phrase, or possibly the entire text, however no formal distinction is made between comments at any of these levels.

A passage from a highly agglutinative language, Tundra Nenets, is given two different interlinear treatments. The Susoi version, shown in (5), has three levels, with words aligned to glosses, and a separate phrasal translation of the entire text. However, while the glosses are segmented into morphs using full stops, the words themselves are not, making it impossible to identify which portion of the word corresponds to the morphemic gloss.

(5) Tundra Nenets Text (Susoi, 1990)

Nyew˚xi˚	nyenecvøyeq	syoq.	
ancient.ABS.NOM.SG	*person.ABS.GEN.PL*	*song.ABS.NOM.PL*	
Xurkaryi	lax˚naku,	yarøbc˚	...
what kind.LIM.ABS.NOM.SG	*tale.ABS.ACC.PL*	*yarabts.ABS.ACC.PL*	...

Free Translation:

Traditional folk songs. Besides presenting various kinds of tales (*lax ˚naku*), lament recitatives (*lax ˚nako*), and heroic recitatives ...

The second treatment of the same text by Paakkan, shown in (6), gives a five-level analysis, with the word level (\TEXT) segmented into morphs (\UNIT) which are then aligned to inflectional coding (\MNNG). A base form of the word is then given (\BASE), and a phrasal translation (\MITA) which is aligned to the phrase level (\ref). No gloss is provided.

(6) Tundra Nenets Text (Paakkan, 1997)

```
\TEXT nyewøxi       nyenæcyøjeq                    syoq.

\UNIT nyewøxi       nyenæcyøje          q          syo          q

\MNNG {noun+∅+vbs} {noun+n/j+acpl+cbs} {suf+genpl} {noun+jo+gbs} {suf+nompl}

\BASE nyewøxi       nyenæcyøh                      syo

\MITA Traditional folk songs.
```

The Diyari and Yidinj texts in (7) and (8) have similar layouts, each with three rows: text, gloss and free translation. The phrases are numbered, intonation groups are marked with a slash (/), and comments appear between after some phrases.

(7) Diyari Text (Austin, 1981:252)

5. *ŋaḍa-n̪i* *kaṯi* *ḍukaṛa-ṇḍa* *pudi-yi* *paḷu* *wapa-ṇḍa*
   then-LOC clothing-ABS take off-PART AUX-PRES naked go-PART
   pudi-l̪ali *ŋan̪ṯi* *ṯuŋka-li* *widi-ṇḍa* *pudi-ṇḍa*
   AUX-IMPL$_{SS}$ meat rotten-ERG paint-PART AUX-REL$_{SS}$
   Then they took their clothes off to walk around naked painted with some rotten meat

   The *ṭinanipa* rubbed their bodies with the rancid fat of a dead animal and hence gave off a strong smell (see also line 17).

(8) Yidinj Text (Dixon, 1977:527)

97. *ŋayu* *ḍuŋga:na /* *ŋaɲaɲ* *gula* *baga:ḍina*
    I-SA run-PURP I-O body-ABS pierce-*:ḍi*-PURP /
    'I had to run [in the fight], and as a result my body got speared'.

Comments are distinguished from interlinear text using alignment to the outer left margin. Despite these visual similarities between these two samples, there is an important structural difference: the Diyari text segments the words into morphs using hyphens, while the Yidinj does not. Thus, the transcription of the Diyari text is at the level of the morph, while for the Yidinj text it is at the level of the word.

Extra information can be included in interlinear texts without adding much complexity. The sample from South Efate in Fig. 1 represented in Shoebox format (SIL, 2002) has more than a dozen rows (some optional). At the text level, we see commentary on the source `\nt` and a reference to the audio/video file `\aud`). Each phrase contains the audio start and end (`\as`, `\ae`), then two rows of text: one at word level (`\tx` may correspond to an orthographic representation) which is then segmented into morphs using hyphens (`\mr`). These are vertically aligned to a gloss (`\mg`) and labelled with a part of speech tag (`\POS`). Free translation is given in two languages, English (`\fg`) and Bislama (`\fgb`).

```
\nt Story from tape 20001bx told by Kalsarap Namaf. Transcribed and
translated into Bislama by Manuel Wayane. The story concerns a natopu
or spirit called Litrapong, also known in Bislama as a Lisepsep.
Story is also told on video.

\aud kalsrap.mov
\as 0
\ae 13.0002
\tx Akit    tumaui              tae   esan  ipi,       go
\mr akit    tu-       mau       tae   esan  i     - pi go
\mg 1plincS 1plincRS- all       know  place 3sgRS - be and
\POS pron    pron-     quantifier vambi n     pron  - v  conj

\fg We all know that place, and this Litrapong, I want to tell you
about Litrapong. She is of grandfather's clan. Those two, grandfather
and Litrapong, would talk every now and then.
\fgb Yumi evriwan isave ples ia. Mo Litrapong (Lisepsep) ia. Mi wantem
talem long yufala abaot Litrapong ia. Hemi naflak blong olfala. Hem mo
olfala (apu) tufala istap storian samtaem.
```

Figure 1: South Efate Text (Namaf, 2001)

```
Source:    \dn14.003.01

Speakers:  Bindie West (Moreland) (speaker code: B), Stumpy George (speaker code: S)
           In the transcript, Flint actually uses three code letters: B, S and G.
           G may stand for "Stumpy George".

The sound recording of this material is not available. The
conversation recorded on the audio-tape (\cm R297C) presumably
corresponds to the text transcribed in \dn 14.003.01.  However, the
informants' speech was very rapid and somewhat unclear, and no clear
connections between the recording and the transcript could be made.

This text material is unanalysed, so the analysis, gloss and "other
sources" fields are not included.


\sp   G
\ft   nanyi??/nganyi??     ngaranjan    jilajbaya     ngabayan          yanba   cont.
\fg   my                   mother       go            white man         talk    cont.
\ft   ngangangi
\fg   want
\ncft It is difficult to determine word order here, and which sentences words belong to.
\ncfg 'white man' refers to a 'Mr Haely??', whose name is written beneath ngabayan.
\fft

\sp   G??
\ft   kudbinji     jilabaya kingkarina    (namukiya    yan ba lajba)     cont.
\fg   happy        go       down          paint        paint??           cont.
\ft   yalungka
\fg   all fellows down at [camp]
\ncft
\ncfg
\fft
```

Figure 2: Garrwa Text (Laughren et al., 2002)

Another example of a text which incorporates supplementary information comes from Garrwa, shown in Fig. 2. In this sample, a dialogue between two speakers (labelled 'B' and 'G') and the notes made by a researcher in the 1960s are supplemented by rows containing notes by later analysts. Most notes apply to a particular constituent (e.g. the entire text, a particular phrase or word), some notes pertain to multiple occurrences even though they overtly refer to a single instance (e.g. a phonological feature which appears throughout a text). Annotations at multiple levels and by multiple analysts raise issues of consistency in the markup, which is an issue affecting a wide cross-section of linguistic and humanities researchers.

The previous text samples from South Efate and Garrwa suggest that there is no upper bound on the number of rows an interlinear text may have. Is there a lower bound? As noted earlier, the most common basic form of interlinear text has three rows, however texts having two rows can be found. The Latin example in (9) contains a row of English translation above the Latin source text.

(9) Latin Text (Valiulis and Wasson, 1998:12)

Epicurus	indeed	from	the souls	of men	tore out	by its roots
Epicurus	vero	ex	animis	hominum	extraxit	radicitus
religion	when	from	the immortal gods	both	generosity	and benevolence
religionem	cum	dis	inmortalibus	et	opem	et gratiam
he removed	For while	the best	and	most excellent		nature
sustulit.	Cum enim	optimam	et	praestantissimam		naturam
he says is god's,	he denies	this very thing	to be	in god --		
dei dicat esse,	negat	idem	esse	in deo		
benevolence.	he removes	the very thing which is	most	natural		to the best
gratiam:	tollit	id quod	maxine	proprium	est	optimae
and most excellent		nature	...			
praestantissimaeque		naturae.	...			

Observe that the glosses can almost be read as a phrasal translation, and that the alignment of gloss words to source text words is many-to-many (e.g. *radicitus* = by its roots). This interlinear text is atypical, probably because it is intended to demonstrate the interlinear function of a database system.

So far all the source texts have been represented in Roman scripts. Scripts which are written from left-to-right pose a further challenge. Consider the Hebrew text in Fig. 3, where right-to-left Hebrew text is translated using the left-to-right system of English. Observe the text-wrapping within the English glosses, and the additional annotations, e.g. *'Elohim* = God (plural of excellence). Above the source text there is a row of cross references to Strong's Concordance (Strong, 1996). The numbers in the left margin refer to the canonical biblical versification,

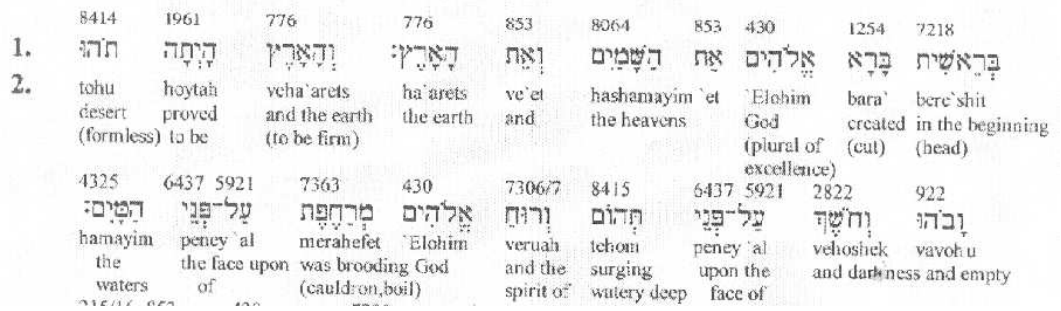


Figure 3: Hebrew Text (Mullins, 2001)

---

although the verse boundary appears within the line and is not explicitly marked (e.g. the break between verses 1 and 2 occurs between the words *ha'arets* and *veha'arets* in the first line). Correct handling of bidirectional interlinear text falls outside the scope of this paper; we return to the issue briefly in Section 6.2.

2.1 Interlinear texts with higher level annotations

A sample of Indonesian interlinear text in (10) gives a different perspective on interlinear text. The interlinear text is unusual in that it represents nested syntactic structure.

(10) Indonesian Text (Simon Musgrave, pers. comm. 2003)

```
Morphology:  dia meN-lihat NP[wanita DC[yang perg1 AP[ke Lombok]]]
Gloss:       3   ACT-see      woman     REL  go       to Lombok
Word Class:  0       V        N         HLnk V        P  N
```

More complex examples of interlinear text may include higher level analysis, such as discourse features. The Denya example in Fig. 4(a) has two lines of text, with words separated into morphs using hyphens, aligned with the gloss. A phrasal translation is given separately on the page, however it is numbered according to each phrase of the text, and therefore conceptually belongs to each phrase, where visually it appears to correspond to the entire text. The numbered phrases themselves however are further divided into sub-phrases indicated by letters, and an extra dimension is added through the charting of the text on the page such that the verb phrase appears in a second column. This relates to the purpose of this particular text, which is part of a monograph examining verb functions in this language. Similarly, the Yele example in Fig. 4(b) is laid out in clauses, with notes on the semantic relationships between

11(a)		*ɔbɛ-gé*	*mɛ́* *muú* *ŋgbɛ*
	(if)	you-are(Cond)	already person Ngbɛ
(b)	*jyɛ́ɛ́*	*ɔ́-jyɛ-ɛ́*	*ndé* *melɔ*
	even-if	you-who-went(RelPst)	what village
(c)	*ɛwí* *ŋgbɛ*	*á-lú*	
	which Ngbɛ	it-is(Attr-Rel)	
(d)		*ɔ-kpɛ-ne*	*wiɛ́* *retu*
		you-go-in(NonPst)	there nothing

(a) Denya Text (Abangma, 1987:112f)

16 a.	*Kwo-d:o.* to.him-I said.IM.PST	ADDITION to 15
b.	*Daa* *tóó.* CI.PRES.3.SB.NEG sitting	CONTENT of a (answer to 15)
17 a.	[*Ndoo apɛ̂*]$_{M}$ [*d:aa*]$_{TOP-NEG}$ [*k:âmo*]$_{COM}$ maybe maybe I'm.not good.fisherman	CONTENT of 16a and CONCLUSION to 16b

(b) Yele Text (Henderson, 1995:92f)

Figure 4: Interlinear Texts Containing Discourse Structure

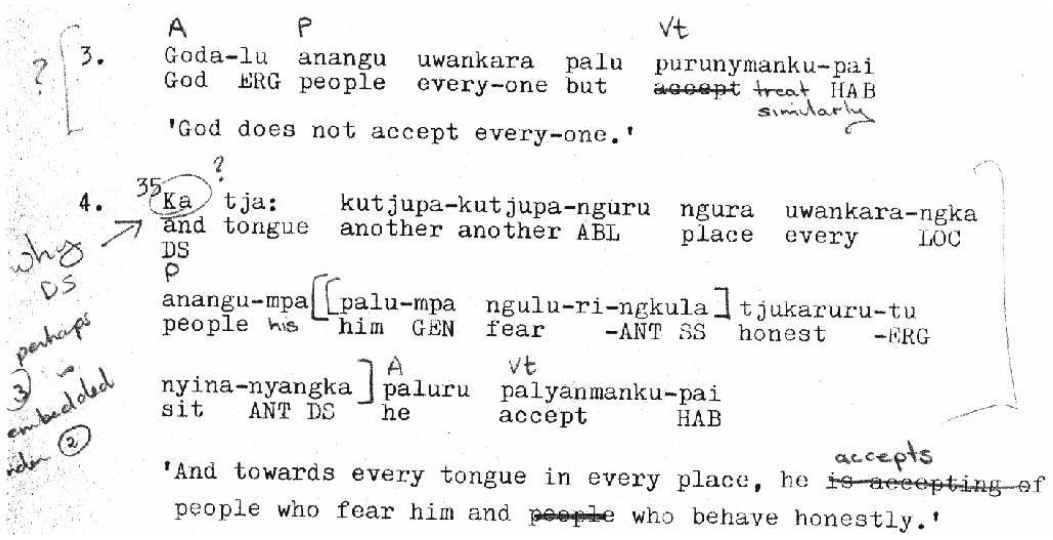


Figure 5: Pitjantjatjara Text (Heather Bowe, pers. comm. 2003)

the clauses. While such features may be useful for some interlinear texts, the question remains as to whether or not interlinear text processing engines should include such functionality.

All of the texts we have considered are complete; they are in their disseminated form and not works in progress. However, fieldwork involves incomplete analysis of data, which must still be incorporated in the process of interlinearisation. An example combining full and partial analysis is shown in the Pitjantjatjara text in Fig. 5. The printed version resembles the standard 3-line texts already described (words, glosses, phrasal translation), however additional information is hand-written on the page, such as use of square brackets to link certain constituents, part of speech annotations on some elements, evidence of corrections and revisions, and some queries noted in the margin. Such a text reflects the practice of a field worker (who annotates the text in various ad hoc ways) more accurately than many of the published texts previously described (which formalise the material for presentation). These annotations raise questions of how much additional information needs to be incorporated in interlinear texts, as well as how ambiguous information should be incorporated.

3 Requirements for a General Model of Interlinear Text

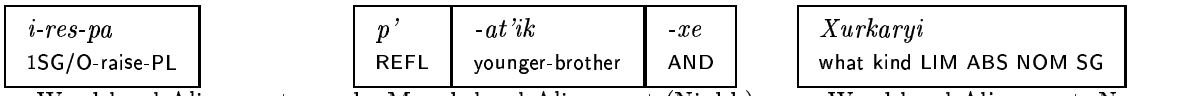
This survey has raised many issues surrounding the structure and presentation of interlinear texts, in the context of observations about formatting, alignment, grouping and wrapping. This section considers each of these issues, and identifies core requirements on any general model for representing interlinear text.

3.1 Content

The simplest examples of interlinear text give three types of information: a row of source text, a row of glosses, and a phrasal translation. It is common to include metadata relating to the text as a whole, such as title and author. Some texts (e.g. Nivkh, Diyari, Yidinj) have commentary on particular phrases or on cultural practices, some include part-of speech information (e.g. South Efate, Indonesian), and some include prosodic information (e.g. intonation boundaries in Diyari and Yidinj). Interlinear texts may contain material in multiple languages (e.g. South Efate), material that is contributed by multiple speakers and/or accorded multiple analyses (e.g. Garrwa). Texts may reference an external resource such as a concordance (e.g. Hebrew). Specific aspects of a source text can be analysed in detail, such as phonetic, syntactic, or discourse information (e.g. Yidinj, Indonesian, and Denya respectively). Interlinear text may be incomplete and subject to revision, as illustrated by the corrections and queries in the Pitjantjatjara text. We conclude that any general-purpose model must support an arbitrary number of user-specified rows, each having a user-specifiable type (e.g. source text, gloss). Moreover, we conclude that commentary is intrinsic to interlinear text and that no structural distinction can be drawn between linguistic analysis and more general commentary; commentary is thus a first-class member of the data structure.

3.2 Presentation

The samples covered in our survey manifest great variety in their presentation. The most consistent feature is that the row of source text appears first, and beneath it the row of glosses. The phrasal translation may appear together with each phrase or sentence, or else as a separate block which may be coindexed with the source text. Line numbers support this coindexing, as well as external citation (e.g. Ainu, Hebrew). Most of the texts use typeface to distinguish the role of a particular row of interlinear text, for example presenting the text line



a. Word-level Alignment; Morph-level segmentation (Ainu)

b. Morph-level Alignment (Nivkh)

c. Word-level Alignment; No Morph-level segmentation (Susoi's Tundra Nenets)

Figure 6: Examples of Alignment

in bold (e.g. Nepali) or italic (e.g. Diyari) to distinguish it from rows of analysis or translation. In (5) there is a combination of styles in the gloss row: 'pierce-*:ḍi*-PURP'. In (8) indentation is used to distinguish lines of interlinear text from comments. The use of capital letters or small caps is a common way of distinguishing grammatical information from glosses within the text (e.g. the Nepali gloss dog-GEN). A general-purpose model of interlinear text should abstract away from all such matters of presentation, while still encoding the necessary distinctions so that presentation software can generate the desired presentation.

3.3 Vertical Alignment

The content of two or more rows may be aligned at either the word or morph level. Fig. 6 illustrates a representative set of cases from our survey, using rectangles to identify the multi-row units which are aligned.

The correspondence between the word or morph and its gloss is one-to-one in (a) and (b), and one-to-many in (c). Observe that (a) contains hyphens to indicate the morphs; while these morphs are not visually aligned, there is a clear one-to-one correspondence between the source text and the glosses at the morpheme level. Thus (a) exhibits both word- and morph-level correspondence. In the case of (c) we cannot tell whether the source text can be segmented into morphemes. Such a one-to-many mapping may result when we have a word whose internal structure is unknown, or where the word is a portmanteau, or some combination of the two. In the Nepali text we saw a word containing two one-to-many mappings: *nisk-yo* = come.out-3smL.PST.

Both kinds of one-to-many mapping are correctly handled by the model. In cases such as (c) we simply have a one-to-many mapping from one unsegmented word to many morph-level glosses. In the second case, we have a one-to-many mapping from one portmanteau morph to many morph glosses. Further examples are easy to find, e.g. in the Nivkh text in (4) we

see a form *r* glossed as GER-3SG and in the Diyari text in (7) we see a form *kaṯi* glossed as clothing-ABS.

Other mappings are possible, though less frequent: a many-to-one mapping appeared in the Ainu text (*ran-pes* = cliff), and a one-to-zero mapping occurred in the Latin text (*est* = 0). Another such mapping arises in the case of epenthetic morphemes (e.g. *t* in French *y a t-il*). The hardest case, which we have not seen attested and which is not covered by the model, arises for non-concatenative morphology, where two or more non-contiguous morph fragments are mapped to a single morph gloss. We believe that the morpheme intercalation and infixation processes should not be implemented inside an interlinear text model, and that it is sufficient to represent all such cases concatenatively (e.g. representing Arabic *k-a-t-a-b* as *ktb-a*).

Above we observed that the content of two or more rows may be aligned at either the word or morph level. We have also seen another kind of alignment between source texts and their translations: in (7) the translations are interspersed between every phrase,[1] while in (4) the translation comes after the entire text. We conclude that the underlying structure of interlinear text must permit information to be associated at any of these four levels, namely morph, word, phrase, and text.

3.4 Line Wrapping

Several texts in the survey illustrated line wrapping. For instance, in the Diyari text in (7), words and their glosses are carried forward to the next line as a unit, while the phrasal translation is wrapped independently. These multirow units which are indivisible for the purposes of line wrapping will be termed *interlinear text cells*. Fig. 7 reproduces this text simulating two different page widths; each interlinear text cell is framed.

This wrapping behaviour follows from the most basic property of interlinear text, that the annotations of source text must be vertically aligned. While some exceptions to this behaviour can be found (e.g. in the second line of the Nivkh text in (4) where part of a gloss has been wrapped), the behaviour is pervasive and must be supported by any general-purpose presentation model.[2]

[1] The connotation of the word *phrase* is imprecise. It may correspond to what Simons has described as 'units of convenient size for translation, often more than a single clause but less than a sentence' (Simons, 1989), and to what Selkirk has called the *sense unit* (Selkirk, 1984).

[2] There is a parallel between our notion of interlinear text cell and Sproat's *Small Linguistic Unit* (Sproat, 2000:25), which he developed in the context of an analysis of non-Roman scripts.

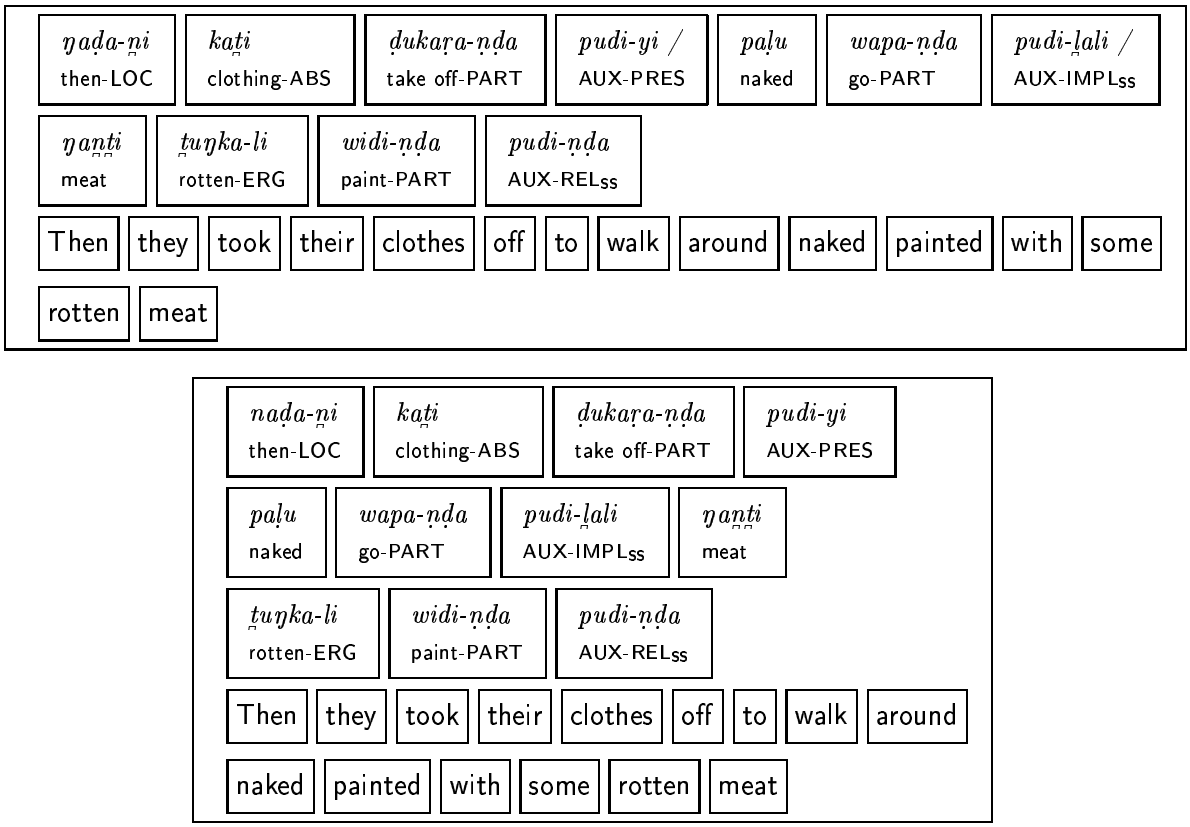


Figure 7: Diyari Text with Wide and Narrow Formatting, Marking Interlinear Text Cells

3.5 Audio alignment

The source for many interlinear texts is an audio or video recording which has been transcribed and analyzed. In the simplest cases, the source media can be documented as a text-level property, so that it can be located if any question arises as to the veracity of the transcription. For example, in Fig. 1 the source was transcribed from a movie file `kalsrap.mov`.

For extended recordings it is often convenient to link the interlinear text to the digitized source media, phrase by phrase. Rather than segmenting the source into clips, it is common to indicate the start and end offset as properties at the phrase level. This can also be seen in Fig. 1, where the `\as` and `\ae` fields contain the audio start and end times.

4 A Four-Level Model of Interlinear Text

The preceding survey of a range of interlinear text types represents an attempt to deduce a conceptual model of an interlinear text from its presentation on the page. In this section we develop a formal model and show how it can represent the examples we have seen. We begin by examining the hierarchical structure of interlinear texts, leading to a formulation based on simple tree diagrams.

4.1 The hierarchical structure of interlinear text

The alignments and groupings observable in interlinear text can be represented using an hierarchical model. One row is represented as dominating another (e.g. words dominating morphs) if the dominated content is aligned as a group to the dominating content. Two rows aligned with each other (e.g. morphs and their glosses) are represented within a single node of this hierarchy. To see how this works, consider Susoi's Tundra Nenets text. Here, a row of source text appears above a row of glosses. A translation of the entire text appears at the bottom of the page. We can abstract away from the idiosyncracies of the layout and represent the logical structure of this text as the hierarchy in Fig. 8(a). The root node of this tree contains the translation of the entire text. This node dominates three child nodes, each containing words. Each word in turn dominates several morph nodes. In contrast to this, Paakkan's version segments the source text into morphs, and so the corresponding tree in Fig. 8(b) represents the text at the morph level. Observe that morphs and their glosses appear together inside the leaf nodes of this tree.

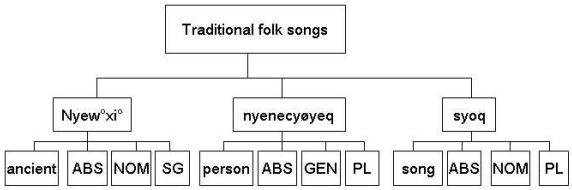
It is often the case that the content at the word level is fully predictable from the morph level. Consider the Nepali example in (2). We contend that the representation of this text has no content at the word level. Instead, morphs are concatenated with intervening hyphens. However, this text is more than a concatenation of morphs, since word-level units are clearly apparent (serving as the basis for alignment). Accordingly, we employ word-level constituents in the tree structure, but leave them empty as shown in Fig. 8(c).

Now that we have introduced a distinction between structure and presentation, it is instructive to consider texts which are presentationally similar yet structurally different. Consider again the Yidinj and Diyari texts, in (8) and (7) respectively. The presentation of the two texts is virtually identical, except that words are segmented into morphemes in the Diyari text. Accordingly, the structure for the Yidinj text in Fig. 8(e) has source text at the word level and glosses at the morph level. In contrast, the structure for the Diyari text has both source text and glosses at the morph level, and nothing at the word level. Even though the word level is empty, it is still required in order to represent the grouping of morphs into words, information that is clearly present in (7).

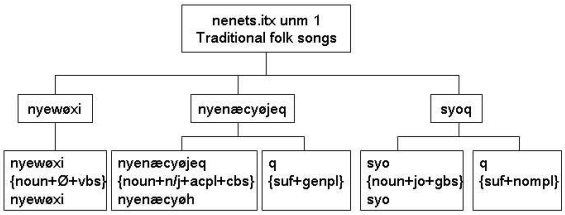
Just as we have examples with similar presentations and different structures, we can try to find cases of different presentations for similar structures. Consider again the Ainu example in (3). The translation appears after all the text, as in the Nivkh example in (4). Yet the sentences of the translation are co-indexed with the phrases of the source text. Therefore, we treat these as translations at the phrase level which have simply been concatenated to form a text-level translation. Part of the corresponding tree is shown in Fig. 8(d), and the reader will observe that it has the same structure as the Nepali tree in Fig. 8(c).

Note that the highest level (for the entire text) has been elided from these trees. Since we have only been displaying one phrase at a time this top level has been redundant. We conclude the discussion with an example which has content at four different levels. The South Efate text includes metadata at the text level, plus another eight lines of information, all of which fit into the tree shown in Fig. 8(g). As implied by dotted lines in this example, further phrases can be attached to the text node and analysed accordingly.

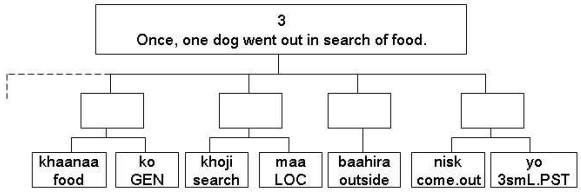
Generalizing over these examples, we propose the four-level model shown in Fig. 8(h). The root level corresponds to the text as a whole, and it is comprised of a sequence of phrases. Each phrase is made up of a sequence of words, where each word is a sequence of morphs.



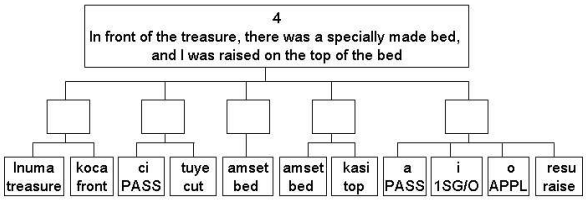
(a) Tundra Nenets (Susoi) Tree, cf. (5)



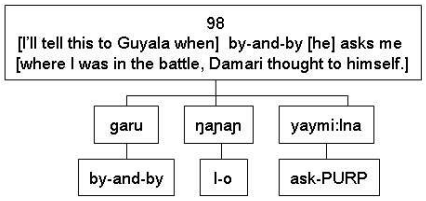
(b) Tundra Nenets (Paakkan) Tree, cf. (6)



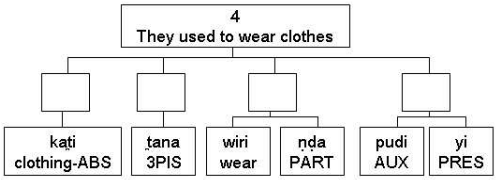
(c) Nepali Tree, cf. (2)



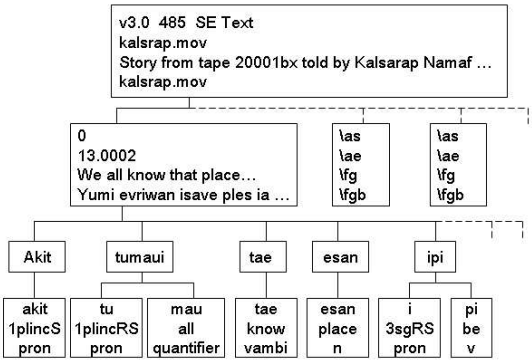
(d) Ainu Tree, cf. (3)



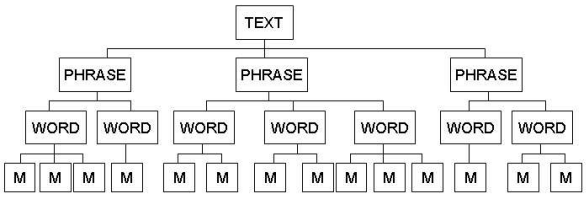
(e) Yidinj Tree, cf. (8)



(f) Diyari Tree, cf. (7)



(g) South Efate Tree (cf. Fig. 1)



(h) Schematic Representation (M = morph)

Figure 8: Examples of Interlinear Text Trees

4.2 Formal definition

An interlinear text is comprised of an idiosyncratic set of rows populated with different types of information that are meaningful to the analyst. The details of this content and its interpretation in the subject language or in the linguistic metalanguage are outside the scope of the model. In order to get started, we simply assume that each such type is defined as a set of all the permissible content. For instance, an IPA transcription type would be defined as IPA*, the set of all strings that can be built on the IPA alphabet. Similarly, an orthographic transcription would be built on some character repertoire represented using (extended) ASCII or Unicode (this encompasses non-Roman and right-to-left scripts). Similarly, a morph level gloss type would be defined as any morph from the lexicon or any morphosyntactic tag. The type for a media file or a start time would be a formatted string.

An *interlinear text signature* is a four-tuple $\langle T, P, W, M \rangle$ where each of T, P, W and M is a set of types. (Note that T, P, W and M denote text, phrase, word and morpheme levels respectively.)

We illustrate this definition by means of an example for the Diyari text. Let Σ be the letters used in the English phrasal translation; let L_d be the set of Diyari morphs (possibly from the lexicon); let L_e be the set of English morphs (from the same lexicon); and let L_m be the inventory of morphosyntactic categories. Then the interlinear text signature for Diyari is as follows:

$$\begin{aligned} T &= \emptyset \\ P &= \{p_1\}, p_1 = \Sigma^* \\ W &= \emptyset \\ M &= \{m_1, m_2\}, m_1 = L_d^*, m_2 = (L_e \cup L_m)^* \end{aligned}$$

An *abstract interlinear text* over $\langle T, P, W, M \rangle$ is a tree of depth four where each node contains a set of sequences. Nodes of depth 1 must have content drawn from T, nodes of depth 2 must have content drawn from P, and so on. Note that an abstract interlinear text contains all the information that the presentation of an interlinear text is intended to convey.

The Diyari tree in Fig. 8(f) is now an example of an abstract interlinear text. To underscore the fact that there is nothing intrinsically significant about the tree diagram, we represent the same abstract interlinear text as a nested bracketing:

```
(
  ( 4 They used to wear clothes ...
    ( ( kati ; clothing-ABS ) )
    ( ( tana ; 3PIS ) )
    ( ( wiri ; wear ) ( nda ; PART ) )
    ( ( pudi ; AUX ) ( yi ; PRES ) )
  )
)
```

The relationship between abstract interlinear texts and their presentation forms can also be systematized. We need to select the types to be displayed, their ordering, and their grouping. The simplest way to specify ordering and grouping for a selection of items is to specify a tree.

An *interlinear text presentation signature* over $\langle T, P, W, M \rangle$ is a tree of depth at most four, whose leaves are a subset of the types from $\langle T, P, W, M \rangle$.

To illustrate this definition consider again our running example. The Diyari text in (7) displays the source text and gloss, grouping them at the word level. The translation appears below each phrase. The required interlinear text presentation signature is the following tree:

```
(
  (
    ( m_1 m_2 )
    p_1
  )
)
```

We can interpret this structure as a set of commands to navigate an interlinear text instance down to the word level, concatenate and print the sequence of morphs m_1 and their glosses m_2 (ignoring any morph-level structure); once this has been done for a phrase, display the translation p_1. Note that the process of realizing an interlinear text as a human-readable document involves further steps: selecting fonts and styles, rendering the text (in the required

directionality), and delivering it in a format that is supported on the user's platform. We do not attempt to formalize this final step.

4.3 Implementation

The four-level model of interlinear text is a conceptual model; it defines the information and structure that must be present in an interlinear text. It generalizes away from instances of interlinear text, from presentation formats, and from the interpretation of the content either in the subject language or the linguistic metalanguage.

We implement the model as follows: an interlinear text signature can be implemented as an XML DTD or Schema; an interlinear text instance can be implemented as an XML document which is valid according to that DTD or Schema; and an interlinear text presentation signature can be implemented as an XSLT transform. Full details of this implementation are beyond the scope of the paper; interested readers are invited to contact the authors for the implementation.

The four-level model can also be implemented using Annotation Graphs (Bird and Liberman, 2001); it is a specialization of the interlinear text model presented by (Maeda and Bird, 2000) and implemented in the InterTrans tool (Bird et al., 2002).

5 Conclusion and Further Research

We have presented a conceptual model of interlinear text that is able to represent the information content of many kinds of interlinear text. The model has four levels, text, phrase, word and morph, organized in a hierarchy. In this section we rebut a common objection to the four-level model, and briefly report ongoing work on implementation.

The artificial limitation of the model to four levels has been identified as a possible shortcoming by some readers. Indeed, three of our examples show one or more levels of structure between the existing word and phrase levels: the Pitjantjatjara text in Fig. 5 contains nested syntactic phrase markers, while the Denya and Yele texts in Fig. 4 have lines of interlinear text numbered at both the sentence and phrase level. However, we believe are not counter-examples to the four-level model since there is enough flexibility in the mapping from abstract interlinear texts to presentation formats to generate these particular presentations. In the case of the Pitjantjatjara text, there is no alignment to the syntactic phrase markers, so the markers can be represented directly in the text. The four lines of the Denya text, identified as (a)

through (d), can be presented as phrase level constituents in the four level model. The phrase level must also carry a boolean property (say, major vs minor phrase) which is used in the presentation process to generate the outer enumeration (e.g. 11) or the inner enumeration (e.g. (c)). A similar method works for the Yele text. We expect that true counter examples would be hard to find, since they would need to demonstrate simultaneous alignments at five levels.

In ongoing work we are developing an interlinear text tool as part of the Annotation Graph Toolkit, extending the existing InterTrans tool (available from `agtk.sourceforge.net`). We are also extending our XML-based implementation to support dynamic rendering, and delivery of archived interlinear texts with linked audio.

In presenting this model of interlinear text we hope to have demonstrated the merit of the distinction between structure and presentation. More than this, we hope that our simple formalization will provide a secure foundation for developing interoperable tools for creating interlinear text, and a simple model for interchanging and archiving interlinear text.

References

Abangma, S. N. (1987). *Modes in Dényá Discourse.* Summer Institute of Linguistics and University of Texas at Arlington.

Austin, P. (1981). *A Grammar of Diyari, South Australia.* Cambridge: Cambridge University Press.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

Bird, S., Maeda, K., Ma, X., Lee, H., Randall, B., and Zayat, S. (2002). TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse tools built on the annotation graph toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 364–370. `http://arXiv.org/abs/cs/0204006`.

Comrie, B. (1981). *The Languages of the Soviet Union.* Cambridge: Cambridge University Press.

Dixon, R. M. W. (1977). *A grammar of Yidinj.* Cambridge: Cambridge University Press.

Genetti, C., editor (1994). *Aspects of Nepali Grammar.* Santa Barbara: Department of Linguistics, University of California.

Henderson, J. (1995). *Phonology and Grammar of Yele, Papua New Guinea.* Canberra: Research School of Pacific and Asian Studies, Australian National University.

Huttar, L. A. (2003). Constituent charting for discourse analysis: Information model and presentation model. Master's thesis, Graduate Institute of Applied Linguistics, SIL.

Laughren, M., Keith, N., and Hughes, B. (2002). Garrwa: Australian Aboriginal language data from the University of Queensland Flint Archive. `http://emsah.uq.edu.au/linguistics/austlang/garrwa/index.html`.

Maeda, K. and Bird, S. (2000). A formal framework for interlinear text. In Bird, S. and Simons, G., editors, *Proceedings of the Workshop on Web-based Documentation and Description.* `http://www.ldc.upenn.edu/exploration/expl2000/papers/`.

Mullins, M. (2001). Genesis Hebrew interlinear. `http://community.webshots.com/album/9436265MwWLblfMhl`.

Namaf, K. (2001). Litrapong. Transcribed from audio tape by Nick Thieberger, University of Melbourne.

Paakkan, J. (1997). A proposal for an arrangement of an interlinear morphological corpus of Tundra Nenets. `http://www.helsinki.fi/~jpaakkan/Interlinear.html#Appendix2`.

Selkirk, E. O. (1984). *Phonology and Syntax.* Cambridge, MA: MIT Press.

Shibatani, M. (1990). *The Languages of Japan.* Cambridge: Cambridge University Press.

SIL (2002). The Linguist's Shoebox. `http://www.sil.org/computing/shoebox/`.

Simons, G. (1989). Proposed framework for encoding analysis and interpretation of running text. `http://tei-c.org/Vault/AI/aiw01.txt`.

Sproat, R. (2000). *A Computational Theory of Writing Systems.* Cambridge: Cambridge University Press.

Strong, J. (1996). *Strong's Exhaustive Concordance of the Bible.* Nashville: Thomas Nelson Publishers.

Susoi, E. G. (1990). Nenècie literatura. Transcribed at http://www.helsinki.fi/~tasalmin/text.html.

Valiulis, D. and Wasson, G. (1998). *Adobe FrameMaker Template Series Template Pack 4 Guide*. Adobe Systems Incorporated. http://www.adobe.com/products/framemaker/tempseries/pdfs/tempac4.pdf.

Acknowledgements

The authors would like to thank Heather Bowe, Simon Musgrave and Nick Thieberger for furnishing us with their unpublished interlinear texts. This paper has benefited from feedback from participants at the EMELD Workshop on Digitizing and Annotating Texts and Field Recordings, July 2003. Our terminology for 'abstract interlinear text' was inspired by parallel work by Lars Huttar on discourse charts (Huttar, 2003). This material is based on work supported by the National Science Foundation under Grant No. 0094934 *Electronic Metastructure for Endangered Languages Data.*