

A Report on the Automatic Parsing of Tigre

Lots of Morphology and Bits of Syntax

Paper prepared for the International Workshop
“History and Language of the Tigre-Speaking Peoples”,
Università di Napoli “L’Orientale”,
February 8th-9th 2008
by Klaus Wedekind

Purpose of the presentation

- The main purpose of this presentation is to make linguists aware of the existence of data bases which contain lexical as well as morphological information about Tigre.
- This information might be useful and could be made available for further studies and further development of the language.

Status of the parser

- The Tigre parser is part of a machine translation project which was first presented at the *“International Conference on Language and Literature”*, Asmara 2000, and which had been launched at the *Ministry of Education, Asmara*, by Saleh Mahmud and Klaus Wedekind.
- In the early stages, the Tigre grammar by *Shlomo Raz* has served as a basis. This was supplemented by informal talks with the late Prof. Raz.
- Starting point was a trilingual Eritrean dictionary, to which Tigre entries were added.

Transcription

- For practical reasons, the Tigre parser starts from the Geez script, accepting its limitations.
 - So gemination had to be disregarded. Because of this, the parser produces a larger numbers of ambiguities.
- The input is transliterated into the Latin script and back again, by means of “visual basic macros”.
 - The next frame shows a few lines from the macro which changes the Eritrean “Geez Type” syllables into Latin script letters:

Automatic transliteration

- The Eritrean “Geez Type” syllabary often requires more than one ASCII symbol for one “fidel” symbol:
 - ASCII 68 is ለ Tigre la, but
 - ASCII 68 followed by ASCII 234 is ሊ Tigre lu
 - ASCII 70 is ለ Tigre l
 - (ASCII 32 is space)
- Thus, the “macro” uses command such as these:

• If L1 = 68 And L2 = 32	Then L\$ = "la " : Goto A2
• If L1 = 68 And L2 = 234	Then L\$ = "lu" : Goto A3
• If L1 = 68 And L2 = 239	Then L\$ = "li" : Goto A3
• If L1 = 69 And L2 = 32	Then L\$ = "laa " : Goto A2
• If L1 = 68 And L2 = 237	Then L\$ = "le" : Goto A3
• If L1 = 70 And L2 = 32	Then L\$ = "l " : Goto A2
• If L1 = 68 And L2 = 247	Then L\$ = "lo" : Goto A3

Transliteration of some consonants

Cons.	
h	ʋ
l	ɒ
h`	ɬ
ʃ	ɸ
k`	ɸ
ʼ	ʒ
&	ɔ
t`	ʈ

Transliteration of vowels

Vow.	
-a	ʊ
-u	ʊ̣
-aa	ɣ
-i	ɰ
-e	ɰ
	ʊ
-o	ʊ̣

Statistics of the Data Bases

- Currently the parser uses “CARLA” software, three data bases (one for 32 prefixes, one for 7985 roots, one for 66 suffixes), and “rules”.
 - Among the roots, there are about 1300 verbs, 5180 nouns, and 2600 “others” (i.e. 96 prepositions, 79 pronouns, 12 numerals, 75 names, 39 interjections, 13 demonstratives, 50 conjunctions, 362 adverbs, and 779 adjectives), many of them tagged by *Saleh Mahmud*.
 - Another 9200 English entries have been given Tigre glosses by *Abu Harish* and *Mohammed Idris*, but are waiting to be tagged.

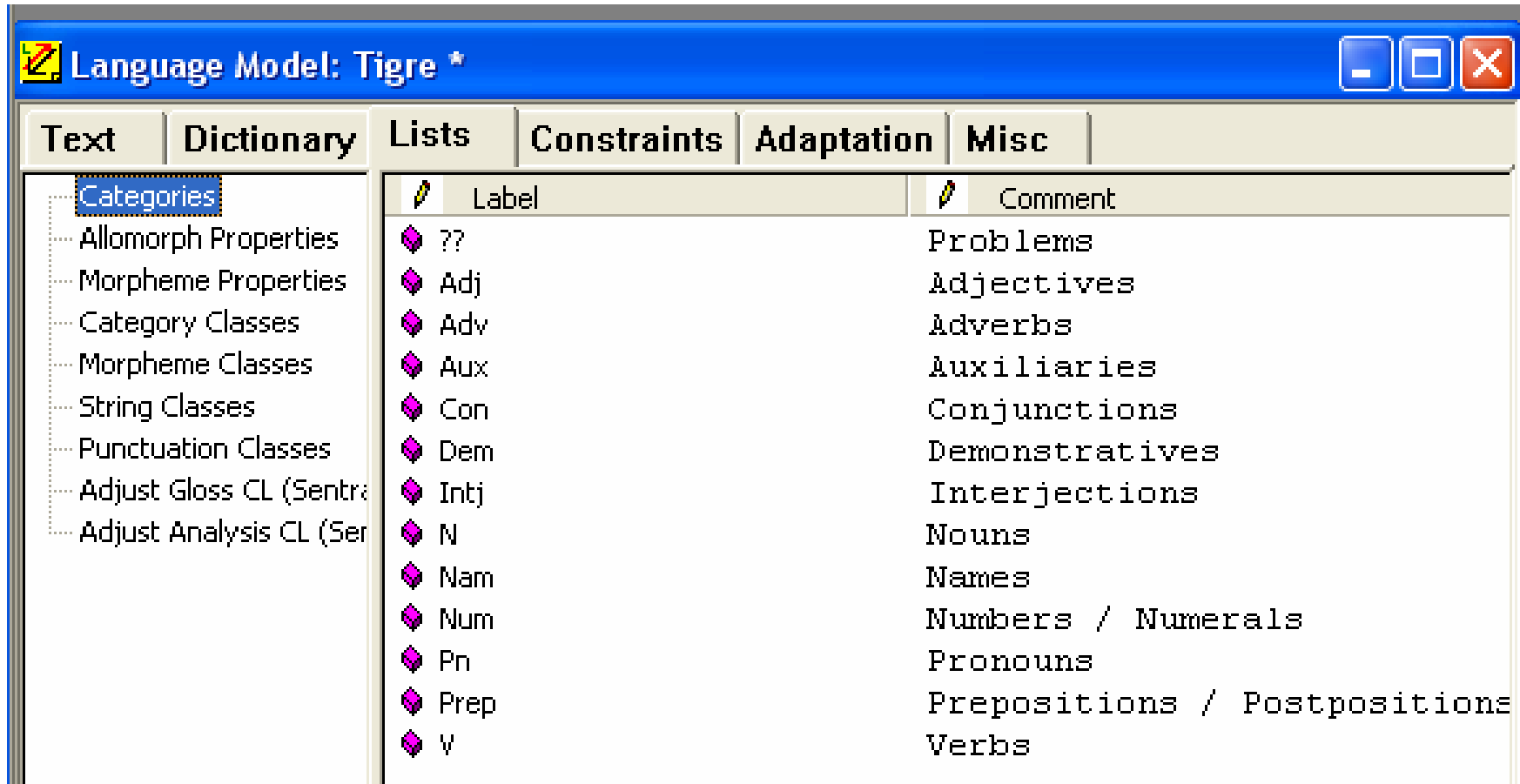
Overview of three data bases

- Here is a “list” view of the three data bases:
 - Prefixes (left), roots (centre), and suffixes (right)
 - For “roots” (data base in the centre), four columns are shown:
Tigre in Latin script | Tigre in Fidel script | English gloss | Morpheme

TEPF7.DB		ABCJun2. db:1				TESF7.DB			
\ENGL EnglishGl	\TEPREF Tigre	\TEC TigreC	REL Tigrel	\TEG Tigre	\ENGL English	\TEC TigreT	\ENGL EnglishGl	\TESUFF TigreT	\TEC TigreClas
ImpP11pf	h	V/V		*no field*	six**3	Num	DerAbstr	nat	N/N
ImpP12Fpf	t	V/V	ssa	ሰሳ	sixty	Num	DerAbstr2	naN	N/N
ImpP12Mpf	t	V/V	stur	ሰቲር	close**6	Adj	DerGerund	ot	V/N
ImpP13Fpf	l	V/V	styat	*no field*	drinking**1	N	DerivDem2	it	N/N
ImpP13M	t	V/V	suk'	*no field*	market**4	N	DerivDim	atN	N/N
ImpP13Mpf	l	V/V	surat	ሰረት	picture	N	DerivDim3	etaay	N/N
ImpfSg1pf	'	V/V	swaasw	*no field*	grammar**	N	DERIVEnum	t	N/N Num/Nur
ImpfSg2Fpf	t	V/V	swes'	*no field*	Swes	Nam	DerivFem	at	N/N Adj/Adj
ImpfSg2Mpf	t	V/V	syaasat	ሰያሰት	policy	N	DerivPauc	&taam	N/N
ImpfSg3Fpf	t	V/V	syaasi	ሰያሰ	political	Adj	DerivSg	aay	N/N V/N
ImpfSg3Mpf	l	V/V	s'aal	*no field*	question**2	N	ImpP11	0	V/V
JussP11pf	n	V/V	s'anat	*no field*	effort**1	N	ImpP12F	a	V/V
JussP13Fpf	l	V/V	s'id	*no field*	Said	Nam	ImpP12M	u	V/V
JussP13Mpf	l	V/V	s&lat	ሰ-ለት	tuberculosis	N	ImpP13F	a	V/V
JussSg1pf		V/V	t	*no field*	beImpf	V	ImpP13M	l	V/V
JussSg3Fpf	t	V/V	t'aaf	*no field*	Name+of+	N	ImpfSg1	0	V/V
JussSg3Mpf	l	V/V	t'aafh'	ጥፍት	accommodate	Adj	ImpfSg2F	i	V/V
NEG	'	V/V Adj/Adj	t'aagat	ጥገት	sbolt**2	N	ImpfSg2M	a	V/V
NEGImpv	t	V/V	t'aaha	*no field*	Taha	Nam	ImpfSg3F	a	V/V
p	p	N/N V/V	taajr	ታጅር	business+m	N	ImpfSg3M	0	V/V
PtcpAct	ma	V/N	taaki	ታኪ	vertical/strai	Adj	ImpvP12F	aa	V/V
PtcpPass	0	V/N	taamm	ታምም	enough**3	Adj	ImpvP12M	o	V/V
PtcpPassMS	u	V/N	taamm	ታምም	whole	Adj	ImpvSg2F	i	V/V
RelFron	la	V/V	taarik	*no field*	history**1	N	ImpvSg2M	0	V/V
			t'aawlat	ጥ* 1 4 5 ላ	table	N	JussP11	0	V/V
			t'aay	*no field*	dear**5	Intj	JussP13F	a	V/V
			t'aayh'uka	*no field*	dear**7	Intj	JussP13M	u	V/V
			t'aa'irat	ጥአረት	airplane/aer	N	JussSg1	0	V/V

The word classes

- There are 12 “categories” (classes) of roots
 - They are needed by the “rules” of the Tigre parser



The screenshot shows a software window titled "Language Model: Tigre *". The window has a menu bar with "Text", "Dictionary", "Lists", "Constraints", "Adaptation", and "Misc". The "Categories" item in the left sidebar is selected. The main area displays a table with two columns: "Label" and "Comment".

Label	Comment
??	Problems
Adj	Adjectives
Adv	Adverbs
Aux	Auxiliaries
Con	Conjunctions
Dem	Demonstratives
Intj	Interjections
N	Nouns
Nam	Names
Num	Numbers / Numerals
Pn	Pronouns
Prep	Prepositions / Postpositions
V	Verbs

Structure of a root entry

- Here is the “entry view” of a numeral:

- English Gloss

- ...

- Tigrigna

- ...

- Tigre

- Class

- Geez scr.

- Latin scr.

- ...

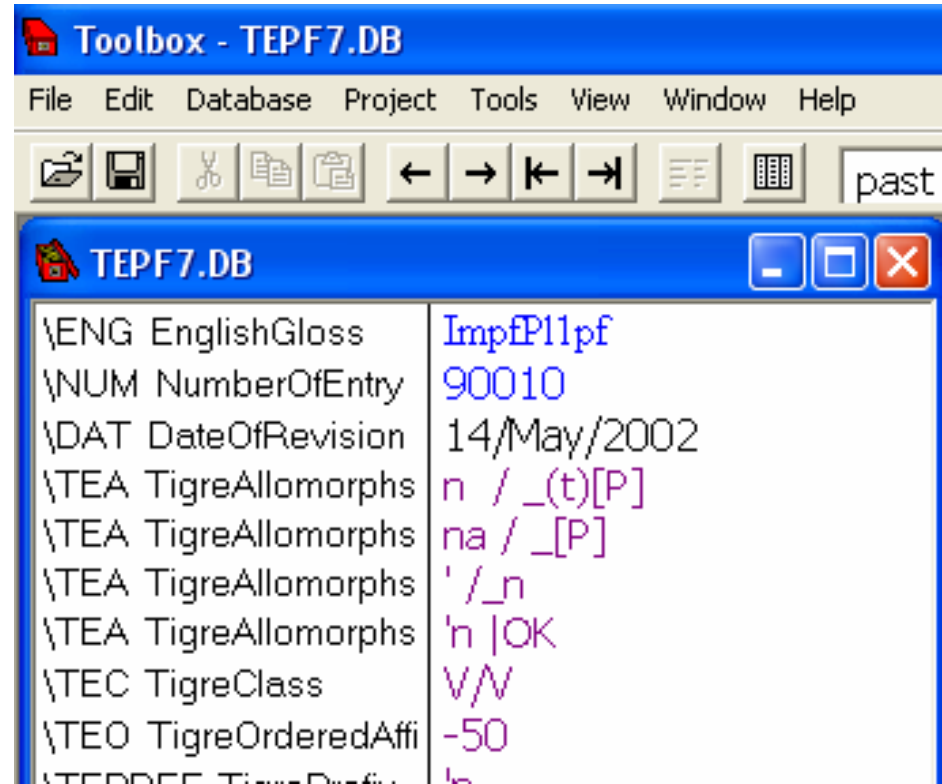
- Arabic ...

The screenshot shows a software window titled "Toolbox - [ABCJun2.db:1]" with a menu bar (File, Edit, Database, Project, Tools, View, Window, Help) and a toolbar. The main area displays a table with two columns: field names and their corresponding values. The field names are listed on the left, and the values are listed on the right. The values are color-coded: blue for English gloss, green for Tigrigna, purple for Tigre, and orange for Arabic.

\ENG EnglishGloss	sixty
\NUM NumberOfEntry	29111
\ENC EnglishClass	n
\TIC TigrignaClass	Num
\TIG TigrignaGeezScript	ሥሳ
\TIL TigrignaLatinScript	ssa
\TIR TigrignaRoot	ssa
\TIA TigrignaAllomorphs	ssa
\TIA TigrignaAllomorphs	susa
\TIP TigrignaP?	pl.none
\TIP TigrignaP?	no-gender
\TEC TigreClass	Num
\TEG TigreGeezScript	ሥሳ
\TEL TigreLatinScript	ssa
\TEA TigreAllomorphs	ssa
\ARB ArabicB?	ستون
\ARBEXP ArabicExplanation	= القعد السادس
\ARBEXP ArabicExplanation	(من الممر أو القرن)
\FND FoundInList	= AE 16978A
\FRQ Frequency	#3TGI
\THS ThesaurusNumber	OneTwoThreeCardinals
\DAT DateOfRevision	06/Nov/2000

Structure of a prefix entry

- Here is the “entry” view of a verb prefix “we”:
 - English Gloss
 - Impf 1st Ps Plural Prefix
 - Number
 - Date
 - Allomorphs
 - n
 - na
 - '
 - Class conditions
 - Prefix order

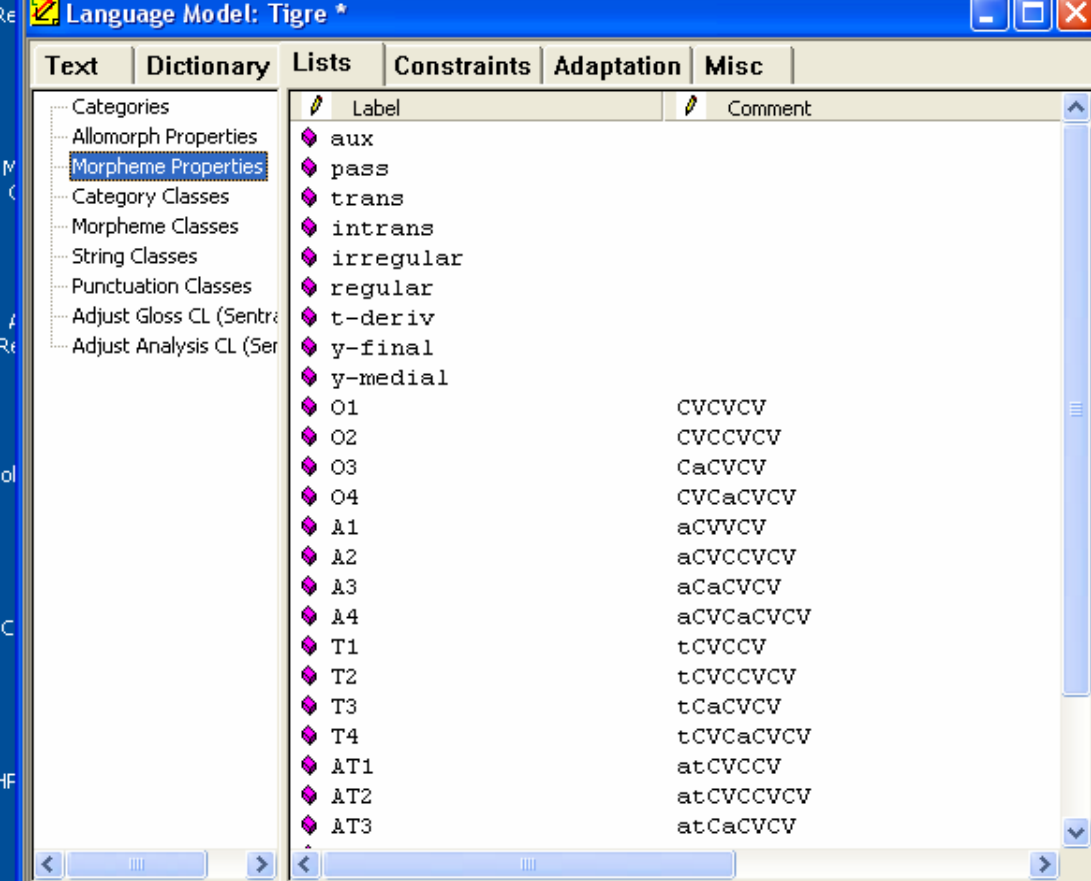


The verb system

- The most demanding aspect in the preparation of the parser is the verb system, especially its harmonization with the verb systems of *Arabic*, *Tigrigna* and *Beja*.
- Presently, the parser of the Tigre verb system works on the assumption that there are 9 main verb classes with several subclasses.

Cross-classification by properties

- The verb classes are defined by their “properties”:
 - auxiliaries
 - passives
 - intransitives
 - ...
 - y-final
 - ...
 - class 01 CVCVCV
 - class 02 CVCCVCV
 - class 03 CaCVCV
 - class 03 CVCaCVCV
 - class A1 aCVVCV etc.



The screenshot shows a software window titled "Language Model: Tigre" with a menu bar containing "Text", "Dictionary", "Lists", "Constraints", "Adaptation", and "Misc". The "Lists" tab is active, displaying a table with two columns: "Label" and "Comment". The table lists various verb classes and their corresponding properties.

Label	Comment
aux	
pass	
trans	
intrans	
irregular	
regular	
t-deriv	
y-final	
y-medial	
O1	CVCVCV
O2	CVCCVCV
O3	CaCVCV
O4	CVCaCVCV
A1	aCVVCV
A2	aCVCCVCV
A3	aCaCVCV
A4	aCVCaCVCV
T1	tCVCCV
T2	tCVCCVCV
T3	tCaCVCV
T4	tCVCaCVCV
AT1	atCVCCV
AT2	atCVCCVCV
AT3	atCaCVCV

Some verbs in Latin transliteration

- Here are some verbs from a list of verbs in Latin transliteration, with English glosses:
 - Ṣak`a *work*
 - Ṣarh`a *describe/advertise*
 - &aaba *make+grow*
 - &aarafa *rest*
 - &ac`da *chaff*
 - &agna *curve*
 - ...

A verb entry with its allomorphs

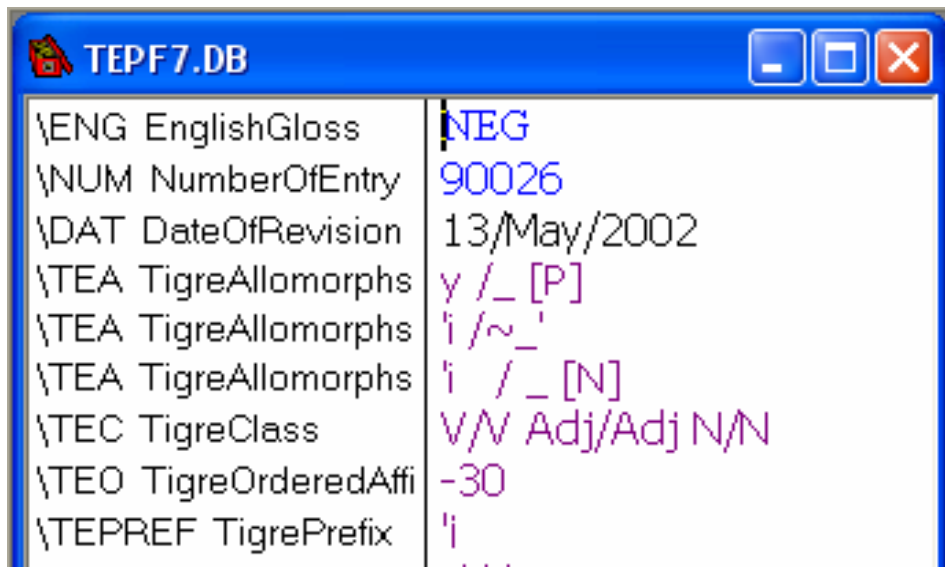
- Here follows a typical verb entry with verb allomorphs (rather than skeleton plus infixes)
- “Allomorphs” (TEA) are set up for the different “aspects” or “tenses” like “perf”, “impf” etc:
 - \ENG accomplish/finish
 - \TEG ገደደ | Geez scr.
 - \TEL wada | Latin scr.
 - \TEA wad perf | 4 Tigre [Allomorphs](#) for perfect etc.
 - \TEA wad impf
 - \TEA wd impv

Results of the parsing project

- The rigor of machine parsing has led to a few refinements of existing analyses.
 - One of them is Saleh Mahmud's discovery of the assimilation of 3rd person prefixes before laryngeals (an areal feature also valid for Beja).
 - Here follows an entry illustrating the phonological rule:

Phonology of pharyngeals

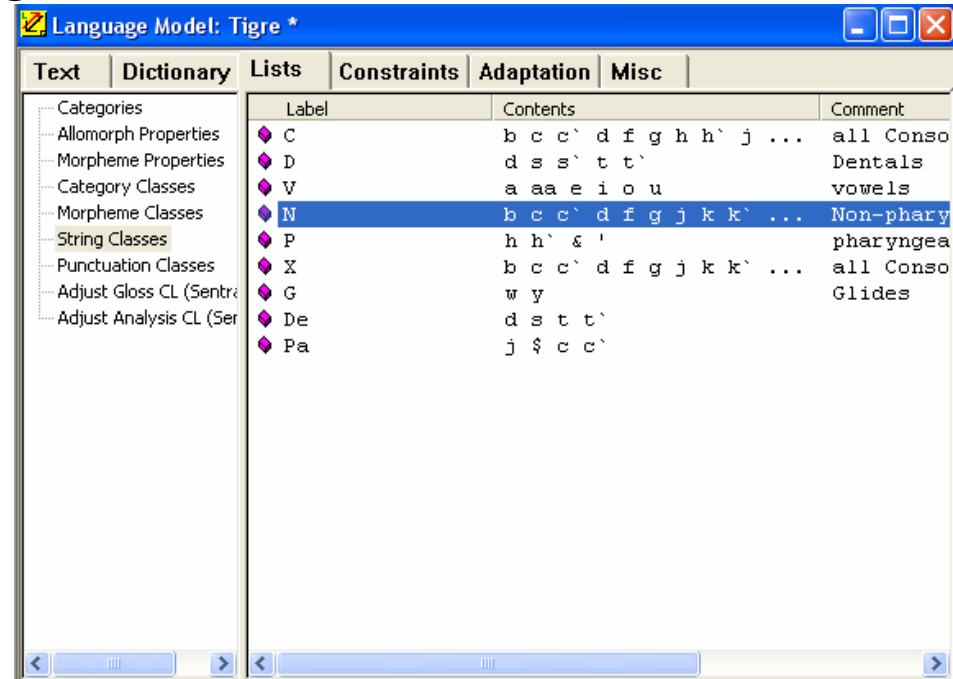
- The prefix “Negation” (NEG) has the following Tigre allomorphs (TEA):
 - y before Pharyngeals [P]
 - ‘i not before ‘ | i.e. glottal stop/hamza
 - ‘i before Non-Pharyngeals [N]



Field	Value
\ENG EnglishGloss	NEG
\NUM NumberOfEntry	90026
\DAT DateOfRevision	13/May/2002
\TEA TigreAllomorphs	y /_ [P]
\TEA TigreAllomorphs	'i /~_'
\TEA TigreAllomorphs	'i /_ [N]
\TEC TigreClass	V/N Adj/Adj N/N
\TEO TigreOrderedAffi	-30
\TEPREF TigrePrefix	'i

Sample phoneme classes

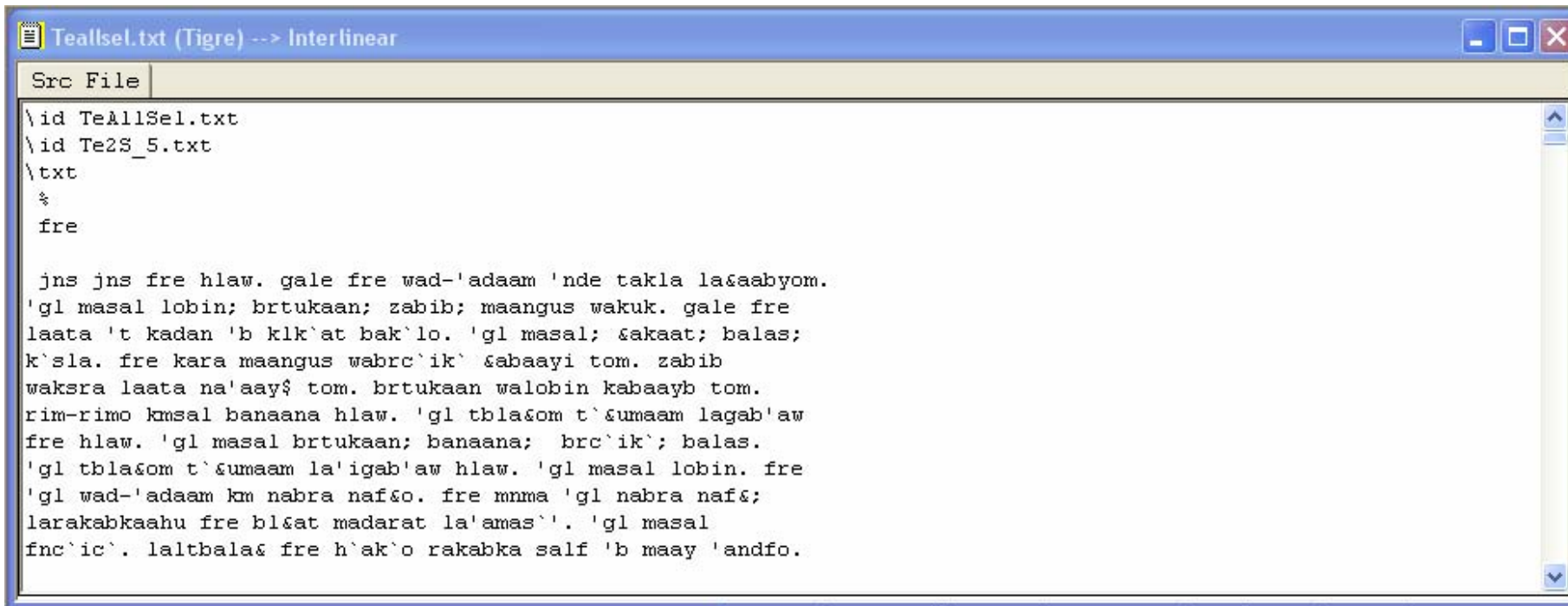
- Several phoneme classes had to be established:
 - C consonants
 - ...
 - N Non-pharyngeals
 - P Pharyngeals



Text	Dictionary	Lists	Constraints	Adaptation	Misc
	Categories				
	Allomorph Properties				
	Morpheme Properties				
	Category Classes				
	Morpheme Classes				
	String Classes				
	Punctuation Classes				
	Adjust Gloss CL (Sentra				
	Adjust Analysis CL (Ser				
		Label	Contents		Comment
		◆ C	b c c` d f g h h` j ...		all Conso
		◆ D	d s s` t t`		Dentals
		◆ V	a aa e i o u		vowels
		◆ N	b c c` d f g j k k` ...		Non-phary
		◆ P	h h` & `		pharyngea
		◆ X	b c c` d f g j k k` ...		all Conso
		◆ G	w y		Glides
		◆ De	d s t t`		
		◆ Pa	j \$ c c`		

Parsing of a sample text

- The parsing of texts is based on the Latin transcription of the original “fidel” text – i.e. gemination is disregarded. Here is a sample:



```
Teallsel.txt (Tigre) --> Interlinear
Src File
\id TeAllSel.txt
\id Te2S_5.txt
\txt
%
fre

jns jns fre hlaw. gale fre wad-'adaam 'nde takla la&aabyom.
'gl masal lobin; brtukaan; zabib; maangus wakuk. gale fre
laata 't kadan 'b klk`at bak`lo. 'gl masal; &akaat; balas;
k`sla. fre kara maangus wabrc`ik` &abaayi tom. zabib
waksra laata na'aay& tom. brtukaan walobin kabaayb tom.
rim-rimo kmasal banaana hlaw. 'gl tbla&om t`&umaam lagab'aw
fre hlaw. 'gl masal brtukaan; banaana; brc`ik`; balas.
'gl tbla&om t`&umaam la'igab'aw hlaw. 'gl masal lobin. fre
'gl wad-'adaam km nabra naf&o. fre mnma 'gl nabra naf&;
larakabkaahu fre bl&at madarat la'amas`. 'gl masal
fnc`ic`. laltbala& fre h`ak`o rakabka salf 'b maay 'andfo.
```

Results of parsing: initial failures

- Words which fail to parse are listed in a LOG file as “Analysis Failures” (AF)
 - The statistics for this text of 7105 words showed a result of 6045 successful analyses, 1060 failures, and 36% ambiguities.

```
Teallsel.txt (Tigre) --> Interlinear
Src File | Phonrule Log | Ample ANA | AmpleDLL Log | ST-Disamb ANA | SENT Disambig Log | PrintANA Log | Interlinear
AMPLE Message:
  ROOT DICTIONARY: Loaded 4399 records

AMPLE: A Morphological Parser for Linguistic Exploration
Version 3.6.5 (October 17, 2002), Copyright 2002 SIL, Inc.
Compiled Oct 21 2002 09:35:21

Analysis Performed Sat Feb 02 17:27:05 2008

Input file: C:\Carla\TeTxt\Teallsel.txt
Output file: C:\DOCUME~1\Wedekind\LOCALS~1\Temp\CSTUDIO\Tigre0\Teallsel-ample.ana
AF: tbla&om [ tbla&om | ]
AF: tbla&om [ tbla&om | ]
AF: laltbala& [ lalt | bala& ]
AF: tbla&om [ tbla&om | ]
AF: 'at`raafka [ 'at | `raafka ]
AF: tbla&om [ tbla&om | ]
```

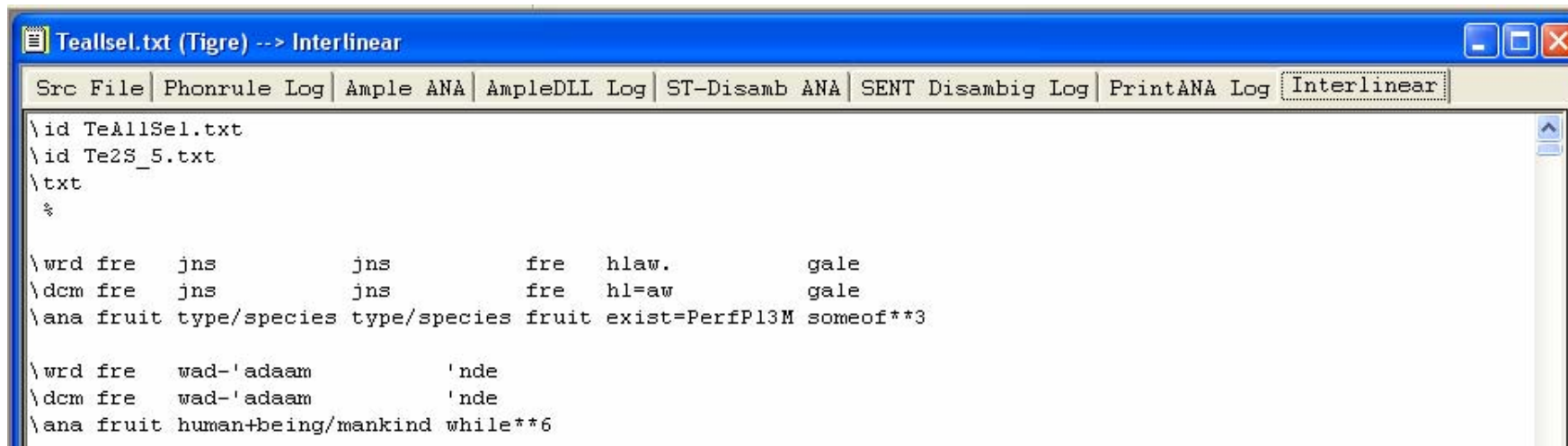
Analysis display

- The analysis of each word is displayed in this pattern:
 - analysis by class and gloss (<N fruit>)
 - category (N noun)
 - property (sg singular)
 - word (phonological form “fre”)

```
\a < N type/species >  
\d jns  
\cat N N  
\p sg  
\w jns  
  
\a < N fruit >  
\d fre  
\cat N N  
\p  
\w fre
```

Interlinear display of the analysis

- An interlinear text is produced
 - \wrd “word” \dcm “decomposition” \ana “analysis”
- Note, for example, the parsing of “**hlaw**”:
 - hl=aw
 - exist=PerfPl3M (third person plural masculine)



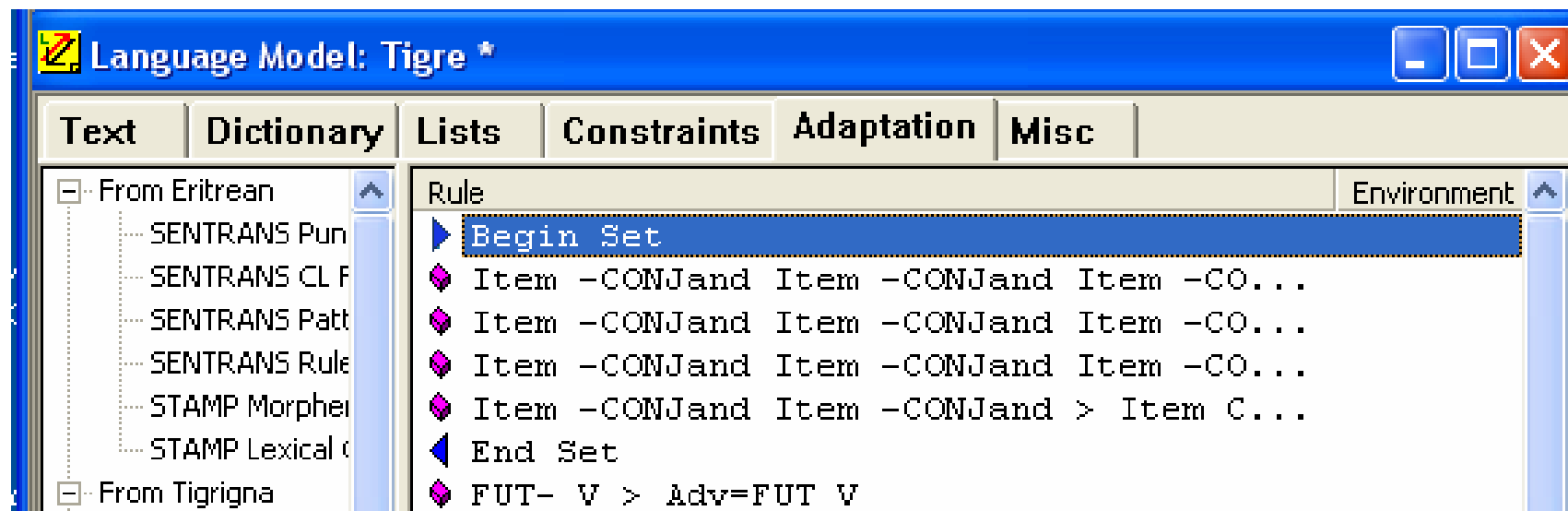
```
\id TeAllSel.txt
\id Te2S_5.txt
\txt
%

\wrd fre   jns           jns           fre   hlaw.           gale
\dcm fre   jns           jns           fre   hl=aw          gale
\ana fruit type/species type/species fruit exist=PerfPl3M someof**3

\wrd fre   wad-'adaam      'nde
\dcm fre   wad-'adaam      'nde
\ana fruit human+being/mankind while**6
```

Sample syntax rules

- Finally, it should be noted that syntactic and substitution rules have the form shown as below:
 - First line:
 - The Tigrigna suffix “and” is changed to a Tigre prefix
 - Last line:



References and data bases

- The main reference works were the following:
 - Hoefner, Maria, and Enno Littmann, 1965, „Woerterbuch der Tigre-Sprache“, Wiesbaden
 - Nakano, Aki'o, 1982, “A Vocabulary of Beni Amer Dialect of Tigre”, Tokyo
 - Raz, Shlomo, 1984, “Tigre Grammar and Texts“, Malibu
- Access to the Tigre data base can probably be arranged through Saleh Mahmud.
 - **THE END**