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 ሉ Tigre lu
    - ASCII 70 is ል 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	ሀ
l	ለ
h`	ሐ
$	ሸ
k`	ቀ
'	ኀ
&	ዕ
t`	ጥ

Transliteration of vowels

Vow.	
-a	Ս
-u	ՍՒ
-aa	Կ
-i	Կ
-e	Կ
	Ս
-o	ՍՐ

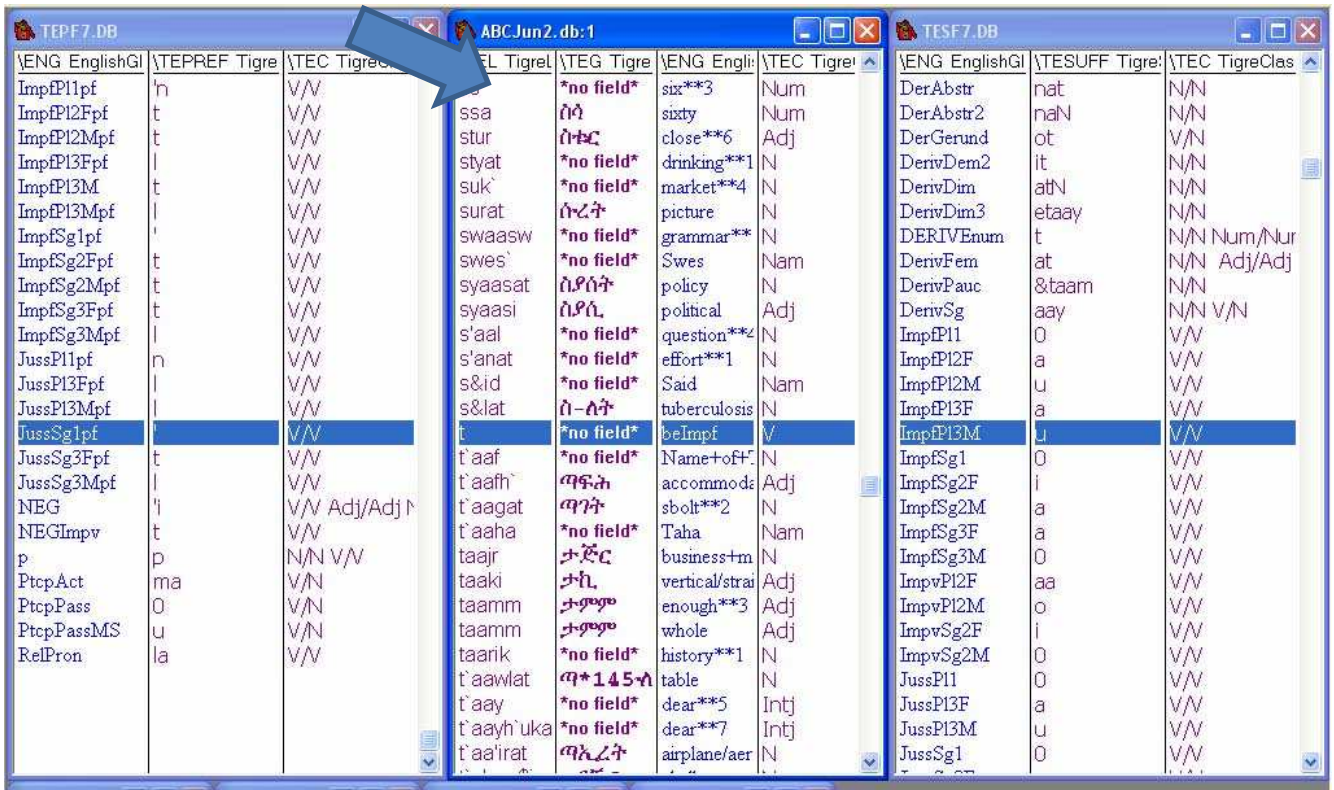
Statistics of the Data Bases

- 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

\ENG EnglishGl	\TEPREF Tigre	\TEC Tigre
ImpfP11pf	'n	V/V
ImpfP12Fpf	t	V/V
ImpfP12Mpf	t	V/V
ImpfP13Fpf	I	V/V
ImpfP13M	t	V/V
ImpfP13Mpf	I	V/V
ImpfSg1pf	'	V/V
ImpfSg2Fpf	t	V/V
ImpfSg2Mpf	t	V/V
ImpfSg3Fpf	t	V/V
ImpfSg3Mpf	I	V/V
JussP11pf	n	V/V
JussP13Fpf	I	V/V
JussP13Mpf	I	V/V
JussSg1pf	'	V/V
JussSg3Fpf	t	V/V
JussSg3Mpf	I	V/V
NEG	'i	V/V Adj/Adj N
NEGImpv	t	V/V
p	p	N/N V/V
PtcpAct	ma	V/N
PtcpPass	0	V/N
PtcpPassMS	u	V/N
RelPron	la	V/V

ABCJun2.db:1

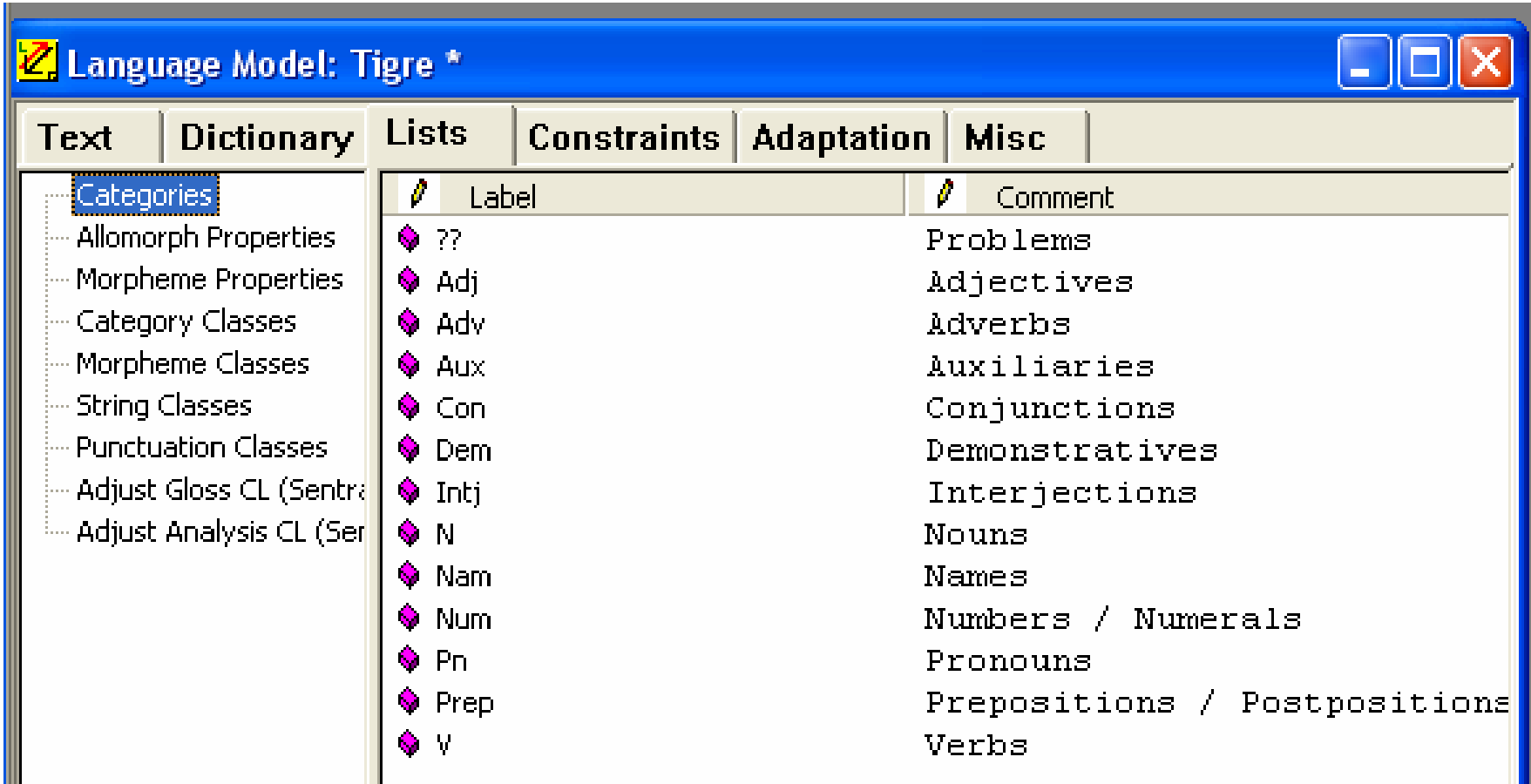
FL TigreL	\TEG Tigre	\ENG Engli	\TEC Tigre
	no field	six**3	Num
ssa	ስሳ	sixty	Num
stur	ስቱር	close**6	Adj
styat	*no field*	drinking**1	N
suk`	*no field*	market**4	N
surat	ሱረት	picture	N
swaasw	*no field*	grammar**	N
swes`	*no field*	Swes	Nam
syaasat	ስያሳት	policy	N
syaasi	ስያሲ	political	Adj
s'aal	*no field*	question**4	N
s'anat	*no field*	effort**1	N
s&id	*no field*	Said	Nam
s&lat	ስ-ለት	tuberculosis	N
t	*no field*	beImpf	V
t`aaf	*no field*	Name+of+T	N
t`aafh`	ጣፍሕ	accommoda	Adj
t`aagat	ጣገት	sbolt**2	N
t`aaha	*no field*	Taha	Nam
taajr	ታጅር	business+m	N
taaki	ታኪ	vertical/strai	Adj
taamm	ታምም	enough**3	Adj
taamm	ታምም	whole	Adj
taarik	*no field*	history**1	N
t`aawlat	ጣ*145ስ	table	N
t`aay	*no field*	dear**5	Intj
t`aayh`uka	*no field*	dear**7	Intj
t`aa'irat	ጣኢረት	airplane/aer	N

TESF7.DB

\ENG EnglishGl	\TESUFF Tigre	\TEC TigreClas
DerAbstr	nat	N/N
DerAbstr2	naN	N/N
DerGerund	ot	V/N
DerivDem2	it	N/N
DerivDim	atN	N/N
DerivDim3	etaay	N/N
DERIVEnum	t	N/N Num/Nur
DerivFem	at	N/N Adj/Adj
DerivPauc	&taam	N/N
DerivSg	aay	N/N V/N
ImpfP11	0	V/V
ImpfP12F	a	V/V
ImpfP12M	u	V/V
ImpfP13F	a	V/V
ImpfP13M	u	V/V
ImpfSg1	0	V/V
ImpfSg2F	i	V/V
ImpfSg2M	a	V/V
ImpfSg3F	a	V/V
ImpfSg3M	0	V/V
ImpvP12F	aa	V/V
ImpvP12M	o	V/V
ImpvSg2F	i	V/V
ImpvSg2M	0	V/V
JussP11	0	V/V
JussP13F	a	V/V
JussP13M	u	V/V
JussSg1	0	V/V

The word classes

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



Structure of a root entry

- Here is the “entry view” of a numeral:
  - English Gloss
    - …
  - Tigrigna
    - …
  - Tigre
    - Class
    - Geez scr.
    - Latin scr.
    - …
  - Arabic …

Toolbox - [ABCJun2.db:1]

File Edit Database Project Tools View Window Help

past | TigreRootsFull

\ENG EnglishGloss	sixty
\NUM NumberOfEntry	29111
\ENC EnglishClass	n
\TIC TigrignaClass	Num
\TIG TigrignaGeezScript	ስሳ
\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	ስሳ
\TEL TigreLatinScript	ssa
\TEA TigreAllomorphs	ssa
\ARB ArabicB?	ستون
\ARBEXP ArabicExplanation	= العقد السادس
\ARBEXP ArabicExplanation	(من العمر 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

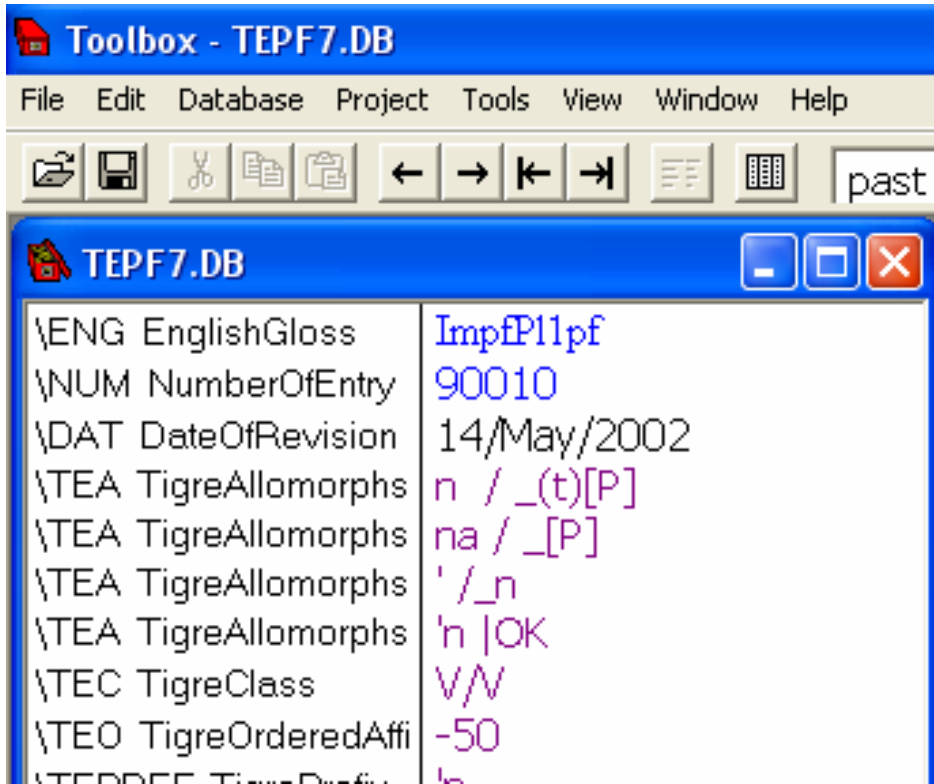
    n

    na

    ‘

    Class conditions

    Prefix order

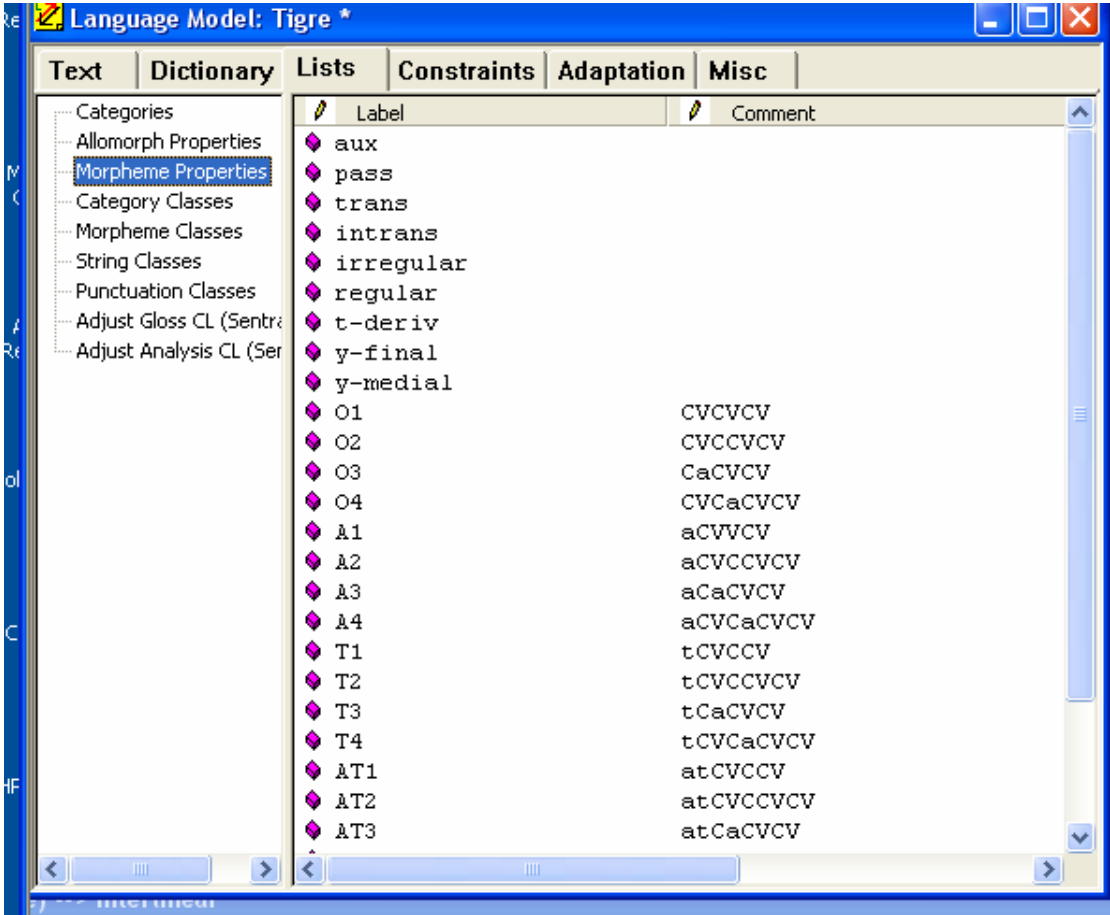


The verb system

- 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.



Some verbs in Latin transliteration

- Here are some verbs from a list of verbs in Latin transliteration, with English glosses:
  - $ak`a *work*
  - $arh`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 ወደ |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

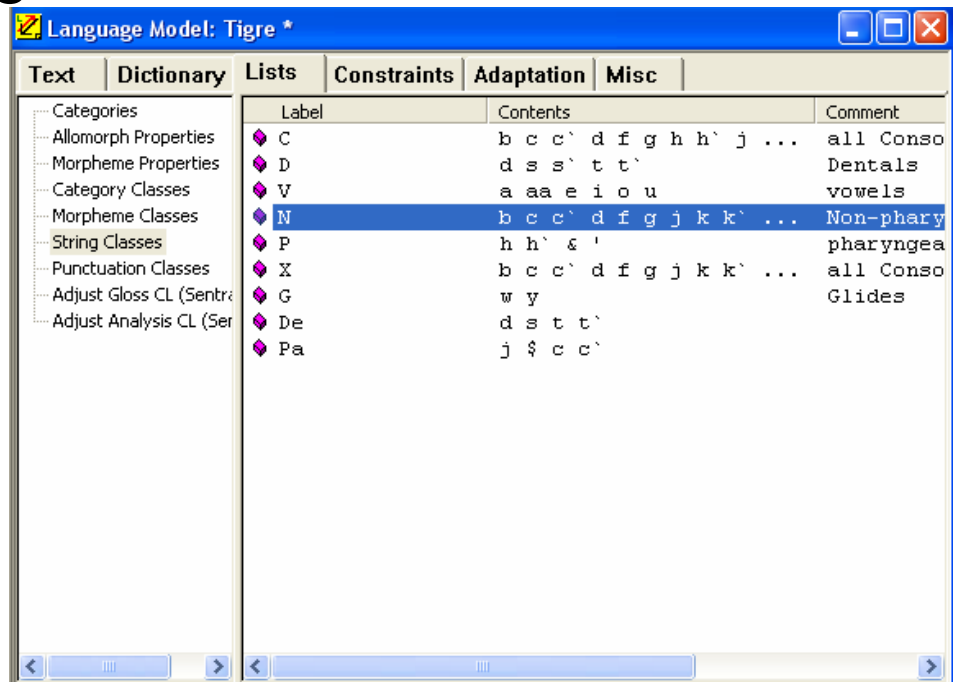
  ‘i before Non-Pharyngeals [N]

TEPF7.DB

\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/V Adj/Adj N/N
\TEO TigreOrderedAffi	-30
\TEPREF TigrePrefix	'i

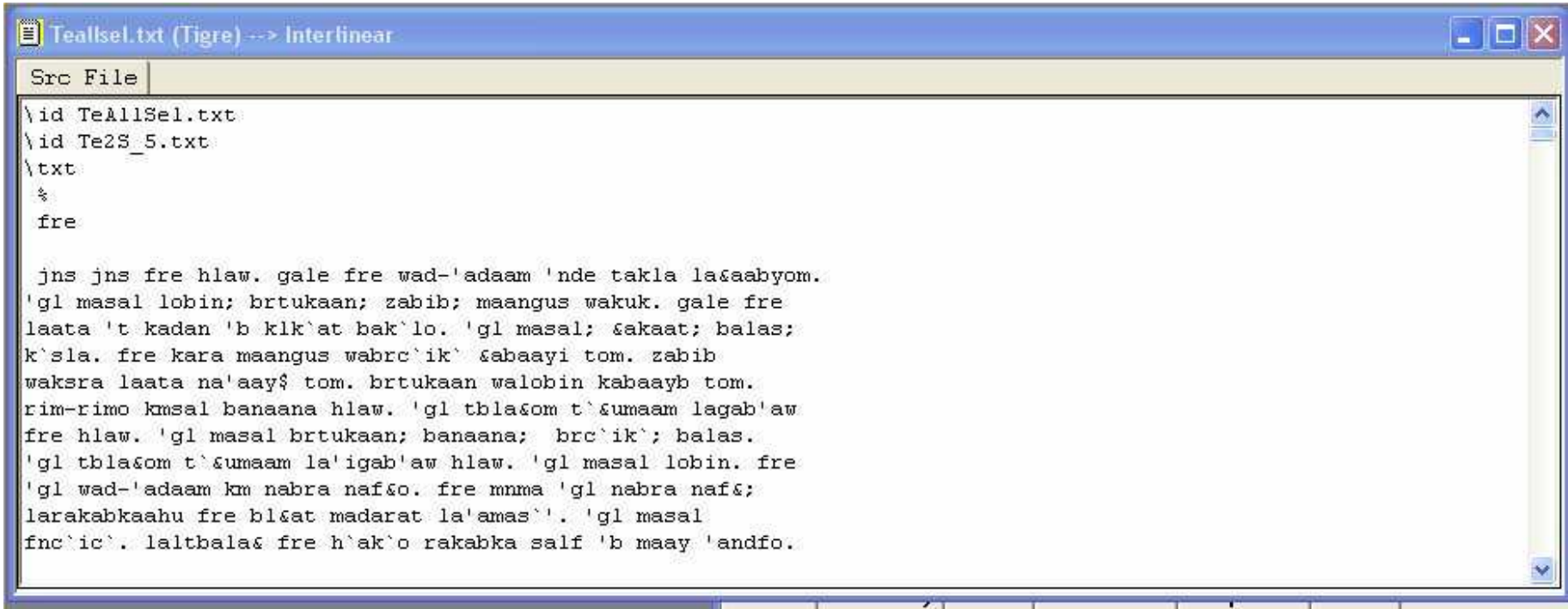
Sample phoneme classes

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



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:



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\TigreO\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)

TeAllSel.txt (Tigre) --> Interlinear

Src File | Phonrule Log | Ample ANA | AmpleDLL Log | ST-Disamb ANA | SENT Disambig Log | PrintANA Log | Interlinear

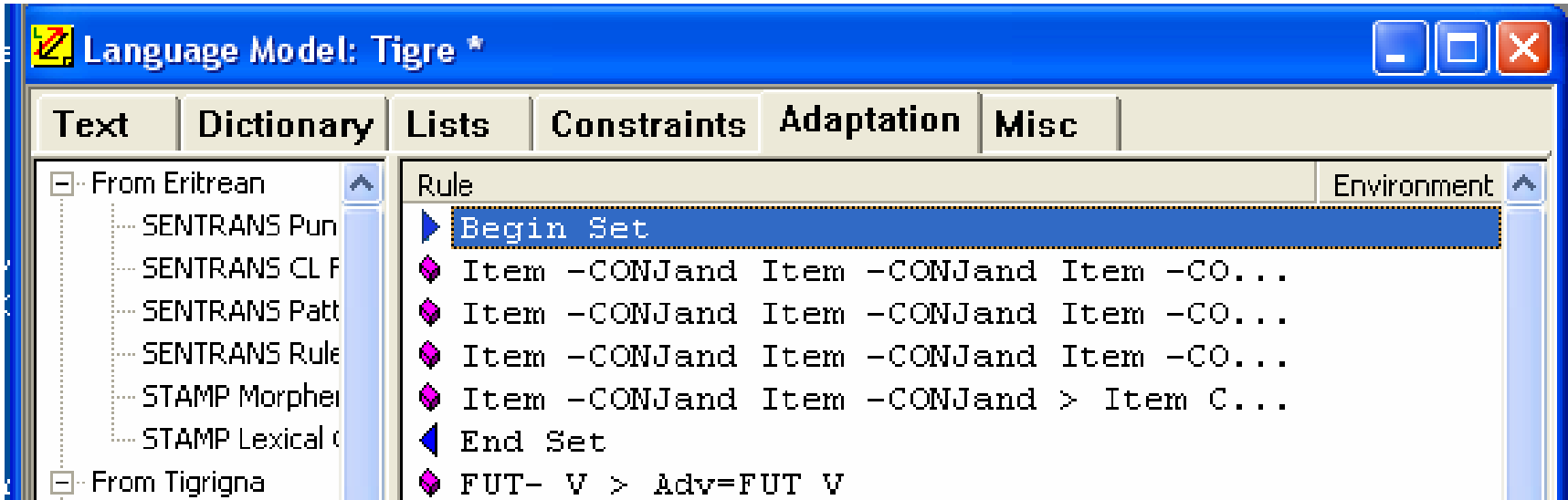
```
\id TeAllSel.txt
\id Te2S_5.txt
\txt
 %

\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