Spoken Corpus Mozambique

1 BASIC INFORMATION

1.1 Corpus composition

The Spoken Corpus Mozambique contains approximately 121,958 running words of spoken Portuguese from Mozambique. It includes 40 transcriptions of spoken recordings (in a total of 40 hours of recordings) that were recorded between 1986 and 1987 (Annex 1, below, shows the informant's metadata).

1.2 Representation of the corpora (flat files, database, markup)

The file formats of this corpus are txt and cqpweb.

1.3 Character encoding

The characters have been encoded in UTF-8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Amália Mendes
Address: Complexo Interdisciplinar da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal
Affiliation: Centro de Linguística da Universidade de Lisboa
Position: Researcher
Telephone: +351 21 790 47 00
Fax: + 351 21 796 56 22
e-mail: amalia.mendes@clul.ul.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be available on the MetaShare platform.

2.3 Copyright statement and information on IPR

The resource is free licence-based for research purposes and free licence-based for commercial purposes. It is planned to be distributed under a MetaShare Commons BY SA license.

3 TECHNICAL INFORMATION

3.1 Directories and files

The Spoken Corpus Mozambique is composed by by a text file (corpus) and a cqpweb file (corpus with annotation).

3.2 Data structure of an entry

The txt version has one sentence per line, an identification number for each text and no further annotation. The cqpweb file has one token per line, followed by PoS tag and lemma, and is annotated for NP chunks.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

The corpus contains 121,958 tokens (???? sentences and ???? noun phrases) and needs about 659 KB for disk storage for the text file and about ???? for the cqpweb file.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is Portuguese from Mozambique.

4. 3 Domain(s)/register(s) of the corpus

This is a spoken corpus, recorded in a situation of spontaneous oral communication, on different themes of everyday life, with speakers of different ages and social and professional backgrounds, aiming at collecting the real Portuguese spoken in Mozambique.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

The corpus is PoS-annotated at token level, including punctuation. Noun phrases were recognized and annotated with specific tags.

4.4.2 Tags (if POS/MSD/TIME/discourse/etc –tagged or parsed),

The corpus was automatically PoS-tagged with MBT tagger (http://ilk.uvt.nl/mbt/), and lemmatized with MBLEM (http://ilk.uvt.nl/mbma/), following the annotation scheme of the

Corpus of Reference of Contemporary Portuguese (Généreux et al., 2012). YamCha software (http://chasen.org/~taku/software/yamcha/) was used to recognize chunks that consist of noun phrases and identifies the elements that are in the beginning, in the middle and in the end of a noun phrase.

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

n.a

4.4.4 Attributes and their values (if annotated)

The following tags were applied in the POS-tagging:

POS codification	
Tag	**Category**
ADJ	Adjectives
ADV	Adverbs
CARD	Cardinals
CJ	Conjunctions
CL	Clitics
CN	Common Nouns
DA	Definite Articles
DEM	Demonstratives
DFR	Denominators of Fractions
DGTR	Roman Numerals
DGT	Digits
DM	Discourse Marker
EADR	Electronic Addresses
EOE	End of Enumeration

EXC	Exclamatives
GER	Gerunds
GERAUX	Gerunds as auxiliary verbs
IA	Indefinite Articles
IND	Indefinites
INF	Infinitive
INFAUX	Infinitive auxiliary verb
INT	Interrogatives
ITJ	Interjection
LTR	Letters
LADV1…LADVn	Latin Multi-Word Adverbs
MGT	Magnitude Classes
MTH	Months
NP	Noun Phrases
ORD	Ordinals
PADR	Part of Address
PNM	Part of Name
PNT	Punctuation Marks
POSS	Possessives
PPA	Past Participles not in compound tenses
PP	Prepositional Phrases
PPT	Past Participle in compound tenses
PREP	Prepositions
PRS	Personals

QNT	Quantifiers
REL	Relatives
STT	Social Titles
SYB	Symbols
TERMN	Optional Terminations
UM	"um" or "uma"
UNIT	Measurement units in abbreviated form
VAUX	Finite "ter" or "haver" in compound tenses
V	Verbs (other than PPA, PPT, INF or GER)
WD	Week Days
LADV1…LADVn	Multi-Word Adverbs
Contracted forms	**Combinations of :**
CL+CL	Two clitics
PREP+ADV	Preposition and Adverb
PREP+DA	Preposition and Definite Articles
PREP+DEM	Preposition and Demonstratives
PREP+IND	Preposition and Indefinite
PREP+INT	Preposition and Interrogative
PREP+PRS	Preposition and Personal pronoun
PREP+QNT	Preposition and Quantifier
PREP+REL	Preposition and Relative
PREP+UM	Preposition and "um" or "uma"

The following tags were applied in the NP chunker:

Position	Description
B-NP	Beginning
I-NP	Inside
E-NP	End
O	Outside

4.5 Intended application of the corpus

The corpus can be either used in linguistic research and as a teaching and learning material.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The POS-tagging and NP chunker were done automatically.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Gonçalves, M. P. (1990), *A construção de uma gramática de português em Moçambique: aspectos da estrutura argumental dos verbos*, Dissertação de Doutoramento em Linguística Portuguesa apresentada à Faculdade de Letras da Universidade de Lisboa (com Anexo).

ANNEX1 - Informant's metadata

Informant Code	Age	Sex	L1	Practice of L1	Poruguese learning's age	Date of the recording
AGO	21	M	chope	S	5	1986
ALE	20	M	sena	N	5	1986
AND	19	M	"cindau"	S	9	1986
ARI	21	F	"tsonga"	S	6	1986
AUG	18	M	"xitshwa"	N	5	1986
AZA	31	M	"tsonga"	S	9	1987
BEL	20	M	"tsonga"	S	5	1987
CAM	20	M	"tsonga"	S	5	1987
CHO	21	M	"cindau"/sena	N	9	1986
COS	22	M	"tsonga"	S	7/8	1987
CUN	23	M	"tsonga"	S	5	1987
DEZ	21	M	sena	S	9	1986
DOR	19	M	"cindau"	S	8	1986
DUM	21	M	"tsonga"	N	6	1987
FEL	29	M	"cindau"	S	8	1986
GAB	19	M	"manyika"	S	6	1987
GUN	22	M	suaili	N	8	1986
IDA	21	F	"tsonga"	N	5	1987
JAM	31	F	português	S	-	1987
JOR	20	M	"tsonga"	N	12	1986
LIM	18	M	maconde	N	7	1986

LUI	20	M	“cindau”	N	9	1986
MAN	20	M	chope	S	7	1986
MEQ	20	M	“cinyungwe”	N	10	1986
MIL	26	M	“echuwabo”	S	5	1986
MON	18	M	“tsonga”	S	7	1986
MUL	31	M	“tsonga”	S	5/6	1986
NAT	29	M	“tsonga”	S	8/9	1986
OAL	20	F	português	S	-	1987
PAS	20	M	“xitshwa”	S	5	1986
PRE	20	M	“tsonga”	S	5	1987
SAM	210	M	“xitshwa”	S	5	1986
SAU	20	M	sena	S	6	1986
SEV	29	M	macua	N	15	1986
SIB	21	M	“tsonga”	S	5	1987
TAN	30	M	macua	N	7	1986
TEM	20	F	“tsonga”	N	6	1986
UEL	20	M	“xitshwa”	S	8	1986
WAI	19	M	macua	S	8	1986
ZIT	32	M	“tsonga”	S	10	1986