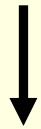
Building Resources: Experiences from Amharic Cross Language Information Retrieval

Lars Asker
Stockholm University

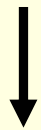
A simple approach to CLIR

Query (in source language)



translation

Keywords (in target language)



retrieval

Retrieved documents

Naïve Translation

- Word by word
- Dictionary lookup
- Disambiguation
- Stopword removal

What if there's no Dictionary? (or other resources)

Approaches to Dictionary Construction

- From parallel electronic corpora
- From printed dictionaries using OCR
- From soft copies of dictionaries

From parallel corpora

- The Bible
  - Old fashioned language
  - Too small
- Aligned new articles
  - Fuzzy alignment
  - Too small

የማቴዎስ ወንጌል

የኢየሱስ የትውልድ ሐረግ

1:1-17 ተን ምብ - ሉቃ 3:23-38
1:3-6 ተን ምብ - ሩት 4:18-22
1:7-11 ተን ምብ - 1ዜና 3:10-17

1 የዳዊት ልጅ፣ የአብርሃም ልጅ የሆ
ነው የኢየሱስ ክርስቶስ የትውልድ ሐረግ
የሚከተለው ነው፤

2አብርሃም ይስሐቅን ወለደ፤
ይስሐቅ ያዕቆብን ወለደ፤
ያዕቆብ ይሁዳንና ወንድሞቹን ወለደ፤
3ይሁዳ ከትዕማር ፋሬስንና ዛራን ወለደ፤
ፋሬስ ኤስሮምን ወለደ፤
ኤስሮምም አራምን ወለደ፤
4አራም አሚናዳብን ወለደ፤
አሚናዳብ ነአሶንን ወለደ፤
ነአሶን ሰልሞንን ወለደ፤
5ሰልሞን በዔዝን ከረዓብ ወለደ፤
በዔዝ ከሩት ኢዮቤድን ወለደ፤
ኢዮቤድ እሴይን ወለደ፤
6እሴይ ንጉሥ ዳዊትን ወለደ።

1፥1 ዘፍ22፥18፤ 2ሳሙ7፥12-16፤ኢሳ 9፥6-7፤11፥1፤ኤር23፥ 5-6፤ሚክ5፥2፤ ሉቃ 1፥32-69፤ ገላ3፥16፤ ሥራ22፥16
1፥2 ዘፍ21፥3፤12፤ 25፥26፤29፥35፤ 49፥10
1፥3 ዘፍ38፥27-30
1፥5 ሩት1፥31
1፥6 1ሳሙ-16፥1፤ 17፥12፤2ሳሙ-12፥24
1፥10 2ነገ20፥21
1፥11 2ነገ24፥14-16 ፤ኤር27፥20፤ 40፥1፤ዳን1፥1-2
1፥12 1ዜና3፥17፤ 19፤ዕዝ3፥2

ማታን ያዕቆብን ወለደ፤
16ያዕቆብ ዮሴፍን ወለደ፤
ዮሴፍም የኢየሱስ ክርስቶስ እናት
የማርያም እጮኛ ነበር።

17እንግዲህ ከአብርሃም እስከ ዳዊት ዐሥራ አራት ትውልድ፣ ከዳዊት እስከ ባቢሎን ምርኮ ዐሥራ አራት ትውልድ፣ ከባቢሎን ምርኮ እስከ ክርስቶስ ልደት ዐሥራ አራት ትውልድ ይሆናል።

የኢየሱስ ክርስቶስ ልደት

18የኢየሱስ ክርስቶስ የልደት ታሪክ እን ዲህ ነው፤ እናቱ ማርያም ለዮሴፍ ታጭታ ሳይገናኙ፣ በመንፈስ ቅዱስ ፀንሳ ተገኘች። 19እጮኛዋ ዮሴፍ ጻድቅ ሰው ስለ ነበርና ማርያምን በሰው ፊት ሊያጋልጣት ስላል ፈለገ በስውር ሊተዋት መሰነ።

20በዚህ ሐሳብ ሳለ፣ የእግዚአብሔር መልአክ በሕልም ተገለጠለት፤ እንዲህም አለው፤ "የዳዊት ልጅ ዮሴፍ ሆይ፤ እጮ

<seg id="b.MAT.1.1" type=verse> **The book of the generation of Jesus Christ, the son of David, the son of Abraham.** </seg>

<seg id="b.MAT.1.2" type=verse> **Abraham begat Isaac; and Isaac begat Jacob; and Jacob begat Judas and his brethren;** </seg>

<seg id="b.MAT.1.3" type=verse> **And Judas begat Phares and Zara of Thamar; and Phares begat Esrom; and Esrom begat Aram;** </seg>

<seg id="b.MAT.1.4" type=verse> **And Aram begat Aminadab; and Aminadab begat Naasson; and Naasson begat Salmon;** </seg>

<seg id="b.MAT.1.5" type=verse> **And Salmon begat Booz of Rachab; and Booz begat Obed of Ruth; and Obed begat Jesse;** </seg>

<seg id="b.MAT.1.6" type=verse> **And Jesse begat David the king; and David the king begat Solomon of her that had been the wife of Urias;** </seg>

<seg id="b.MAT.1.7" type=verse> **And Solomon begat Roboam; and Roboam begat Abia; and Abia begat Asa;** </seg>

ምዕራፍ አንድ
ጠቅላላ ድንጋጌዎች

አንቀጽ ፩
የኢትዮጵያ መንግሥት ስያሜ

፩/ ይህ ሕገ መንግሥት ፌዴራላዊና ዴሞክራሲያዊ የመንግሥት አወቃቀር ይደነግጋል ። በዚህ መሰረት የኢትዮጵያ መንግሥት የኢትዮጵያ ፌዴራላዊ ዴሞክራሲያዊ ሪፐብሊክ በሚል ስም ይጠራል ።

አንቀጽ ፪
የኢትዮጵያ የግዛት ወሰን

፩/ የኢትዮጵያ የግዛት ወሰን የፌዴራሉን አባሎች ወሰን የሚያጠቃልል ሆኖ በዓለም አቀፍ ስምምነቶች መሰረት የተወሰነው ነው።

አንቀጽ ፫
የኢትዮጵያ ሰንደቅ ዓላማ

፩/ የኢትዮጵያ ሰንደቅ ዓላማ ከላይ አረንጓዴ ፣ ከመሐል ቢጫ ፣ ከታች ቀይ ሆኖ በመሐሉ ብሔራዊ ዓርማ ይኖረዋል ። ሦስቱም ቀለማት እኩል ሆነው በአግድም ይቀመጣሉ ።

፪/ ከሰንደቅ ዓላማው ላይ የሚቀመጠው ብሔራዊ ዓርማ የኢትዮጵያ ብሔሮች ፣ ብሔረሰቦች ፣ ሕዝቦች እና ሃይማኖቶች በእኩልነትና በአንድነት ለመኖር ያላቸውን ተስፋ የሚያንጸባርቅ ይሆናል ።

፫/ የፌዴራሉ አባሎች የየራሳቸው ሰንደቅ ዓላማና ዓርማ ሊኖራቸው ይችላል ። ዝርዝሩን በየራሳቸው ምክር ቤት ይወስናሉ።

CHAPTER ONE

GENERAL PROVISIONS

Article 1

Nomenclature of the State

This Constitution establishes a Federal and Democratic State structure. Accordingly, the Ethiopian state shall be known as *The Federal Democratic Republic of Ethiopia.*

Article 2

Ethiopian Territorial Jurisdiction

The territorial jurisdiction of Ethiopia shall comprise the territory of the members of the Federation and its boundaries shall be as determined by international agreements.

Article 3

The Ethiopian Flag

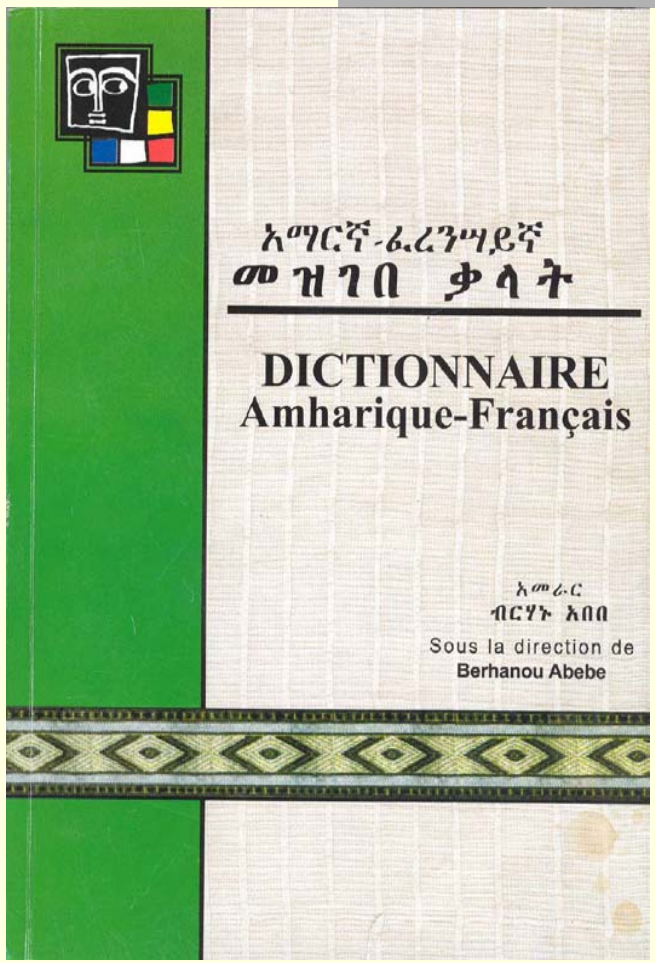
1. The Ethiopian flag shall consist of green at the top, yellow in the middle and red at the bottom, and shall have a national emblem at the centre. The three colours shall be set horizontally in equal dimension.
2. The national emblem on the flag shall reflect the hope of the Nations, Nationalities, Peoples as well as religious communities of Ethiopia to live together in equality and unity.
3. Members of the Federation may have their respective flags and emblems and shall determine the details thereof through their respective legislatures.

Using OCR

- No Amharic OCR software available
- Copyright issues

Soft copies of Dictionaries

- Complicated
- Made for humans
- Copyright issues



H]k — a. pensée, idée b. motion, proposition c. inquiétude, souci ; H]k ´kqoG il se fait du souci.

H]k „dUi — proposer, suggérer.

H]i hXº — résolu, décidé, déterminé.

H]i d| — droit, sensé.

H]i ¶pZ — a. intransigeant, rigide, opiniâtre b. tenace.

H]i Ë{ñ{p — fermeté d'âme, constance, détermination.

H]lu ¶ºOºPb — association d'idées.

ˆH]k „]UÑ — soulager.

MW H]k — idée directrice.

eZº H]k — résolution.

H]j'õ{p — idéalisme.

Hô]k — a. comptabilité b. arithmétique, calcul c. addition de restaurant, note d'hôtel.

Hô]k ¤º — comptable.

¡Hô]k Mš´k — registre de comptes.

¡Hô]k `ïO — comptable.

H\p ou „\p — mensonge, *subst.* faux.

H\p m|´U — mentir.

iH\p M\ˆU — faire un faux témoignage.

¡H\p — faux, mensonger.

¡H\p ^O — faux, mensonger.

H\m — menteur.

Ethiopic script

- A written language for ~600 years
- No standard for representing the letters until 1997 (Unicode standard in 2000)
- More than 70 different encoding systems (all incompatible with each other)
- Encoding of some fonts can change while the font names stay the same

Dictionary lookup

- Encoding &Transliteration
  - Lack of standards
  - 70 different encoding systems
- Stemming
  - Complex morphology
- Phrases & multiple words
- Proper names
- non-dictionary words

Transliteration (SERA)

□ = he □ = hu □ = hi □ = ha

□ = hE □ = h □ = ho

□ = le □ = lu □ = li □ = la

□ = lE □ = l □ = lo □ = lWa

□ = He □ = Hu □ = Hi □ = Ha

□ = HE □ = H □ = Ho □ = HWa

□ = me □ = mu □ = mi □ = ma

□ = mE □ = m □ = mo □ = mWa ...

Amharic morphology

bEt	house
bEt-u	the house
ye-bEt-oc-E	my houses’
bEt-acew	their house
ke-bEt-u	from the house
ye-bEt-um	the house’s also
ye-bEt-oc-achu	your houses’
le-bEt-oc-acn	for our houses

sebari - one who breaks
sbari - a fragment
sebara - broken
sebere - he broke
asebere - he made somebody to break something
sebabere - he breaks something again and again
tesebere - it has got broken
asabere - he helped in breaking something
asebabere - he helped in breaking something into pieces
seberku - I broke
seberec - she broke
seberu - they broke
sebern - we broke
seberk - you broke
seberachu - you(pl) broke
isebralehu - I will break
sebrealehu - I have been breaking
iyeseberku - I am breaking
siseber - while it was being broken
yemiseber - something that can be broken

Dictionary lookup

- Encoding &Transliteration
  - Lack of standards
  - 70 different encoding systems
- Stemming
  - Complex morphology
- Phrases & multiple words
- Proper names
- non-dictionary words

□□□□□ □□□ □□□□ □□□□□ □□□□ □□□□□□□ □□□□□ □□□□□ □□□□□ □□□□ □□□ □□□ □□□□ □□□□□□□ □□□□□□ □□□□□ □□□□□□ □□□□ □□□□□□□ □□□□□
□□□□□ □□□□□□□□ □□□□□ □□□□ □□□□ □□ □□□□□ □□□□ □□□ □□□□□ □□□□□□ □□□□ □□ □□□□□ □□□□□ □□□□□ □□□□□□□□ □□□□ □□□□□ □□□□□□ □□□□
□□□□□□ □□□□ □□□□□□ □□□□□□ □□□□□ □□□□□ □□□□□□□□ □□□□□ □□□□□ □□□□ □□□□□□□□ □□□□□□ □□□□□ □□□□□□□□ □□□□□□ □□ □□□□□ □□□□ □□□□□□ □□ □□□ □□□□□ □□ □□□□ □□□□ □□□□□ □□□□□□□ □□□□□□□□ □□□□ □□ □□□□□□ □□□□□ □□□□ □□□□ □11 □□□□□□□ □□□□□ □□□
□□□□□ □□□□ □□□□ □□□□□ □□□ □□□□□□□ □□□□ □□□□□

beasmera ketema yemigeNu yeawropa `hebret ambasaderoc yeisayas afewerqi meng`st bemeriw parti wsT yeneberu bale`slTanat yejemerutn yeteHedso `InqsqasE mafenun Indeteqawemu tegele`Se:: zegebawoc Indameleketut yeawropa `hebret begudayu lay weqtawina yetemWala aqWam Indiyz diplomatocu tnant wede brasles mastawexa lkewal:: yeambasaderocu teqawmo leErtra yemiseTewn yelmat projektoc `Irdata liyazegeyew Indemicl riportoc Tequmew; yedEnmarkna yeamErika tewekayoc agerocacew yemiseTut `Irdata Indemayleqeq leasmeraw meng`st mastaweqacewn Teqsewal:: yh beIndih Indale yeferensay wC guday mini`stEr qal aqebay megleCa yeErtra bale`slTanat kedEmokrasiyawi teHedso gar IndaygWazu yetewesede Irmja yalewn ye11 bale`slTanet metaserna yegl gazETocn Htmet metaged meqawemun walta informExn ma`Ikel zegbWal::

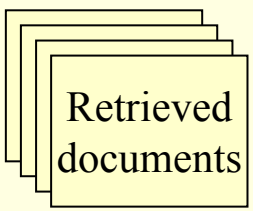
A simple approach to CLIR

‰ìŠN µ‰Dz!K SlsRbÃ y¯œ õRnT m¶ ...

radovan karadzik sleserbeya yego`sa Tornet meri ...

radovan karadzik sle-serbeya ye-go`sa Tornet meri ...

Radovan Karadzic Serbe armée chef conflit crime …



The way forward...

- Standards
  - Encoding
  - Transliteration
  - Representation
  - Tag set
- Shared resources
  - Annotated corpora, tree-banks
  - Morphological analysers, POS-taggers, parsers, ...
- Communication, collaboration, coordination...

Acknowledgements

- Daniel Yacob
- Philip Resnik
- French Ministry of Foreign Affairs
- Jean-Baptiste Chauvain
- Gerard Prunier
- Former and current staff and students at the Departments of Information Science and Linguistics at Addis Ababa University

Thank You!

References

- A. Alemu Argaw, and L. Asker "Web Mining for an Amharic - English Bilingual Corpus", in Proceedings of the 1st International Conference on Web Information Systems and Technologies (WEBIST 2005), 2005.
- A. Alemu Argaw, L. Asker, R. Cöster and J. Karlgren. Dictionary-based Amharic - English Information Retrieval, in Proceedings of Cross Language Evaluation Forum (CLEF 2004), 2004.
- A. Alemu Argaw, L. Asker, and G. Eriksson. Building an Amharic Lexicon from Parallel Texts, in Proceedings of First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a Workshop at LREC2004, 2004.
- A. Alemu Argaw, L. Asker, and G. Eriksson. An Empirical Approach to building an Amharic treebank, in Proceedings of TLT 2003 - The Second Workshop on Treebanks and Linguistic Theories, Växjö, Sweden. November, 2003.
- Atelach Alemu, and Lars Asker “Natural Language Processing with Few Computational Linguistic Resources: An Experiment with Automatic Sentence Parsing for Amharic Texts” Proceedings of SCI 2003.
- Atelach Alemu, Lars Asker, and Mesfin Getachew. “Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward”. In Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, Batz-sur-Mer, France, June, 2003.

Amharic electronic corpora

- Ethiopian News Headlines (ENH)
  - Unicode character set
- Ethiopian News Agency (ENA)
  - non Unicode
- Walta Information Center (WIC)
  - Transition to Unicode from 2003
- Web pages, books, the Bible...

Word sense disambiguation

- Mutual information
- Parallel corpora
- Target language corpora

Copyright Issues

- In the past, no copyright or intellectual property laws
- Recently (2004) passed a strict copyright proclamation that covers a wide range of media and gives the author copyright for life plus 50 years
- The new laws are not yet well understood by the public nor the judicial system
- Possibly, a "fair use" policy whereby electronic articles may be reused, even reprinted, so long as the source is acknowledged and that they are used in a non-commercial context

mehon-u-n
le-walta
bEt-u
ye-kll-u
bEt-oc
halafi-w
be-kll-u
mengst-awi
wereda-woc
be-mehon-u-m
bhEr-awi
ye-tmhrt
be-mehon-u
guba-E

sra-woc
drjt-u
bale-fut
b-alefe-w
le-madreg
mehon-acew-n
ager-oc
maheber-awi
be-ahun-u
be-tekahEde-w
temari-woc
cgr-oc
askiyaj-u
ye-hzb

le-mekelakel
be-debub
newari-woc
k-alefe-w
ministr-u
ye-amErika
ye-drjt-u
ader-oc
le-and
guba-E-w
yemibelT-u
guday-oc
ketem-oc
be-tgray