Creating a parallel Corpus from the „Book of 2000 Tongues“

Presentation of the paper of Resnick, Olsen, Diab
Outline

- Why the bible?
- Annotation
- Research Uses
- Conclusion
Why the bible?
Why the bible?

- Widely available
- Cohesive document
  - 66 books
  - 30-40 authors
- Different language styles
  - Narrative
  - Poetry
  - Correspondence
Availability

- Hebrew, Aramaic, Greek translation between 1000 DC and 100 AD
- Complete translation: 4th century (Latin „vulgate“)
- 1804: 67 languages at least one book
- 2000: 2100 languages
- Today 2575 (Dr. Lupas)
Electronic Availability

• Check if available in a given language:
  – Bible Society
  – http://worldbibles.org/
  – www.worldscriptures.org

• Bible
  – Bibleworks
  – Techflow biblical software
  – Bibles.org
  – Biblegateway.com
Availability

• No parallel corpus available
  – No existence of a standard set of conventions
  – Not explicitly aligned across languages
Careful translation

• Word of god
• Numerous papers on bible translation
• Lots of translation from missionaries
• Conservative translations
• But different styles:
Style differences

• Original Greek phrase in Colossians 1:20
  „The blood of his cross“

• Good News Bible
  „God made peace through his Son's sacrificial death on the cross“
Structure and alignment

- Most parallel corpora not explicitly aligned
- Bible fully standardized
  - Book
  - Chapter
  - Verse
- But: Variations within verses
  - Consistent verse-level alignment provide enough training material for algorithms that learn to do lower level alignments.
Annotation
Intermediate Format

• 3-level hierarchy
  – Book
  – Chapter
  – Verse
Mathew 1:7

ENGLISH: And Solomon begat Roboam; and Roboam begat Abia; and Abia begat Asa;

FRENCH: Salomon engendra Roboam; Roboam engendra Abia; Abia engendra Asa;

DANISH: og Salomon avlede Roboam; og Roboam avlede Abia; og Abia avlede Asa;

FINNISH: Salomolle syntyi Rehabeam, Rehabeamille syntyi Abia, Abialle syntyi Aasa;

GREEK: solomwn de egennhsen ton rohoam rohoam de egennhsen ton abia abia de egennhsen ton asa;

LATIN: Salomon autem genuit Roboam Roboam autem genuit Abiam Abia autem genuit Asa;

SWEDISH: Salomo födde Roboam, Roboam födde Abia. Abia födde Asaf;

SPANISH: Salomón engendró a Roboam; Roboam engendró a Abia; Abías engendró a Asa;

VIETNAMESE: Salomôn sinh Roboam, Roboam sinh Abya, Abya sinh Asa,
Intermediate Format

<b id="MAT">
  <c id="MAT:1">
    <v id="MAT:1:7">
      someverse
    </v>
  </c>
</b>
Target Format

• Corpus Encoding Standard
  http://www.cs.vassar.edu/CES/
  – Guidelines for encoding
  – Different layers of conformance
    (aiming at level 2: correct paragraph level markups +
     consistent marking of some sub-paragraph-el.)

• cesAlign
  – Recording of alignment external
  – Encode 1:1 links:
    – <link xtargets=“GEN:1:1; GEN:1:1”>
Target Format

• Disadvantages of CES
  – Word sense and semantic issues are outside the pure view (CES Sec. 0.2.4)
  – Verse structure does not respect subdivision (cesDec) into paragraph, sentence and token

Example:

<v id="GEN:10:13"> And Mizraim begat Ludim, and Anamim, and Lehabin, and Napthuhim, </v>
Inputs Formats

• Formats of available versions of the Bible are fairly consistent

• Problems:
  – Line breaks as verse delimiter or not?
  – Labels in text or as file names?
  – Header information available?
  – What formatting code?


\c 2
\s Wageni kutoka mashariki
\p
\v 1 Yesu alizaliwa mjini
alipokuwa mfalme. Punde tu
kutoka mashariki walifika
\v 2 wakauliza, <<Yuko wap
Tumeiona nyota yake ilipot

Matthieu 2
1. J'esus \'etant n\'e roi H\'erode, voici des m J\'erusalem,
2 et dirent: O\'u est le : nous avons vu son \'etoil l'adorer.
Annotation Progress

• Minimal effort to handle different input formats
• Authors claim that...

... every input format could be handled via simple pearl scripts within a few days. In simple cases the implementation and execution could be done in about an hour.
Status

- Following bibles (66 books) where annotated: English, French, Danish, Chinese, Finnish, Greek, Latin, Swedish, Spanish and Vietnamese.

(Book/Chapter/Verse-Format)

- Translated into CES format
Research Uses
Translation and Comparative Linguistics

- Helpful for languages with no living speakers
  - Study language for translation
  - Variety of translated forms
Resource acquisition for natural language processing

- Cross language information retrieval
- Machine translations
- Computational lexicography
- Multi-way parallel corpus
  - Variety of language families
  - Carefully translated
  - Nearly sentence level aligned
  - Variety of linguistic phenomena and domains
Conclusion
Conclusion

- Rough sketch
- Out of date
- Helpful:
  - Intermediate format
  - Target format discussion
- Project stopped
- No script resources available :(