Tools for Text Technology

Loïc Boizou

Lithuanian University of Education Sciences
&
Vytautas Magnus University
Centre of Computational Linguistics
Let’s start

Thank you all.

Purpose

– to provide some basics about tools to process textual information;
– to introduce some basics about structuration formats used by these tools;
– to introduce some key concepts related to textual resources.
First Part

Basics
Data acquisition

- Text files:
  - Clean them…
  - Simple text (txt.)
  - utf8 encoding
- OCR (optical character recognition):
  - image ↔ (written) text
  - drawings ↔ letters
- Typing…
- Recording (see Speech Technology)
What is a corpus?

• A collection of texts selected according to certain principles (compare with archives).

• Main question: what for?
  • in applied linguistics, it is always a very important question (compare with traditional linguistics)

• The question of representativeness (and balance)
  • important to keep in mind...
  • but not always easy to achieve.

• What about general corpora?
Metadata

• Proper corpora should have metadata, that is, data about each text included in the corpus.
  • Title,
  • Date,
  • Genre,
  • and others (ex. for spoken corpora information about speakers, see Speech Technology)…
• Important to keep it in line with the issue of balance.
Corpora

- Rough corpora
  - Text + Metadata

- Parallel corpora (alignment)
  - Text + Metadata + Text in another language/media

Annotated corpora

- Text + Metadata + information about the text (ex. named entities, parts of speech, syntactic function)

- Some parallel corpora are annotated.
Annotations
Annotation process (1)

- Fully manual (by human annotators).
- Fully automatic (by automatic annotation tools).
- Mixed:
  - usually, automatic, then human correction.
Annotation process (2)

• Tools that rely on machine learning are usually improved by an iterative process of automatic annotation + human correction
  • human annotated resources $\rightarrow$ machine learning $\rightarrow$ automatic annotation $\rightarrow$ human correction $\rightarrow$ machine learning…

• Often (but not always) starting from initial human prepared data.
  – Supervised (vs unsupervised) methods.
Types of annotations items

- Segments
- Segment labels
- Relations (links)
- Relation labels
Example of segments

- Structural blocks (sentence, paragraph, verse, page…)
- words
  - lemma, parts of speech, transcription, gloss…
- syntagmas
  - named entities, mwe (multiword expressions), phrases/constituents…
Examples of relations

- Syntactic dependency.
- Anaphoric link.
- Rhetorical link.
- Parallel text fragments.
- Link to external resources (ex. to manuscript image).
Second Part

Annotation Formats
Annotation structure (1)

• Tabular:
  – easy to use for human and computer, but cannot describe very complex structures.
  – can be imported to Excel or LibreOffice Calc.
• Tree-based structures, ex. XML, Json:
  – heavy-weighted, unconvenient for humans (except for very simple data), but allows for complex structures.
Annotation structure (2)

• On place (inline annotation)
  – annotations are in the document.
  – suitable for humans.

• Pointer-type (stand-off annotation)
  – annotations are in a different file than the document, with a pointing system.
  – allows to have stable textual source (no mixing between document and analysis) and different concurring structural analyses.
  – unsuitable for human annotators.
Some annotation formats

- Tabular
  - csv/tsv (comma/tab separated values)
  - conllu (inline)
  - brat (stand-off)
- XML
  - TEI (stand-off or inline)
  - FoLiA (mainly inline)
  - Paula (stand-off)
Example

Brat: http://brat.nlplab.org/examples.html#annotation-examples

TEI: https://www.kielipankki.fi/support/recommended-xml-annotation-of-text-corpora/

FoLiA: https://proycon.github.io/folia/example.xml
Third Part

Tools for Textual Resources
Let’s try Linux

• Ubuntu or other distributions:
  • https://www.ubuntu.com/download/flavours
• Many high-quality free tools.
• It can be installed near Windows if you have enough memory.
• Some Linux versions are made for old computers (ex. Lubuntu).
About this selection of tools

• There are very many available tools for processing texts.
• Here is a small set of tools (all tested, but one).
• Presented tools are:
  • free,
  • actively supported/developped.
Editors

• Very many free tools…
• Text editors:
  • Should be compatible with UTF8 and all newlines symbols.
  • For Windows, NotePad++, SublimeEditor…
    • Basic NotePad is bad, do not use it…
  • For Linux, standard Kate or Gedit are fine.
• In case you need it, XML editors (BlueFish…) but the mentioned text editors do well for XML.
OCR (optical character recognition)

• Not much free tools, Tesseract probably the best option.
• Supported by Google.
• Scanning:
  • 300 dpi or better 600 dpi (dots per inches);
  • tiff better image format for Tesseract.
• Best command tested with option -psm 4 or -psm 11:
  – tesseract -psm 4 KhmerSong1.tif KhmerSong_psm4_ocr -l khm
• Some technics to improve recognition:
Tesseract training (1)

- What about old characters? Tesseract has a training system, but it requires patience...

1. Prepare text pages with sufficient variety and redundancy. → if you can…
2. Correct text and picture alignment (semi-automatic).
   - A tool such as jTessBoxEditor can help...
   - http://vietocr.sourceforge.net/training.html
3. Run all machine learning commands.
Tools for human annotators (1)

• Some tools makes annotation easier with graphical interfaces.
• Easy to use, but should be installed on servers by IT specialists.
• Several annotators can work online on the same documents.
Tools for human annotators (2)

• For simple annotation:
  • Brat

• For rich annotation:
  • WebAnno,
  • Flat.
Brat

- [http://brat.nlplab.org/](http://brat.nlplab.org/)
- University of Tokyo
- Annotation categories (segments and relations) must be defined.
- Unconvenient for overlapping annotations or for annotating big spans of text.
Flat

- https://lamas.science.ru.nl/software/flat/
- Maarten van Gompel (Radboud University)
- Access to the Flat server through all main browsers.
- Multi-annotator.
- Multi-layered annotations.
- FoLiA as main format.
- Support seems to weaken.
WebAnno

- https://webanno.github.io/webanno/
- Darmstadt University of Technology
- Access to the WebAnno server through Chrome or Safari (other browsers not fully supported).
- Multi-annotator.
- Multi-layered annotations.
- TEI as main format.
Automatic annotation tools

• Depends on the annotation you need, so they usually need to be programmed for you (no ready to use solution).

• Machine learning library (Scikit in Python) and tools (TensorFlow from Google) allow programmers to learn from data made by linguists or other researchers in Humanities.

• For segmentation, tagging (not really an issue for Khmer), syntactic parsing, UDPipe might be a useful free tool.
UDPipe

• Charles University (Czechia)
• runs as command line.
• Segment, tag and parse data (together or separately).
• learn from previously annotated data (of course scarce data gives bad results).
• Q. Is it suitable for Khmer word segmentation?
Interfaces to data: visualisation only

• For conllu files:
  • several tools, ex. conllu_viewer (online).

• For digital editions:
  • EVT, versioning Machine.
  • Can run on server (IT specialist needed) or on your computer as ready to use softwares.
  • Rely on complex XML structures (TEI or similar).
  • Allow for visualisation of different variations of a text/manuscript.
conllu_viewer

- [link](http://universaldependencies.org/conllu_viewer.html)
- Charles University (Czechia)
- Select and view.

- RMQ better options for conllu files should arrive (some other tools are still not stable).
Versioning Machine

- [http://v-machine.org/](http://v-machine.org/)
- Susan Schreibman (Maynooth University)
- Several projects use it (see webpage).
- Easy to use on your computer.
- Provided with examples (useful as a starting point).
EVT2

- [http://evt.labcd.unipi.it/](http://evt.labcd.unipi.it/)
- University of Pisa
- Easy to use on your computer.
- Provided with examples (useful starting point).
- Younger than Versioning Machine (still in test phase), but easier to modify/customise and better designed.
Interfaces to data: searching data

• Indexation for searching:
  • organize data in an efficient way for searching, like an index in a book.

• Tools
  • AntConc (easy to use for small to medium corpora with very simple annotation).
  • BlackLab (for big corpora with simple annotations).
  • Annis (for small corpora, about 1 milion words, with rich annotation)
  • Textual databases, ex. Elasticsearch.
AntConc

- http://www.laurenceanthony.net/software/antconc/
- Laurence Anthony
- Unzip and use it.
- Good for starting to use corpus (small personal projects or for teaching students the basics of corpus analysis)
- Simple annotation at the word level (ex. parts of speech)
- Q. Is it fine Khmer word writing?
BlackLab

- [http://inl.github.io/BlackLab/](http://inl.github.io/BlackLab/)
- Institute of Dutch Language
- Command line and web application (can be integrated in your own tools).
- Limited annotation (segments only).
- For big corpora:
  - fast,
  - powerful search (with CQL, corpus query language).
Annis

- [http://corpus-tools.org/home/](http://corpus-tools.org/home/)
- Humboldt University in Berlin, Georgetown University
- For personal use, unzip and use (for server you need IT specialist).
- For small to medium corpora (but improving).
- For rich annotation (multi-layered, segments and relations), powerful search (CQL).
- Rich set of tools comes together with Annis (ex. converter Pepper).
Elasticsearch

- https://www.elastic.co/products/elasticsearch
- Elastic NV
- Powerful document search in very big amount of data (possibly billions of words).
- Full text (not annotated).
- Easy to run and to use if you know json format (both data, commands and results are in json).
- You need to prepare a tokenizer for Khmer.
What is left behind

For example:

- Alignment tools
- Lexical tools
Resources

• About corpora
  – Lancaster university:
    • Unit 1 Corpus linguistics: the basics
    • Unit 2 Representativeness, balance and sampling
    • …
  – Developing Linguistic Corpora: a Guide to Good Practice

• Some online corpora
  • British National Corpus
  • Prague Dependency Treebank
  • Universal Dependency project
  • Annis-based projects
Thank you for your attention!
Time for questions, remarks, new ideas…

contact:

lboizou@gmail.com