Provenance in relation to language corpora

Nelleke Oostdijk
3 September 2018

CLS | Centre for Language Studies
Radboud University

Introduction

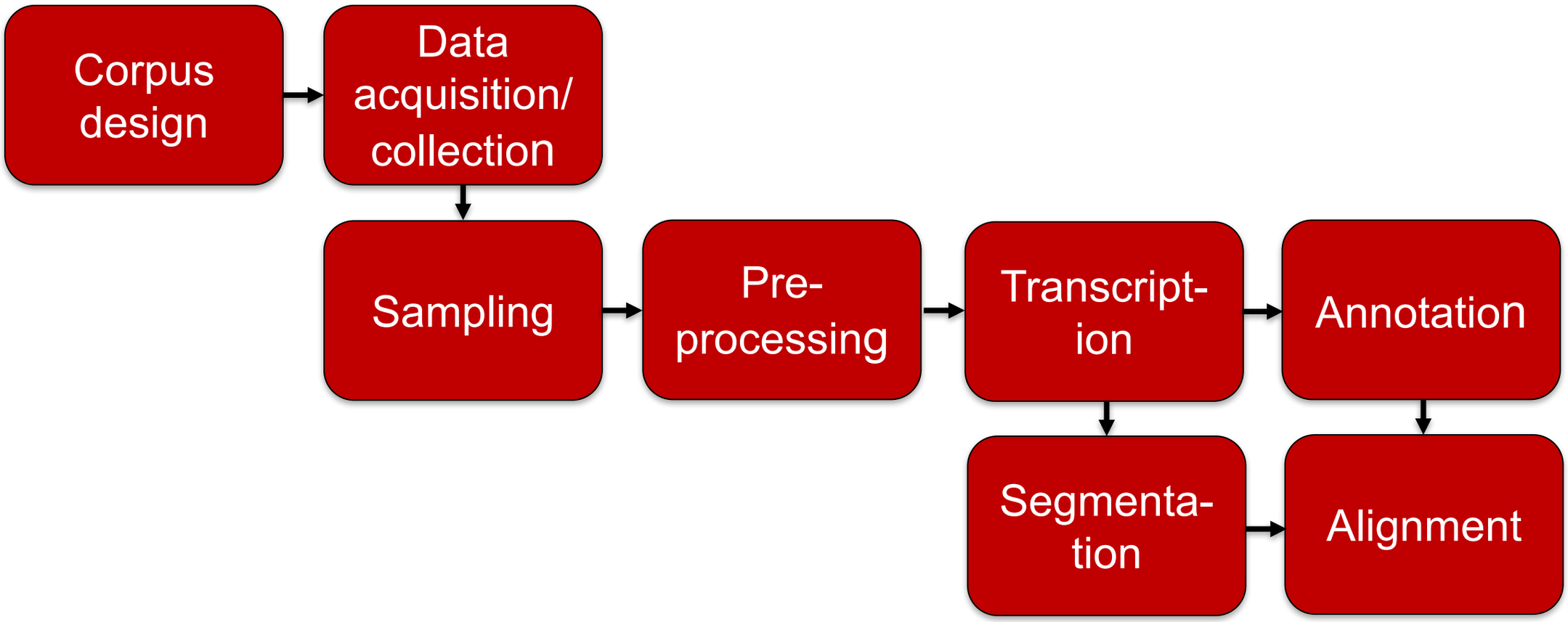
Language corpora

- motivated/balanced data collections for (re-)use by wider research community
- generally designed to serve multiple purposes
- may include data from multi-media and different modalities
- often come with rich annotations

Data provenance

The documentation of where data come from, and the processes and methodology by which it was produced.

Corpus creation



- Definition of design parameters (types of data to be included, types of transcriptions/annotations)
- Developing methods and procedures for various stages – from corpus design to transcription/annotation, segmentation and alignment)
- Documenting the corpus and the corpus creation process

Documenting the corpus and the creation process

Different forms of documentation

- Protocols, guidelines, manuals
- Metadata
- Papers, articles

Observations

- Documentation available with a corpus tends to be mostly user-oriented
- Part of the documentation & information that was available during the time the corpus was in the process of being created may be lost (incl. results from intermediate steps)
- Important information/details needed for re-producing/curating/extending a corpus is/are generally missing

Documentation in support of the creation process

Protocols, guidelines, manuals; for example

- Specification of the corpus design, formats, symbol and label sets
- Protocols for (pre)processing
- Guidelines for transcription and annotation
- Description of dependencies between different stages in the creation process and tasks performed between these stages

Records of data origin, characteristics and workflow progress

Documentation in support of the creation process

Protocols, guidelines, manuals
Records of data origin, characteristics and workflow progress

Despite extensive documentation

- Humans may interpret guidelines differently
- Humans may interpret the data differently
- Tools will not necessarily produce the output they should according to the documentation
- Human verification may be biased by the output presented to them to begin with
- Description of dependencies between different stages in the creation process and tasks performed between these stages found to be lacking relevant information

User-oriented documentation

Aim

Provide what information is deemed useful for those using the corpus, i.e. information enabling users to associate the data with contextual information needed for understanding and use.

Cf. ANDS Provenance

> “For data users, the scienticific basis of their analyses and accountability of their research rely largely on the credibility and trustworthiness of their input data and so they may want to check data quality along with expected level of imprecision.”

In so far as researchers want to be able to check the data quality, they can do so against the documentation or consistency between the original text/audio/video and the transcription and/or annotation layers

User-oriented documentation

Meta-data

Descriptive in nature, providing information on the origin of the data and its characteristics

For example, metadata with the Spoken Dutch Corpus:

Comprises information about speakers and samples - with link to documentation on corpus design and selection methods

- speakers:
  speaker ID, age, place of birth, place of residence, level of education, etc.
- samples:
  sample ID, how, where and when produced/collected, equipment used, types of transcriptions/annotations available, parties responsible for recording, selecting, transcribing, annotating, segmenting, etc., sample size, duration, ...

Re-producing a corpus? (curating or extending it)

So far has not been the focus of corpus compilers:

- Focus has been on collecting data and making data available, not on re-producing them
- Corpus is reference: replicability of research, verifiability of research results

Documentation needed for re-producing corpus is

- generally insufficient
- distributed (over metadata, manuals, papers, ..)
- not always directly linked

Moreover, reproduction in many cases is impossible

- Fugitive data cannot be captured again
- Recording and other conditions cannot be reproduced/recreated/replicated
- Resources, scripts and tools used in the process of creating the corpus may no longer be available (in the exact same version)

Summary

- Corpus creation involves multiple workflows, distributed processes, automatic and manual involvement
- Provenance has the attention of corpus creators, witness the information available with present-day corpora
  - Information provided is
    - mostly user-oriented
    - distributed over metadata, manuals, papers, ...
    - not always explicitly referred to or linked

    Information a user seeks might not/no longer be available
  - Information available with different corpora varies considerably