VALIDATION OF SCIENTIFIC RESULTS AND FAIR PRINCIPLES IN LINGUISTIC RESEARCH WITHIN THE FRENCH CORLI CONSORTIUM

Christophe Parisse and Carole Etienne

CORPUS RESEARCH IN LINGUISTICS

Working in linguistics is working on language

- Research material can come from introspection
- **Research material can come from data collection**

- If data is the basis for research, then the nature and the characteristics of data is very important
  - Checking the quality of the data (controlling the results)
  - Recreating the data (reproducing the results)
  - Reusing the data (adding to the results)

CORPUS: A COLLECTION OF LINGUISTIC DATA

- A corpus is a collection of data recorded at a certain time and with certain conditions.
- The collection itself can be used to testify about the existence of some linguistic facts, at some time and in some conditions.
- The collection can be used to test a linguistic theory, to perform an analysis (manual or automatic).

- Further uses of the corpus can target reassessment of the data, or recreation of the data (with the same assessment)
- This can very much improve the quality of linguistic research.

THE EXAMPLE OF LANGUAGE ACQUISITION DATA

- Language acquisition (how children learn / acquire their mother language) is one of the first field where introspection data was impossible to use.
  - So the use of corpus data is mandatory since the 1960s. Some of the first large collections of linguistic data were collected in the 1960s and published in the 1970s (work by Brown, Bloom, Braine).
- A lot of child language acquisition data is available since 1986 in the CHILDES database (MacWhinney and Snow, 1986; MacWhinney, 2000).
  - These data have all the same format (a simple, but fully described, text format: the CHAT format).
  - These data are free of access (but you have to cite the people that did create the data).

REUSE IN CHILD LANGUAGE ACQUISITION DATA

- The work of Brown (1973), Bloom (1970, 1973) are seminal works and the original data is available and has been used and reused many times for research and teaching.
- The data can also be used in contradictory fashion
  - For example:
    - Schütze and Wexler (1996) produced a paper based on three corpuses from CHILDES
    - Pine, Rowland, Lieven, and Theakston (2005) made a response based on the three corpuses and on a new corpus that was added to the CHILDES database.
    - All this work can be controlled because all the data is available! The old data and the new data. Also different results can be obtained using the same data.

WHY DOES THIS WORK?

- Because the format is the same for all the data (although this is not a fancy recent XML-like format).
- Because all the data can be processed using the same tool (CLAN) that is available since thirty years.
- Because the data is free of access.
- Because the data is located in one place and so it is easy to find data (even though it is not perfect).
- Because conditions for the reuse of the data are clear.

THE FAIR PRINCIPLES AND LINGUISTIC DATA

- CHILDES and the language acquisition research done with the database and the tools follow most of the FAIR principles since more than 30 years
  - Automatic processing is limited but nearly all other properties of FAIR data are in place since the beginning
- The principle exemplified by CHILDES can be extended to all linguistic research based on corpora:
  - This makes it possible to significantly control, replicate, and enrich research

CORLI AND FAIR DATA

- The CORLI Consortium: CORpus, Languages and Interaction
  - A network of researchers and laboratories in France (and Belgium) whose goal is to promote corpora (creation of new data, dissemination of old data) and good practices in corpus linguistic research.
- Promotion of the FAIR Data principles
  - Findability
  - Accessibility
  - Interoperability
  - **Reusability**

ACTIONS OF CORLI

- **Findability:** TEIMETA: A tool for creating specific metadata editors – designing sets of metadata description
  - Make it possible to distribute metadata coding and enrich data deposition
- **Accessibility:** Promotion of deposition in public repositories, free of use for research, using standard and interoperable formats
- **Interoperability:** TEICORPO: A tool for conversion to the TEI and back from various frequently used tool formats:
  - CLAN, ELAN, PRAAT, Transcriber
  - Allowing multiple use of the same data though multiple tools

REUSING THE DATA

- An example: "Corpus linguistics conference JLC2017 in Grenoble, France"
  - Gathering data from 9 corpora of spoken language
    - All these corpora were made public in the last 10 years
    - They follow various methods for metadata (not yet uniform and not all data can be processed fully automatically)
    - We used our tools for format conversions (TEICORPO)
  - The result is a 9 million word corpus, all data come with sound or video, all data is free for research use
  - The data can be used with tools for textometric analyses

- Not yet a perfect FAIR project, but getting close!

CONCLUSION

- Data deposition and reuse in linguistic research is not new
  - But it was limited to certain projects or sub-fields
- It is slowly becoming more and more frequent
  - Deposition is easier
  - People know that they should follow standards and they do it
  - Data is becoming more and more available for reuse
- Validation of scientific results is becoming easier and more frequent