How to make data reusable?
Inter-lab workshop on data and standards 29.5.2015
Summary

3 July 2015

The series of workshops "Working with linguistic and spatial data" is intended to structure and unify the support activities of CorpusLab, GISLab, and VideoLab. The goal of the first workshop "How to make data reusable" was to initiate the discussion on exploiting and handling linguistic and spatial data within the URPP. We have therefore invited speakers with good experience in building, managing, and, especially, sharing linguistically annotated corpora (Tomaž Erjavec), spoken language corpora (Thomas Schmidt), and various, widely used geographic data sets (Beat Tschanz). We have aslo solicited our colleagues Robert Schikowski and Steven Moran to share their fresh experiences with reusing language acquisition corpora.

To have an overview of the current issues with data, we had two focused discussion. The first discussion dealt with particular issues in spoken language. In the second session, the results of a survey conducted before the workshop were discussed.

Introduction to the workshop, Tanja Samardžić

The introduction pointed out the need for a more systematic approach to linguistic and spatial data in the context of the URPP, which is particularly oriented towards data and empirical research. An important component of the URPP are three labs, which are in charge of handling and analysing linguistic and spatial data. This workshop is described as an effort in helping researches make good choices regarding storage, formats, and annotation. Making data reusable is underlined as the key to good choices.

Platforms and standards for data sharing, Tomaž Erjavec

The focus of Tomaž Erjavec's presentation was on making linguistic data reusable through data sharing platforms.

The presentation started with distinguishing between open and closed (locked) data. This was followed by an overview of the main factors that make researches lock their data, including unclear copyright terms, perfectionism, feeling of loosing something, keeping the competitive advantage, additional work that needs to be invested in formatting and documentation.

The following issues are discussed in deatail in the talk:

- How CLARIN language resource repositories can help in overcoming the listed obstacles to data sharing, using the newly establish Slovenian CLARIN center as an example.
- Good practice at different levels of text encoding:
  - Character set (how are characters encoded): Unicode
  - Format (what distinguishes annotations from the text): XML

  - Schema (which annotations does the document use): TEI
  - Metadata (how is the information about the document encoded): many standards in use, CMDI for language data
  - Linguistic categories (the vocabulary of linguistic features): ISOcat, universal dependencies

- Git as a good platform for hand annotated datasets and documentation

Standards in spoken language corpora, Thomas Schmidt

This presentation was about standards for spoken language corpora and good practices in improving the interoperability of spoken language corpora.

As a starting point, Thomas Schmidt described vividly the problems with data standardisation and corpora construction with which the researchers of the SFB Multilingualism at the University of Hamburg were confronted when this SFB was founded in 1999: A situation which bares close resemblance to the situation within our URPP, at least with respect to the high degree of interdisciplinarity and the number and variability of different corpora in use. Then he showed how the initial problems led to data standardisation and to the development of tools such as Exmaralda and COMA and finally to the foundation of the Hamburg Centre for Language Corpora (HZSK).

Schmidt presented a number of international initiatives in data standardisation, namely

- TEI, the Text Encoding Initiative, with its detailed guidelines on metadata,
- an ISO project on spoken language corpora, in which the Institut für Deutsche Sprache is participating and which tries to define standards for the transcription of oral data (in compliance with TEI); and
- CLARIN, which already had been addressed to in the talk of Tomaz Erjavec. Thomas Schmidt ended his talk with a number of recommendations on best practices in dealing with language corpora:
- Always obtain informed consent
- Collect and check metadata immediately
- Use uncompressed audio formats
- Use ELAN, EXMARaLDA, FOLKER, Praat or Transcriber for transcription / annotation
- Backup and version control for your data
- Keep in touch with a centre for publication / long-term archiving

In the following discussion, two foci of interest developed: Many questions were asked with respect to how to structure spoken corpora and their metadata, and the other focus of interest was the question of how to use infrastructure and know-how from outside the URPP (especially from members of the CLARIN consortium) to store, manage, and work with own corpora, both existing ones and corpora which will be created in the near future.

What makes a corpus easy to (re)use?, Robert Schikowski and Steven Moran

The purpose of data processing in the ACQDIV project, at the Department of Comparative Linguistics, is to obtain a collection of ten language acquisition corpora that can be searched and analysed in a unified way, including cross linguistic searches. The corpora used in the project are previously compiled by a number of collaborators. Their reuse in the project turns out to be far from straightforward. This presentation summarised the obstacles encountered and gives suggestions on how to avoid such problems in the future.

The following four items are shown to ensures better data reuse, with concrete examples from recent data processing tasks.

- Consistency: marking the same content with the same label
- Separation of independent contents
- Documentation: describing in detail what is done and why
- Explicit coding: using mark-up languages rather than arbitrary marks and punctuation (e.g. bracketing, question marks, multiple points, etc.)

These recommendations might seem intuitive and straightforward, but the attempt to reuse the existing corpora in the ACQDIV project shows that they are often not followed.

Issues in spoken language corpora, Wolfgang Kesselheim

As an introduction to the discussion, Wolfgang Kesselheim stressed the fact that the question of data reusability should be asked right from the beginning of a research project. Data can only be reused if they meet certain quality standards. With respect to the recording of video and audio data, Kesselheim summarised a number of recommendations made by experts and leading institutions in the field (e.g. the recent recommendations made by the German DFG on this respect, or the more technical recommendations on the internet site of the Centre Informatique National de l'Enseignement Supérieur).

Audio:

- Buy recording devices with a good internal microphone and with the possibility to plug in external microphones; use recording devices which produce data in a popular data format such as wav or mp3, not in a proprietary format (like many dictaphones do).
- If there are several recording modes, always use the best one. If possible: uncompressed linear PCM, 16bit/48KHz. If there are only lossy compression formats – normally mp3 – choose the highest bitrate.

Video:

- Buy a camera with the possibility to connect external microphones, if possible with XLR connector.

- Since at the moment there is no widely distributed, open, free lossless compression format for video, record in MPEG2 or MPEG4 (problem: MPEG 4 seems unable to include the linear PCM audio format).
- Choose a high bit rate, which could mean 3.5 Mbit/s to 9 Mbit/s for SD videos, and 9 to 48Mbit/s for videos in HD (the problem here: higher bit rate means more storage space, which with video is still a problem!). Finally Kesselheim draw attention to the importance of ethical and legal questions: In order for the data to be reusable, a written consent of the participants is needed. Participants have to be informed about
- the name of the project, the names of the responsible researchers
- the aim of the project
- the way the recorded data will be used (consent is not only needed for publication, but also explicitly for the inclusion in an archive!) and
- their right to revoke their consent

In the following discussion many practical problems were addressed and discussed with the attendees. The discussion benefited a lot from the experience of Thomas Schmidt in the construction and management of spoken language corpora.

Survey results, Tanja Samardžić

1 Resources mentioned

The following resources were cited as created or used by researches at the UZH (mostly affiliated with the URPP): Swiss SMS corpus, WhatsApp Switzerland, Variantengrammatik des Standarddeutschen, COSMAS, DWDS, Petra Ivanon stories
10'000 vacation postcards, Schaltergesprächen in Zürich (40 or 160 hours)
Transcribed and labelled phonetic data, human judgements
Corpus of Historical American English, own corpus with Sketch Engine and WordSmith, BNC, CHILDES, the Brown Corpus, Youtube transcriptions Transcripts of interviews, transcripts of TV-mediated spoken language (TV-series)
International Component of English (ICE), Corpus of Global Web-Based English (GloWbE)
Paralleles Corpus von Altrussischen und Russischen
Metadata in law sources
Corporus of Chintang, Swiss German ArchiMob corpus
Text+Berg corpus, globally referenced and tagged images (200M), georeferenced and described images for the UK (3M)

2 Geographic component

Most answers cited the fact that data are associated with a certain region (regional varieties) as a geographic component associated with the resource. Some of the resources contain geographic coordinates, postal codes or names of places. Most often, only the geometric primitive point is used for spatial representation.

3 Storage - where

Most of the resources are stored on department servers. Almost as many are stored on local or personal drives. One corpus is a physical (paper) archive.

4 Storage - how

Different data formats have surfaced through the discussion as one of the biggest issues in reusing data sets. Specialised software often determines a specific format and it is hard to convert data sets to be used with different tools. While XML (Extended Mark-up Language) was listed several times as the format of the data in the survey, a whole variety of other formats was listed, including: mySQL database, scans, audio/video, .wav, EMU speech database (similar to TextGrid in Praat), F5, .csv, .doc .txt, maxgda, raw text, word+TAG, columns (word form + TAG + lemma), Toolbox.

5 Data look-up

Data look-up is usually performed with a specialised software through web interface (Sketch Engine, SMS navigator, ANNIS, XQUERY, CQP) or through a local installation (EMU/R, WordSmith, Maxda). Manual inspection and custom scripts are cited either as a primary (sole) or secondary (after an automatic extraction) procedure.

6 Accessibility

Only two resources of those created at UZH are publicly available for non-commercial uses and accessible online. One is shared upon request. All the others are not accessible to other researches at all.

This status is often described as unwanted. Provided good channels and know-how, resources would be made available for reuse.

7 Limitations

Limitations of working with data identified in the survey include: the complexity of data that need to be stored and searched, tool-specific formats, small data sets, the quality of manual and automatic annotation.

8 Training needs

The skills recognised as needed in order to improve work with data include: Multi-layer browsing software, format conversion, handling multiple files, visualising speakers location on a map, integrating spatial data in transcriptions, using standard corpus tools with spatial data, transcription software, other software for linguistic data management, making data accessible to others, managing a publicly accessible resource.

Some of the issues were addressed by brief instructions directly during the discussion, while others are left for future workshops.

Outcomes

1. The workshop brought many researches together, enabling exchange and levelling of data strategies within the URPP community.

2. An important outcome is an agreement among the participants on the following:

   (a) Linguistic and spatial data constitute an issue characteristic of the URPP.
   (b) Proper data management can help unify the URPP activities and increase its visibility.
   (c) For a successful data management, it is important to store and make data available on a central server.

3. Individual researchers are motivated to take measures in order to share their data: As a direct result of the workshop, we have received one request to help format and share an existing data set and one request for general support in data management.

4. Potential topics for future workshops are identified. These topics include writing scripts for data extraction and conversion, measures of data quality, integrating spatial information in corpora, etc.

5. The three labs propose to accompany the future centralised data repository with URPP data manuals, which would help not only in exploiting the existing data, but also in planning optimal data projects.

Registered participants

Simone Uberwasser
Heiko Hausendorf
Hanna Ruch
Michele Loporcaro
Vincenzo Faraoni
Tania Pacciaroni
Elvira Glaser
Agnes Kolmer
Barbara Sonnenhauser
Nathalie Meyer
Danica Pajovic
Sonja Ulrich
Beni Ruef
Uliana Petrunina
Nicolas Fischer
Fabio Rinaldi

Speakers:
Tomaž Erjavec
Thomas Schmidt
Beat Tschanz
Steven Moran
Robert Schikowski

Organisers:
Tanja Samardžić (CorpusLab)
Curdin Derungs (GISlab)
Wolfgang Kesselheim (VideoLab)