Data curation for a VALID Archive of Dutch Language Impairment Data

Henk van den Heuvel[1,] Eric Sanders[1], Jetske Klatter[1], Roeland van Hout[1], Paula Fikkert[1], Anne Baker[2], Jan de Jong[2], Frank Wijnen[3], Paul Trilsbeek[4]

[1]CLS / Centre for Language and Speech Technology (CLST)

Radboud University Nijmegen, The Netherlands

[2]Universiteit van Amsterdam, The Netherlands

[3]Universiteit Utrecht, The Netherlands

[4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

E-mail: h.vandenheuvel@let.ru.nl

1. Introduction

Vast amounts of data have been gathered on the language use, communicative interactions and speech of groups of speakers with language impairments, in a wide variety of countries. Thus, there is a growing need to share data and set up data banks, motivated by the need for scientific integrity on the one hand and by the desirability of data (re)use to make cross-linguistic comparisons possible, on the other. Furthermore, the availability of a variety and multitude of collections of language impairment data is a prerequisite for the advancement of research and application development in the field of language and speech technology for these specific user groups.

In the area of comparative first language acquisition the CHILDES data bank is the most well-known and appealing enterprise of providing supportive and even innovating cross-language data[1]. CHILDES contains many data sets on language acquisition of normally developing children, complemented by a few data sets from clinical groups. In the Netherlands, but also in other countries, funding agencies now require that data collected in research projects is made available for general (re)use. However, as yet, there are no clear and generally accepted guidelines on making data available obtained in different research projects and from various clinical groups.

Research projects commonly collect more data than can be fully analyzed and reported. In addition, time and financial constraints often make it prohibitive to make the data accessible to others, even if an investment has been made on the part of the researchers and funding agencies in the data collection. Data collected in (clinical) professional contexts are commonly not made available to other researchers. The consequences are obvious. In various places unique and precious data exist that could be useful for both existing and new research projects. This situation might even prohibit comparative, perhaps ground-breaking research.

Another problem is that it is usually difficult to recover data that were stored in diverging formats, such as written notes, score sheets, audio tapes, film and video tapes or DVD, plain or extended transcriptions, excel and SPSS files. However, archiving such data in a general, user-friendly, accessible way is a realistic and fruitful enterprise. A concerted

[1] http://childes.psy.cmu.edu

action is expected help to settle ethical and juridical questions about data access and to establish more intensive data exchange.

In this contribution we focus on the curation of five databases dealing with language impairments in Dutch as part of the VALID initiative. Curation is defined as the application of preservation methods and technologies to ensure that digital information of enduring value remains accessible and usable. The curation was carried out in the framework of CLARIN-NL (Odijk, 2010), which provides an infrastructure for archiving, retrieving and sharing linguistic data (among others). In section 2 we will briefly describe the VALID initiative, and in section 3 we explain the core activity of a pilot project: curation of the five initial datasets.

2. The VALID initiative

Dutch research groups in the field of language pathology agreed to join forces to set up a data consortium for making existing data sources accessible and discussing how to handle future data collection activities more efficiently[2]. The aim is to design and set up a multimedia data archive in which old, current and future data can be brought together: the VALID Data Archive (*Vulnerability in Acquisition: Language Impairments in Dutch*). Open access to such data will allow innovative and overarching research questions to be explored in the area of language impairment. The partners are Radboud University Nijmegen, University of Amsterdam, University of Utrecht, and the Max Planck Institute for Psycholinguistics Nijmegen.

There are several compelling reasons for exchanging and sharing data on language pathology. Obviously, in a small country like the Netherlands, less data can be collected than in larger countries on particular language disorders, a highly specific research domain. The combination of a wide range of language impairments in one data archive not only enhances the study of similar impairments but also advances comparisons between different disorders. Moreover, the inclusion of different age groups allows for quasi-longitudinal research designs. Finally, analysis of task properties and effects that are specific to pathological language groups can make a significant contribution to the evidence base of clinical work.

[2] See also http://validdata.org/

In short, sharing data sources helps develop scientific standards of comparative research. In preparing new research projects the data archive can be used to check initial hypotheses. Finally, properties and effects of tasks specific to pathological language groups can be studied by observing how a variable ‘behaves’ under different task designs. For clinicians, possibilities arise for evidence-based practices, supporting a major trend in professional contexts.

The first step in setting up a VALID data archive was a pilot project in which five existing data sets were curated funded by CLARIN-NL[3]. The pilot enabled us to gain experience in conserving different kinds of clinical language data in an enduring format. This process required a thorough analysis of language data properties and structures and intensive discussions on relevant VALID metadata.

Below all data sets are described briefly, principally with the intention of listing the various materials included in the curation.

(1) Spatial relations in the spoken language of children with Specific Language Impairment (SLI)

[3] http://www.clarin.nl: The pilot project started April, 1 2013 and was concluded July, 1 2014

The data were collected by Radboud University Nijmegen and Kentalis. The informants were 63 SLI children and 24 children from a control group: 56 boys and 31 girls, aged between 5 and 12 years old. The data offered for conservation included:

- Photo & Film task, Taaltoets Alle Kinderen narrative task, Frog Story narrative task (audio recordings, Praat[4] transcripts, SPSS files);
- Raven intelligence test, parts of the WISC intelligence test, Peabody receptive vocabulary test (SPSS data files)

(2) Early language development in children with a family risk of dyslexia (FR) compared to SLI children

The data from Utrecht University consisted of developmental language profiles of two longitudinal cohorts: 110 baby's (19 - 37 months) and 140 toddlers (3;2 – 5;0 years). The baby cohort consisted of 70 FR dyslexia children and 40 controls, and the toddler cohort of 70 FR dyslexia children, 40 controls, and 30 children (tentatively) diagnosed with SLI. Follow-up tests were run at age 8 (65 baby's and 107 toddlers) (De Bree et al., 2012).

The materials available were: preferential listening experiment; measurement of listening times in several trials; categorical perception experiment; tests of productive phonology (elicited naming); various procedures (book reading; card matching); digital recordings of speech, (partly) transcribed in IPA and coded for phonological errors; word–picture matching experiment; eye gaze to corresponding pictures (one out of two per trial was recorded); lexical decision experiment: words (presented in combination with pictures)

[4] http://www.praat.org

correctly or incorrectly pronounced (phonemic errors); many speech elicitation experiments (various designs; digital audio recording; partial transcriptions); auditory grammaticality judgment task; all coded responses in Excel / SPSS formats; WISC digit span task (Kort et al., 2000); Snijders-Oomen nonverbal intelligence test (Snijders, Tellegen, & Laros, 1997); N-CDI's (Zink & Lejaegere, 2002): standardized communicative development inventory, completed by participants' parents.

(3) Bilingual language and communication development in young deaf children

The data were collected by Radboud University Nijmegen and Kentalis from young deaf children in Sign Language of the Netherlands (SLN) and Dutch (Kolen, 2009). The 11 participating children in this longitudinal case study were five boys and six girls, between 3 and 6 years old, who were prelingually deaf and had no mental restrictions. The materials available were:

- Spider Story (SLN and Dutch) (SPSS data files);
- Semi-structured conversations with deaf and hearing adults (SLN and Dutch) (104 video recordings; transcriptions in CHAT-like format of 5 minutes per recording);
- Nijmeegse Observatieschaal voor Kleuters (NOK) (SLN and Dutch), Reynell Test voor Taalbegrip (SLN and Dutch), Dutch version Assessing British Sign Language Development (SLN) (SPSS data files)

(4) Language and executive functioning in children with ADHD and children with SLI

The University of Amsterdam compared the language and executive functioning profiles of children with ADHD to children with SLI and normally developing children (Parigger, 2012). 67 children took part: 26 ADHD children, 19 SLI children, and 22 controls, aged between 7 and 8 years. The male:female ratio was 80:20. This project delivered the following data:

- Sentence repetition task;
- Non-Word repetition task;
- Frog story narratives (processed in SPSS on morphological, syntactic and pragmatic measures);
- Children's Communicative Check-list II;
- CANTAB tasks for executive functioning

(5) Deaf adults' writing skills in Dutch

Radboud University Nijmegen collected written data from deaf Dutch adults and compared these to hearing Turkish and Moroccan-Arabic L2-learners of Dutch on morphosyntactic aspects. The subjects were 46 deaf Dutch adults, 38 hearing Turkish adults, 24 hearing Moroccan adults, and 10 Dutch controls, with approximately as many males as females. The results of the standardized C-test of the Instaptoets Anderstalige Volwassenen (IAV) were coded and processed in SPSS. The Writing task The Frog Story was recorded and stored in ScriptLog[5] (see Wengelin, 2012), and the data was coded and processed in Excel and SPSS. Scriptlog is a program for studying writing processes that offers a writing environment with a text editor and a frame in which pictures can be shown to elicit writing activities. it keeps a record of all events on the keyboard, the exact screen position of these

[5] http://www.writingpro.eu/logging_programs.php#Scriptlog

events, and their temporal distribution. Recorded sessions can be played back in real time on the basis of the log file.

Apart from the interesting mix of various contents and diverging topics in the instruments listed above and the substantial range of pathological groups that were subjected to manifold data collection activities, the five data sets cover the many characteristics of the research designs typical of the field.

For all five data sets, written informed consent from the participants or their caretakers had been obtained at the time of the respective investigations. Informants or their caretakers had agreed to share their speech and/or language data and metadata, on the condition of anonymity, which was ensured by the data providers and infrastructure specialists. In the context of the next step, i.e. the preparation and launching of the VALID archive, the consent was verified again with respect to the candidate data sets of the VALID archive (see further section 3 about anonymisation). In VALID only those data were included for which a consent was obtained from the participants or their parents.

3. Curation

Curation of resources was accomplished in five steps (Oostdijk & Van den Heuvel, 2014). These will be discussed as applied to the five initial VALID databases (see also Klatter et al. 2014). Keywords in data curation are interoperability and sustainability of resources.

1. Data collection

All materials of the five data sets had firstly to be assembled and made available to the technology providers at the Nijmegen Centre for Language and Speech Technology (CLST) and the Max Planck Institute for Psycholinguistics Nijmegen. This turned out to be more time-consuming than expected. In the original projects the researchers were assisted by quite a number of other scientists, such as supervisors and statisticians, and data were spread over various computers and servers. When the principal researcher has moved on to a job elsewhere, portions of the data may have moved too. It was often unclear as to which version of a file was the most recent one and the labeling structure sometimes turned out to be inadequate and inconsistent. Eventually all data were found and handed over to the technology providers.

2. Data anonymisation

In VALID we decided to not manipulate the original audio and video files, and to restrict access to researchers sending a motivated request. We did anonymize the transcripts and metadata and made these publicly accessible. In these files we only used pseudonyms for people but not for other entities such as places or events.

The anonymisation of the participants was accomplished either by using a completely unrelated code or an abbreviation of a nickname previously used in publications or a code based on the original name of a maximum of three letters. For each curated database there is an anonymisation file containing the link of the anonymized names to the actual names. This file is available only the database owners.

3. Data conversion

As CLARIN employs a fixed list of formats[6], all audio files were curated into wav (linear PCM) or MP3 files, and video files were kept in the mpg standard format. Scriptlog files (see 2.5) were included as text files.

All transcriptions were as far as possible converted into CHAT[7] format or ELAN[8] format. The transcripts of the bilingual deaf children database (see 3.3) were delivered as doc files (MS Word), which is not a standard in CLARIN. The transcripts are CHAT-like but not genuinely CHAT. Therefore, before anonymisation, they were converted into text files using Linux tools antiword[9] and abiword[10]. In our view, text files are more easily re-usable than PDF files.

SPSS files were converted into CSV text files, as were excel files. We made sure that labels were preserved as meaningful text rather than codes.

[6] See http://trac.clarin.nl/wiki/WikiStart#Formatsandstandards

[7] http://childes.psy.cmu.edu/manuals/chat.pdf

[8] http://tla.mpi.nl/tools/tla-tools/elan

[9] http://www.winfield.demon.nl/

[10] http://www.abisource.com/

4. Metadata

All data sets were allocated appropriate CMDI[11] metadata files (see Broeder, Van Uytvanck, Windhouwer, Gavrilidou, & Trippel, 2012), both at database level and at recording session level (per speaker). Starting from existing metadata profiles for language acquisition data (Sanders et al., 2014) a specific new profile for data resources related to language and speech impairments was established. Care was taken that ISOcat[12] categories were used. This was time-consuming, as existing metadata categories do not organically fit new corpora with their own terminology. It took several meetings between data and technology providers to reach consensus about the categories and the overall metadata profile. Finally, to fill the profile with the a Python script was written to convert excel files to CMDI metadata files. The script was designed to retrieve specific information about the primary media files and additional written resources (such as formats and duration) automatically by accessing the database and inserting the information into the corresponding CMDI files.

5. Persistent identifiers and accessibility

After curation all data were deposited at The Language Archive (TLA) of the Max Planck Institute for Psycholinguistics Nijmegen and all resources and metadata files were issued persistent identifiers. These are long-lasting references to digital objects that are independent of the actual online location of the object. TLA guarantees the availability of

[11] http://www.clarin.eu/content/component-metadata

[12] http://www.isocat.org/

the archived materials for at least 50 years. All metadata are searchable and browsable in the MPI metadata catalogue. The content of the transcriptions in CHAT and plain text format can be searched with the search engine TROVA[13] -(Stehouwer & Auer, 2011).

[13] http://tla.mpi.nl/tools/tla-tools/trova

4. Conclusion

In this contribution we have highlighted the relevance and need for data curation of language resources related to the domain of language impairments. The Dutch VALID initiative is an example of a data curation project in this area, and we have illustrated the activities in VALID by describing the curation of five different datasets.

Our future perspective is twofold. On the one hand we hope to curate further existing datasets and add them to the VALID archive. On the other hand we hope that new databases will be created along the CLARIN guidelines so that they can be included without, or with minimal, intermediary curation effort. This will also enhance the potential of applications in language and speech technology for user groups with language impairments if only because more powerful language models and acoustic models can be created on the basis of these accumulated data sets.

In the course of the pilot project VALID partners have received requests from various university groups and professionals in the field wanting to contribute or link their data resources to the VALID data archive. This makes the VALID initiative worth continuing. The VALID consortium is in the process of finding new funding to acquire new data sets and to develop and disseminate specifically developed (software) tools for data collection and data management in view of participation in the VALID archive.

Acknowledgements

The research for this paper was funded by CLARIN-NL (http://www.clarin.nl) under grant number CLARIN-NL-12-010.

References

Broeder, D, Van Uytvanck, D., Windhouwer, M., Gavrilidou, M. & Trippel, T. (2012) Standardizing a Component Metadata Infrastructure. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2012*, Istanbul, Turkey.

De Bree, E., Snowling, M., Gerrits, E., van Alphen, P., van der Leij, A., & Wijnen, F. (2012). Phonology and literacy? Follow-up results of the Utrecht dyslexia and SLI project.. In: Benasich, A.A., & Fitch, R.H. (Eds.), *Developmental Dyslexia: Cross-Disciplinary Insights on Early Precursors, Expression, and Remediation* (pp. 133 – 150). Baltimore (MD): Paul Brookes Publishing Co.

Klatter, J., Van Hout, R., Van den Heuvel, H., Fikkert, P., Baker, A., De Jong J., Wijnen, F., Sanders, E., & Trilsbeek, P. (2014). Vulnerability in Acquisition, Language Impairments in Dutch: Creating a VALID Data Archive. *Proceedings of the Eighth International Conference on Language Resources and Evaluation LREC 2014*, Reykjavik, 26-31 May 2014.

Kolen, E. (2009). *De tweetalige ontwikkeling van dove kinderen in de Nederlandse Gebarentaal en het Nederlands. Een meervoudige casusstudie [The bilingual* *development of deaf children in Sign Language of the Netherlands and Dutch. A multiple case study]*. Nijmegen: Radboud University (doctoral dissertation).

Kort, W., Compaan, E.L., Bleichrodt, N., et al. (2000). *WISC-III NL. Wechsler Intelligence Scale for Children*. Amsterdam: Harcourt Test Publishers.

Odijk, J. (2010). The CLARIN-NL project. *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC-2010,* Valletta, Malta, 48-53.

Oostdijk, N. & H. van den Heuvel (2014): The evolving infrastructure for language resources and the role for data scientists. *Proceedings LREC 2014, Reykjavik*, 26-31 May 2014.

Parigger, E. (2012). *Language and executive functioning in children with ADHD.* Amsterdam: University of Amsterdam (doctoral dissertation).

Sanders, E., Van de Craats, I. De Lint, V. (2014) The Dutch LESLLA Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik.*

Snijders, J.T., Tellegen, P.J., & Laros, J.A. (1997). *SON-R 5.5-17*. Lisse: Swets Test Services.

Stehouwer, H., & Auer, E. (2011). Unlocking language archives using search. In C. Vertan, M. Slavcheva, P. Osenova, & S. Piperidis (Eds.), *Proceedings of the Workshop* *on Language Technologies for Digital Humanities and Cultural Heritage*, Hissar, Bulgaria, September 16, 2011; Shoumen, Bulgaria: Incoma Ltd, 19-26.

Wengelin, Å. (2012) *Text production in adults with reading and writing difficulties*. PhD thesis, Göteborg University.

Zink, I., & Lejaegere, M. (2002). N-CDI's: *Lijsten voor Communicatieve Ontwikkeling*. Leuven: Acco.