Vulnerability in Acquisition, Language Impairments in Dutch: Creating a VALID Data Archive

Jetske Klatter[1], Roeland van Hout[1], Henk van den Heuvel[1], Paula Fikkert[1], Anne Baker[2], Jan de Jong[2], Frank Wijnen[3], Eric Sanders[1], 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: j.klatter@let.ru.nl

Abstract

The VALID Data Archive is an open multimedia data archive (under construction) with data from speakers suffering from language impairments. We report on a pilot project in the CLARIN-NL framework in which five data resources were curated. For all data sets concerned, written informed consent from the participants or their caretakers has been obtained. All materials were anonymized. The audio files were converted into wav (linear PCM) files and the transcriptions into CHAT or ELAN format. Research data that consisted of test, SPSS and Excel files were documented and converted into CSV files. All data sets obtained appropriate CMDI metadata files. A new CMDI metadata profile for this type of data resources was established and care was taken that ISOcat metadata categories were used to optimize interoperability. After curation all data are deposited at the Max Planck Institute for Psycholinguistics Nijmegen where persistent identifiers are linked to all resources. The content of the transcriptions in CHAT and plain text format can be searched with the TROVA search engine.

Keywords: data sharing; interoperability; data curation; language impairments

1. Introduction

Over the last fifty years, a wide variety of data has been gathered on the language, communication and speech of groups with language problems in different countries, including the Netherlands. The groups in question had difficulties during acquisition or the problems occurred later in life. Data come from children but also adolescents and adults. Oral and written, receptive and productive data have been gathered on the basis of spontaneous language production, language elicitation tasks, (semi-)structured language tests and experiments on speech and language perception and production. Mainly research groups from universities have been responsible for this data collection, but in the course of their work professionals in the fields of education and health, such as teachers, remedial teachers, clinical linguists, and speech therapists, have also accumulated data. In the past, data were registered by hand on paper, but audio and video recordings have become more common. At the present time data are typically stored in a digital format, but the considerable variation in formats does not necessarily allow for data exchange.

Internationally, there is a growing tradition of data sharing for the purposes of cross-linguistic comparison and complementary analyses. In the field of language acquisition the CHILDES database is the best example[1]. This database concentrates on data from first language acquisition of normally developing children although it contains a few data sets from clinical groups. In the Netherlands, funding agencies require that data be shared but as yet there is no tradition of making data from clinical groups available. We want to make a head start in setting up an open multimedia data archive of Dutch language data from such groups. Open access to such data will allow innovative and overarching research questions to be addressed in the area of language impairment.

It is common knowledge that in scientific research projects more data are collected than can be fully exploited. Data collected in the context of large research projects or smaller PhD projects are usually transcribed, coded, and analyzed in the context of the project but they are rarely fully analyzed due to time constraints. Time and financial constraints also often prohibit efforts to make the data accessible to others, even when a considerable investment had been made on the part of the researchers and funding agencies in collecting the data. Data collected in (clinical) professional contexts are not made available to other researchers for the same reasons.

The consequences of this situation are obvious. In various places unique and precious data exist that could be extremely useful for new research projects. The data are difficult to recover since they are stored in different formats (e.g. written notes, filled-in scoring sheets, audio tapes, film/video tapes or DVD, transcriptions, EXCEL and SPSS files). Creating an archive which is user-friendly and accessible would be of great benefit to all. It will also help to address the specific ethical and legal questions about data access with respect to clinical groups.

Dutch research groups have agreed to join forces in a consortium aiming at making existing data resources accessible, and formulating guidelines that will make future data collection activities more efficient. The aim of the VALID project is to design and set up a multimedia data archive in which old, current and future data can be brought together. The VALID project started in April 2013 and will continue until 1 July 2014.

In this paper, we will describe the ideas behind VALID (section 2), the five data sets included in the pilot project

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

(section 3), and the curation and the accessibility of the data sets (section 4).

2. The VALID initiative

There are a number of compelling reasons for exchanging and sharing data on language pathology. Obviously, in a small country like the Netherlands less language data can be collected than in larger countries, especially so with respect to 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 needed for the development of new diagnostic instruments and intervention techniques. Three examples illustrate the scientific merits of a VALID data archive. In each case, the availability of specific data collections can address research questions that could not otherwise be answered or would require collection of new data.

(1) Normative data. The consequences of language impairment can only be assessed by comparison with typical development. This matching procedure is a painstakingly difficult process, often requiring a large pool of potential matches for the inclusion of control (i.e., not-affected) participants. If more data become available from a significant number of controls, from different data sources, matching could be made substantially easier, allowing for a better view on language development processes.

(2) Infrequent clinical conditions. For those researchers who focus on highly specific subgroups of language and speech problems, such as Landau-Kleffner syndrome or aphasia in (Dutch) Sign Language, it is hard to find subjects/informants. Research becomes more feasible when data can be accessed from other data sets.

(3) Comorbidity. The research area of language impairment typically entails verifying comorbid symptoms in a clinical group (like SLI and dyslexia) and comparing them on behavioural variables to groups that have only the 'other' disorder. The availability in one data archive of 'pure' clinical groups and data from individuals or groups who show symptoms of more than one disorder makes such research possible.

In short, sharing data sources means that we can better meet the scientific standards of comparative research. In preparing a research project, the data archive can also be used to check hypotheses that the project departs from. Finally, properties and effects of tasks that are 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.

In the framework of the CLARIN-NL infrastructure (Odijk, 2010) a pilot project for VALID has started in which five data resources are curated. By curation we mean the application of preservation methods and technologies to ensure that digital information of enduring value remains accessible and usable.

The five data sets will serve as the launching platform for a more elaborate VALID data archive, together with the BISLI data set that is currently being curated within the CLARIN-NL framework as well (the CLARIN FESLI project[2]). For all data sets concerned, written informed consent from the participants or their caretakers has been obtained. Informants or their caretakers have agreed to share their speech/language data and metadata, on the condition of anonymity, which will be ensured by the data providers and infrastructure specialists, when the resources are curated.

3. The pilot data-sets of VALID

3.1 The SLI RU-Kentalis data base

The aim of the project was to investigate the expression of spatial relations by children with Specific Language Impairment (SLI) and normally developing children in spoken language production. Data were collected from 63 children with SLI and 24 control children. The total group consisted of 56 boys and 31 girls, aged between 5 and 12 years. In the project various instruments were employed. The tests used were Raven's Progressive Matrices, Wechsler Intelligence Scale for Children (WISC; Block Pattern and Mazes), and the Peabody Picture Vocabulary Test. The results of these tests were analyzed in SPSS and these SPSS data files are available. Especially for this project a Photo/Film Task and a Route Description Task were developed. The Photo/Film Task was audio recorded and the Route Description Task was both audio and video recorded. Both tasks were transcribed in Praat[3], and data were processed and coded in SPSS. Two additional narrative tasks were included as well: one of the narratives of the Taaltoets Alle Kinderen (Language Test for All Children) and one narrative based on the popular Frog series. The narratives produced were audio recorded, transcribed in Praat, and processed and coded in SPSS.

3.2 The UU SLI-Dyslexia project data base

The data were collected to allow for a systematic, longitudinal comparison between children at familial risk of dyslexia, children diagnosed with SLI and age-matched controls, focusing on phonology (speech sounds) and grammar, both in perception and production (van Alphen et al., 2004). Two cohorts of children were included: (a) a 'baby' cohort, consisting of children 19 months of age at the first assessment session to approximately 37 months at the last; (b) a 'toddler cohort', aged 3;2 (years; months) at inclusion to about 5;0 at the last test session. The baby cohort comprised ~70 children at familial risk (FR) of dyslexia as well as ~40 controls. The toddler cohort comprised ~70 FR children, ~40 controls, and ~30 children (tentatively) diagnosed with Specific Language Impairment (SLI). All children were seen at four time points, separated by ~6 month intervals. At each of these sessions, several experimental and observational procedures were conducted. Children from both cohorts returned to the lab at age 8 for follow-up tests on reading and language abilities (toddler cohort: n = 107; baby cohort: n = 65; De Bree et al., 2012). The materials

[2] see http://www.clarin.nl/node/278

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

available include:

- data of a preferential listening experiment addressing children’s recognition of grammatical patterns in Dutch (dependent variable: listening times across trials);
- experiments assessing categorical perception of speech sounds;
- assessments of productive phonology by means of elicited naming, using various procedures (digital recordings of speech; [partly] transcribed in IPA and coded for phonological errors);
- a spoken word–picture matching experiment; eye gaze to corresponding pictures (one out of two per trial) was recorded;
- a lexical decision experiment in which words (presented in combination with pictures) were correctly or incorrectly pronounced (judgments, reaction times);
- several speech elicitation experiments (digital audio recordings; partial transcriptions);
- auditory grammaticality judgment tasks (coded responses in Excel / SPSS formats);
- WISC (Kort, Compaan, Bleichrodt, et al., 2000) digit span task;
- Snijders-Oomen (Snijders, Tellegen, & Laros, 1997) nonverbal intelligence test;
- N-CDI’s (Zink & Lejaegere, 2002): standardized communicative development inventory, completed by participants’ parents.

3.3 The bilingual deaf children RU-Kentalis database

In this longitudinal case-study (Kolen, 2009), 11 deaf children took part: 5 boys and 6 girls in the age range of 3 to 6 years. These children were prelingually deaf (hearing loss of minimally 80dB Fletcher Index on the best ear), and did not have any mental restrictions. The study focused on the bilingual language and communication development of these young deaf children in Sign Language of the Netherlands (SLN) and Dutch. The central part of the data collection was formed by semi-structured conversations of the children with deaf and hearing adults. The conversations were video recorded (SLN and Dutch). Per recording five minutes were selected and transcribed in a CHAT-like format (104 recordings). Furthermore, two narratives were administered: one from the Nijmeegse Observatieschaal voor Kleuters (NOK, Nijmegen Observation Scale for Infants; SLN and Dutch) and the Spider Story (SLN and Dutch). The results of these tasks were processed and coded in SPSS. This also applied to the Reynell Test for language comprehension (SLN and Dutch) and the Dutch version of Assessing British Sign Language Development (SLN).

3.4 The ADHD and SLI corpus UvA database

In this project (Parigger, 2012) data were collected from 26 Dutch children with ADHD, 19 Dutch children with SLI, and 22 typically developing children as controls. All children were aged between 7 and 8 years and had a non-verbal IQ in the normal range. The gender distribution was 80% male and 20% female in all three groups in order to reflect the distribution in clinical groups. The goal of the project was to compare the language and executive functioning profiles of children with ADHD to children with SLI and TD children. This was done on the basis of several language tasks: a Sentence repetition task, a Non-Word repetition task, and the Frog story narrative. The language tasks were recorded on video (mpg), and transcriptions in CHILDES’ CHAT format were provided for the Frog stories. The data from the narratives were analyzed for morphological, syntactic and pragmatic aspects. The Children’s Communicative Check-list II was also used to measure pragmatic behaviour. For executive functioning subtests from the CANTAB battery were used as well as a fluency task. The quantitative data were entered into SPSS and are stored in a SAV file.

.

3.5 The deaf adults RU database

In this project the acquisition of Dutch by deaf Dutch adults (late L1/early L2) was investigated and compared to the proficiency of hearing Turkish and Moroccan-Arabic adult L2-learners of Dutch (late L2) on morpho-syntactic aspects. The informants were 46 deaf Dutch adults, 38 hearing Turkish adults, and 24 hearing Moroccan adults, as well as 10 Dutch controls. The gender distribution was 53 males (22 deaf, 31 Turkish/Moroccan) and 55 females (24 deaf, 31 Turkish/Moroccan). The materials available are the scores on the standardized C-test Instaptoets Anderstalige Volwassenen (IAV), which are coded and processed in SPSS, and the results on the writing task of The Frog Story, that are recorded and stored in ScriptLog (Holmquist; see Wengelin, 2012) and coded and processed in Excel and SPSS.

3.6 IPR of the data

Researchers using the database must declare that they will publish the data in an aggregated or anonymized form, with an acknowledgement to the VALID data archive plus the specific database(s) used.

Consent forms were signed by participants (or their legal caretakers) for all five databases. The forms arrange that metadata and transcriptions are anonymized and are accessible for researchers without further consent. In some cases, informants will be contacted again to check if their initial permission extends to the audio and video recordings upon a motivated request by a researcher.

4. Curation of the datasets

The curation of databases of the type collected in VALID was not a straightforward enterprise. The curation team at CLST encountered issues that are general for the curation of larger databases (cf. Oostdijk, van den Heuvel, 2014), but also issues related to this specific type of language resources.

4.1 Data collection

Obtaining the final data collection turned out to be time-consuming. The researchers who collected and handled the data are no longer available most of the time, and, generally speaking, they do not consider curation their highest priority. As a consequence, considerable time had to be invested in trying to reach them via e-mail, telephone, via supervisors, etc. It obviously was not reluctance to hand over the data, but sheer lack of time that considerably prolonged the process of getting all data available. Particularly when the data resided at various researchers, it appeared difficult to obtain all data sets and to warrant that we succeeded in having the final version of the complete data collection in our hands.

Typically, data sets were submitted in parts by researchers. Therefore multiple submissions were needed to complete the databases. Also, inspection of the data by the curation team evoked further questions as to the nature and completeness of the data leading to additional information and data exchanges.

Particular difficulties arose when we tried to include research analyses and outcomes (such as statistical analyses on the database proper). These analyses and the resulting data abound in varieties of formats and can be expanded arbitrarily. This is why we decided to restrict ourselves to the curation of the basic data (audio, video, written text files) with their transcriptions, and to include only test scores that were obtained as diagnostic metadata for the recorded participants (and were not a result of analyses on the database). We deviated from this principle in exceptional cases only.

4.2 Data anonymization

Anonymity of the persons recorded is of paramount importance (O'Meara & Good, 2010). Anonymity must be secured in the filenames, the metadata, the transcriptions, and in the raw media files (audio, video, text). Anonymization, however, is a rather paradoxical operation: it requires a substantial effort resulting in corruption of the data. Removing names from text is one thing, but removing other information such as date/time/location etc. may render transcripts unintelligible and thus virtually worthless. Furthermore, removing corresponding personal information from audio-files requires introducing beeps in the speech stream. In video files faces must be blurred, which is fatal in recordings involving sign language where face expressions are important (para)linguistic sources of information.

In VALID we decided to keep the audio and video files untouched, and to keep their access restricted to researchers sending a motivated request. We did anonymize the transcripts and metadata and made these publicly accessible. In these files we only anonymized names, but not other named entities, e.g. spatiotemporal references to concrete places or events.

The anonymization of the names was realised either by using a completely unrelated code for the participants or by using an abbreviation of a nickname which was also used in publications involving the database or by using a code based on the participant's name of maximum three letters which still kept the name itself unrecognizable.

For each curated database there is an anonymization file containing the link of the anonymized names to the actual names. This file is in principle not available to others than the database owners.

4.3 Data conversion

CLARIN-NL has a restricted set of permitted formats for various file types[4]. For audio and video it was convenient to stick to wav/mp3 and mpg standard formats. Scriptlog files (see 3.5) were included as text files. SPSS files were converted into CSV text files, as were excel files .

The transcripts of the bilingual deaf children database (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, after anonymization, we converted them into text files by using Linux tools antiword[5] and abiword[6]. In our view, text files are more easily re-usable than PDF files.

Where deemed appropriate the directory structure of the database was modified, e.g. we put the recordings and transcripts of one participant all in the same subdirectory. Names of directories and files were made as uniform and consistent as possible. This is not required for accessing the files via the CMDI metadata files (where links to files are used), but it renders the database better structured independent from CMDI access. CMDI files are described below.

4.4 Metadata

A considerable part of our efforts was devoted to setting up a proper CMDI[7] profile (Broeder et al., 2012) containing a (hierarchical) description of the metadata considered relevant for pathological language data. All metadata categories in the VALID CMDI profile are registered at ISOcat[8]. This profile will be made available via the CMDI registry[9]. The VALID CMDI profile borrowed many building blocks from the LESLLA profile (Sanders et al., 2014). The profile contains information about the research project, individual recording sessions, the content of the sessions (e.g. cycles of various tasks), the subjects and their language impairments and language characteristics, the raw audio/video files, the associated written resources (such as transcripts), and scores of administered tests. The scores of these tests could be added in the metadata per participant or as one CSV file covering the (all) test scores of all subjects.

Newly defined ISOcat categories for VALID are: Language impairments, other impairments, vision status, intelligence, socio-economic status, personal history, characteristics father, characteristics mother, characteristics partner. A full overview of VALID metadata categories is shown in the appendix.

Since it is not feasible to fill the CMDI metadata files (which are XML files) by hand with the database metadata, an excel file was used as intermediary. This

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

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

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

[7] CMDI = component metadata infrastructure

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

[9] http://www.clarin.eu/cmdi

excel file contained rows with the metadata categories and an explanation per category. A written instruction was provided to carry out this task, accompanied with a face-to-face session in which the instruction was orally and visually explained to curation employees at the various sites. Particular attention was given to map the original metadata to the corresponding CMDI category. For each database a file was kept in which these mappings were documented. The files with the original metadata and the file with the mapping information were added as CSV files to the curated database after they were anonymized.

A Python script was written to convert the resulting excel file to CMDI metadata files. The script was designed to retrieve specific information about the mediafiles and additional written resources (such as formats and duration) automatically by accessing the database and inserting the information into the corresponding CMDI files.

4.5 Persistent identifiers and accessibility

Every single file and metadata record that is archived will automatically get a Handle Persistent Identifier assigned to it. This ensures a stable reference that can be used in publications or in other online resources. The Language Archive at MPI has implemented support for part identifiers (suffixes) of the Handle PID system in such a way that different views on the resource can be requested by altering the part identifier. Resolving the handle without a suffix will lead directly to the resource itself, whereas an "@view" suffix will lead to the resource in the context of the archive browser. The latter can be useful for seeing the metadata and other associated resources and is also useful in case that the resource is not directly accessible, such that a user can see what steps are required to get access to the resource. The audio and video files within VALID will not be anonymized and will therefore only be accessible upon request, whereas the transcripts are anonymized and will be made publicly accessible. The content of the transcriptions in CHAT and plain text format can be searched with the TROVA search engine[10] of The Language Archive. Time-aligned transcriptions can also be viewed in sync with the media in a web browser with the ANNEX annotation viewer[11], provided that the user has access to the media.

5. Acknowledgement

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

6. References

Alphen, P. van, de Bree, E., de Jong, J., Gerrits, E., Wilsenach, C. & Wijnen, F. (2004). Early language development in children with a genetic risk of dyslexia. *Dyslexia,* 10(4), 265-288.

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.* Baltimore (MD): Paul Brookes Publishing Co. (pp. 133 – 150).

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.

O'Meara, & Good, J. (2010). Ethical issues in legacy language resources. In Language & Communication, Vol. 30(3), ,pp.162–170.

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

Oostdijk, N. & Van den Heuvel, H. (2014). The evolving infrastructure of language resources and the role of data scientists. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC-2014, Reykjavik..*

Parigger, E.M. (2012) *Language and Executive Functions in Children with ADHD.* Ph.D. thesis, University of Amsterdam

Sanders, E., Van de Craats, I. De Lint, V. (2014) The Dutch LESLLA Corpus In *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, 16 September 2011 (pp. 19-26). Shoumen, Bulgaria: Incoma Ltd.

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.

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

[11] http://tla.mpi.nl/tools/tla-tools/annex/

7. APPENDIX: HIERARCHICAL OVERVIEW OF VALID METADATA CATEGORIES

- Main
  - Documentation
    - Documentation type
    - Filename
  - References
  - Session
    - Name
    - Title
    - Date
    - References
    - Location
      - Continent
      - Country
      - Region
      - Address
    - Project
      - Name
      - Title
      - ID
      - Funder
      - Website
      - Contact
      - Duration
    - Content
      - Cycle
      - task
      - Genre
      - Subgenre
      - modalities
      - subject
      - CommunicationContext
        - Interactivity
        - PlanningType
        - Involvement
        - SocialContext
        - EventStructure
        - Channel
      - Languages
    - Actors
      - Actor
        - Role
        - Name
        - Code

        - FamilySocialRole
        - Ethnic group
        - Age
        - BirthDate
        - Sex
        - Education
        - Grade
        - Anonimyzed
        - Origin participant
        - Origin parents
        - FamilyStructure
        - Family Age
        - Friends structure
        - Residence History
        - Other details
        - Country of Birth
        - Age at Immigration
        - Level of Bilingualism
        - Language Mode
        - Literacy
        - ActorCharacteristics
            - Language impairment
            - Other impairments
            - Hearingstatus
            - Visionstatus
            - Handedness
            - signLanguageExperienceAcquisitionLocation
            - Intelligence
            - SocialEconomicStatus
            - PersonalHistory
            - CharacteristicsFather
            - CharacteristicsMother
            - CharacteristicsPartner
        - Languages
            - MotherTongue
            - PrimaryLanguage
            - HomeLanguage
        - Contact

- Resources
    - MediaFile
        - Size
        - TimePosition
        - Access
            - Contact
    - WrittenResource
        - ResourceLink

- MediaResourceLink
- Size
- Transcription
- Description
- Validation
- Access
  - Contact
- TestScores
  - Access
    - Contact
    - Description
- Anonyms
  - Access
    - Contact