HENK VAN DEN HEUVEL[1], NELLEKE OOSTDIJK[1], CAROLINE ROWLAND[2], PAUL TRILSBEEK[3]

Il *Knowledge Centre for Atypical Communication Expertise* di CLARIN[4]

Abstract

In this chapter we introduce the *CLARIN Knowledge Centre for Atypical Communication Expertise*. The mission of ACE is to support researchers engaged in languages which pose particular challenges for analysis; for this, we use the umbrella term "atypical communication". This includes language use by second-language learners, people with language disorders or those suffering from language disabilities, and languages that pose unique challenges for analysis, such as sign languages and languages spoken in a multilingual context. The chapter presents details about the collaborations and outreach of the centre, the services offered, and a number of showcases for its activities.

1. *Introduzione*

Negli ultimi anni è stata istituita ed è andata consolidandosi CLARIN, un'infrastruttura di ricerca europea che offre risorse e tecnologie linguistiche a studiose e studiosi di scienze umane e sociali (cfr. clarin.eu; Hinrichs & Krauwer 2014; De Jong *et al.* 2018; Krauwer & Maegaard 2022). L'infrastruttura consente agli utenti di accedere a dati e strumenti distribuiti tramite ambiente online ad accesso singolo (De Jong 2019). Oltre a mantenere un'infrastruttura tecnica e i relativi protocolli, CLARIN investe nella cosiddetta *Knowledge Sharing Infrastructure* (KSI)[5], il cui scopo è la condivisione di conoscenza ed esperienza dell'infrastruttura tecnica, del suo funzionamento e del suo utilizzo tra tutte le parti interessate, da chi fornisce le tecnologie e le risorse all'utente finale. A tal fine all'interno della rete CLARIN giocano un ruolo fondamentale i *Knowledge (K-)Centres*, che offrono consulenza sulla raccolta e sulla gestione dei dati, forniscono informazioni sulle risorse e sui servizi

[1] CLS/CLST, Radboud University.

[2] Donders Institute for Brain, Cognition & Behaviour, Nijmegen & The Language Archive, MPI for Psycholinguistics.

[3] The Language Archive, MPI for Psycholinguistics.

[4] La versione originale del contributo è: Van den Heuvel, H. & Oostdijk, N. & Rowland, C. & Trilsbeek, P. 2022. The CLARIN Knowledge Centre for Atypical Communication Expertise. In Fišer, D. & Witt, A. (eds), *CLARIN The Infrastructure for Language Resources*, 373-388. Berlin: De Gruyter. https://doi.org/10.1515/9783110767377. La versione contenuta nel presente volume è stata tradotta dall'inglese da Letizia Cirillo.

[5] https://www.clarin.eu/content/knowledge-infrastructure.

disponibili, su dove trovarli e su come accedervi, nonché supporto all'utilizzo di varie metodologie e applicazioni e, su richiesta, corsi di formazione ciascuno per il suo ambito di intervento.

Esistono attualmente oltre venti *K-centres* certificati[6]. Uno dei più recenti è il *K-centre for Atypical Communication Expertise*[7] (in breve ACE), istituito presso il Centro di Tecnologie Linguistiche (CLST) della Radboud University[8]. La missione del Centro ACE consiste nel fornire assistenza alle ricercatrici e ai ricercatori alle prese con lingue che pongono sfide specifiche in termini di analisi e che rientrano nella macrocategoria denominata "comunicazione atipica". Tale denominazione include produzioni linguistiche di parlanti di L2, di parlanti con disturbi del linguaggio o disabilità linguistiche, ma anche produzioni linguistiche la cui analisi pone questioni specifiche, come nel caso delle lingue segnate o dei contesti multilingui. Fa inoltre riferimento a diversi canali sensoriali e modalità espressive plurime (lingua scritta, parlata, segnata, ecc.) e abbraccia diverse fasi evolutive. I destinatari delle attività del Centro ACE sono linguisti, psicologi, neuroscienziati, informatici, logopedisti e pedagogisti. Una recente pubblicazione che descrive le attività del Centro è quella di van den Heuvel *et al.* (2020a), di cui il presente contributo rappresenta un ampliamento e un approfondimento.

Il paragrafo 2 è dedicato alle collaborazioni del Centro ACE, mentre il paragrafo 3 ne descrive i servizi, e il paragrafo 4 ne esemplifica le risorse. Nel paragrafo 5 è discussa l'importanza della collaborazione nel rendere le risorse accessibili attraverso due centri dati di CLARIN e nel paragrafo 6 vengono delineate le strategie adottate per attività di divulgazione e sensibilizzazione.

2. *Collaborazioni*

Il nucleo del Centro ACE è il CLST[9] della Radboud University. Il Centro ha poi stretti legami con singoli ricercatori e ricercatrici e gruppi di ricerca del *Centre for Language Studies*[10] che si occupano di acquisizione linguistica[11], apprendimento della lingua e terapia linguistica[12] e lingua dei segni[13].

All'interno di CLARIN[14], il CLST ha lo status di *C-centre* e, in quanto tale, fornisce metadati all'infrastruttura e consente l'accesso a strumenti e applicazioni web grazie ai servizi di identità federata messi a disposizione da CLARIN.

[6] https://www.clarin.eu/content/knowledge-centres.
[7] https://ace.ruhosting.nl.
[8] https://www.ru.nl/clst.
[9] https://www.ru.nl/clst/ and https://www.ru.nl/cls/our-research/research-groups/language-speech-technology.
[10] https://www.ru.nl/cls.
[11] https://www.ru.nl/cls/our-research/research-groups/first-language-acquisition.
[12] https://www.ru.nl/cls/our-research/research-groups/language-speech-learning-therapy.
[13] https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics.
[14] https://www.clarin.eu/content/clarin-centres; http://roadmap2018.esfri.eu/projects-and-landmarks/browse-the-catalogue/clarin-eric.

Al fine di archiviare contenuti e corpora relativi alla comunicazione atipica rendendoli accessibili secondo i principi FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*), il CLST ha stabilito una stretta collaborazione con *The Language Archive* (TLA), il quale ha sede presso il Max Planck Institute di psicolinguistica (MPI) di Nijmegen. In qualità di *B-centre*[15] di CLARIN, lo scopo del TLA è quello di osservare come la lingua è utilizzata in situazioni quotidiane. L'attività principale del TLA è pertanto quella di raccogliere materiali audio e video di comunicazioni in lingue parlate e segnate, unitamente alle relative trascrizioni, analisi, annotazioni e ad altro materiale collegato, come foto o appunti. Il TLA si occupa inoltre di archiviare dati sensibili (parlato e trascrizioni) e di coadiuvare l'infrastruttura di metadatazione CMDI[16] (cfr. Windhouwer & Goosen 2022). Il TLA sostiene infine procedure di autenticazione, l'accesso selettivo ai dati e l'identificazione persistente.

Relativamente a corpora di parlato prodotto da persone con disturbi del linguaggio, il Centro ACE lavora in stretto contatto con DELAD[17], la *Database Enterprise for Language And speech Disorders*[18], un'iniziativa volta a condividere tra ricercatrici e ricercatori corpora di lingua parlata da soggetti con disturbi della comunicazione e del linguaggio in archivi sicuri all'interno dell'infrastruttura CLARIN (cfr. Kamocki *et al.* 2022), in conformità a quanto previsto dalla normativa europea (regolamento Generale sulla Protezione dei Dati, GDPR)[19]. DELAD organizza seminari e laboratori su come rendere condivisibile questo tipo di corpora (cfr. Lee *et al.* 2021) e, in particolar modo per i disturbi del linguaggio, incoraggia, attraverso il Centro ACE, forme di collaborazione con la TalkBank e le banche dati cliniche della Carnagie Mellon University[20]. In virtù di tale collaborazione, di cui alcuni esempi sono riportati nel paragrafo 5, è possibile registrare dati presso la TalkBank, con metadati e pagine di atterraggio sul sito corrispondente, lasciando al TLA il compito di archiviarli e autenticare l'accesso ai dati "grezzi" (tipicamente audio e video).

Per garantire accesso a dati sensibili, il Centro ACE è inoltre coinvolto nel Progetto SSHOC[21], che tra i suoi compiti ha quello di creare un inventario di sistemi e tecnologie adatte a condurre attività di ricerca su dati sensibili quali registrazioni audio-video con soggetti affetti da patologie del linguaggio. Si tratta di sistemi e tecnologie che consentono di accedere a dati sensibili depositati in archivi centrali, da cui possono essere scaricati, oppure in archivi protetti fruibili solo da remoto. In quest'ultimo caso è fondamentale che i dati non migrino dal luogo sicuro in cui si trovano. L'utente non potrà scaricarli ma dovrà accedere a una rete sicura dove poter condurre le sue analisi utilizzando strumenti disponibili all'interno della rete stessa; potrà scaricare solo i risultati delle analisi, i quali saranno soggetti a controlli da par-

[15] https://tla.mpi.nl/resources.
[16] https://www.clarin.eu/content/component-metadata.
[17] http://delad.net.
[18] Inoltre *delad* in svedese vuol dire "condiviso".
[19] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.
[20] https://talkbank.org.
[21] https://sshopencloud.eu.

te del gestore della rete o dei dati. In tal modo si evitano perdita e corruzione di dati, questioni di estrema rilevanza per il tipo di dati di norma trattati al Centro ACE. Nel paragrafo 3 esamineremo più da vicino le sfide che il GDPR pone rispetto alla condivisione di dati sensibili.

Nel 2021 ha avuto inizio una nuova collaborazione nell'ambito delle lingue segnate, con il coinvolgimento di altri *K-centres* di CLARIN. A fine 2020, durante un incontro tra i vari centri, era emerso che erano in otto a occuparsi di raccogliere e analizzare dati in lingue segnate. Questi otto centri si sono in seguito riuniti a distanza per scambiarsi informazioni più dettagliate e idee e proposte per progetti finanziati sulla base delle risorse di ciascuno, disponibili sui rispettivi siti grazie a CLARIN. Questo avvio di collaborazione è sfociato nel 2021 in un *Resource Family project for Sign Language*s, finanziato da CLARIN-ERIC[22] e condotto da quattro *K-centres* specializzati in lingue dei segni, con il supporto degli altri (cfr. Lenardič & Fišer 2022).

3. *Servizi*

Come già anticipato nell'introduzione, la missione del Centro ACE è quella di coadiuvare ricercatrici e ricercatori nelle attività di raccolta, annotazione, analisi, archiviazione e condivisione di dati linguistici "atipici", come quelli che coinvolgono apprendenti di una L2, parlanti con disturbi del linguaggio o disabilità linguistiche, lingue segnate e lingue parlate in contesti multilingui, tutti casi in cui è rilevante la dimensione multimodale della comunicazione e possono essere presi in considerazione diversi stadi o fasi evolutive.

Chi lavora con questo tipo di dati affronta due sfide principali. In primo luogo, si tratta di dati con caratteristiche peculiari in termini di riservatezza e più in generale dal punto di vista etico. Le ricercatrici e i ricercatori sono chiamati a rispettare regole e procedure stringenti imposte da comitati etici e autorità a vari livelli, che prevedono che in tutte le fasi della ricerca siano adottate misure atte a ottenere il consenso informato dei partecipanti ed evitare la divulgazione accidentale di dati sensibili. Nell'Unione Europea la normativa di riferimento è rappresentata in primis dal già menzionato GDPR (cfr. van den Heuvel *et al.* 2020b). A titolo esemplificativo, i minori e i soggetti con gravi disturbi dell'apprendimento potrebbero non essere in grado di fornire autonomamente il loro consenso alla raccolta e alla condivisione di dati che li riguardano, e potrebbe essere necessario, dunque, che un tutore intervenga in loro vece. In questi casi le ricercatrici e i ricercatori potrebbero ritenere opportuno limitare l'accesso a utenti registrati anche laddove il tutore avesse acconsentito alla condivisione (potrebbero per esempio permettere l'accesso solo a coloro che hanno accettato per iscritto di garantire l'anonimato dei soggetti e l'utilizzo dei dati per soli scopi accademici). In presenza di dati particolarmente sensibili, o dati per cui non è stato ottenuto il consenso alla diffusione, potrebbe es-

[22] https://www.clarin.eu/resource-families.

sere necessario conservare i dati originari non anonimizzati in un archivio nascosto, senza che questi vengano copiati o distribuiti in alcun modo o forma. Proprietari e utenti di dati necessitano dunque spesso di consulenza su come conservarli in modo sicuro, a partire dal momento in cui viene generato il dato grezzo e fino a quando i dati e le informazioni secondarie che da esso derivano sono condivise con altri.

In secondo luogo, i dati relativi alla comunicazione atipica rappresentano una sfida per quanto concerne la scelta degli strumenti e dei metodi più adatti alla loro annotazione e alla loro analisi. Le indicazioni e gli strumenti sviluppati per dati "standard" non sono sempre adeguati e richiedono di essere adattati, ed è importante che ricercatrici e ricercatori sappiano quali specifiche indicazioni e quali specifici strumenti sono disponibili (Crasborn 2015).

Alla luce di queste due sfide, il Centro ACE si occupa di fornire informazioni e assistenza in tre modi. Innanzitutto, offre assistenza relativa alla raccolta e alla gestione dei dati tramite informazioni e materiali pubblicati sul suo sito (per esempio moduli per il consenso informato conformi al GDPR), un helpdesk dedicato e consulenze ad hoc per progetti estesi. La procedura necessaria a ottenere il consenso per la raccolta, l'analisi e la condivisione di dati, ad esempio, richiede la massima attenzione laddove siano trattati dati sensibili come videoregistrazioni di interazioni in cui sono coinvolti bambini con disturbi dell'apprendimento. In casi come questo, le procedure per il consenso informato richiedono colloqui scrupolosamente preparati, nonché schede informative e moduli attentamente predisposti e redatti in una lingua chiara e immediatamente comprensibile. Solo in questo modo la persona che rilascia il consenso sarà pienamente consapevole dei possibili usi dei dati e delle modalità secondo le quali saranno protetti e tenuti riservati. Procedure di questo tipo non tutelano solo gli informanti ma accrescono le possibilità di condivisione dei dati, poiché gli informanti saranno più propensi a consentirne la condivisione se comprendono le condizioni in cui saranno archiviati, protetti e riutilizzati.

Inoltre il Centro ACE offre informazioni su metodi e strumenti disponibili per l'elaborazione e l'utilizzo dei dati e consulenza su quali tra questi metodi e strumenti sono più adatti ai dati in questione. Per esempio ELAN, che è stato sviluppato dal gruppo del TLA (ELAN 2020), si presta bene all'annotazione delle lingue segnate perché è stato pensato per dati videoregistrati ed è dotato di un sistema flessibile in cui a ogni riga (*tier*) di trascrizione corrisponde una dimensione del parlato (per es. le espressioni facciali, i gesti, la postura, la direzione dello sguardo), che viene dunque rappresentata simultaneamente alle altre in una sorta di partitura (si veda, per esempio, il corpus di lingua dei segni olandese disponibile nel TLA[23]). Per progetti che riguardano le proprietà acustiche del parlato, invece, il sistema di annotazione più adatto è PRAAT (Boersma & Weenink 2021), che mette a disposizione una gamma di strumenti per l'analisi, la sintesi, la manipolazione e l'etichettatura linguistica, mentre per progetti che richiedono accurate analisi morfosintattiche il sistema più indicato è il CHILDES CLAN (MacWhinney 2000), che contiene un software di annotazione morfosintattica automatica di una serie di lingue (si

[23] https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0004_DF8E_6.

veda, ad esempio, il corpus VALID contenente dati di lingua olandese prodotta da parlanti con disabilità linguistiche[24]). È importante sottolineare l'interoperabilità di molti di questi sistemi, per cui se, per esempio, si codificano produzione verbale e gesti utilizzando ELAN è poi possibile convertire in formato CLAN il documento ottenuto per procedere con l'annotazione morfosintattica.

Infine il Centro ACE offre consulenza su come archiviare dati in modo sicuro sul lungo termine, consentendone la condivisione e il riutilizzo, e fornisce assistenza tecnica nelle fasi di progettazione, creazione, annotazione, formattazione e metadatazione, operazioni cruciali per evitare le difficoltà associate alla consultazione e all'interpretazione di raccolte di dati non etichettati o mal etichettati. Per questo tipo di consulenza è fondamentale la collaborazione con gli esperti del TLA, che sul sito dello stesso hanno messo a disposizione un approfondito manuale sull'archiviazione dei dati e allestito una serie di *screencast* che spiegano come creare raccolte di dati etichettati e organizzati in modo tale da facilitarne il riutilizzo[25]. Per coloro che non raccolgono dati nuovi ma intendono riutilizzarne di già esistenti sono disponibili informazioni su dove trovare corpora e set di dati.

4. *Progetti "in vetrina"*

Il sito web del Centro ACE rimanda a una serie di progetti rappresentativi, tra cui i corpora di parlato di minori e adulti con disturbi del linguaggio compilati nell'ambito del già citato progetto VALID (Klatter *et al.* 2014) e archiviati nel TLA. All'interno di VALID, si è intervenuti su quattro set di dati esistenti in modo da renderli disponibili a scopi di ricerca in un formato compatibile con l'infrastruttura CLARIN. Si tratta dei seguenti:

- il database RU-Kentalis di disturbi specifici del linguaggio (SLI) contenente circa 40 ore di registrazioni audio e 150.000 parole trascritte;
- il database RU-Kentalis di bambini sordi bilingui, contenente circa 9 ore di registrazioni video e 19.500 parole trascritte;
- il Corpus UvA database di ADHD e disturbi specifici del linguaggio (SLI), contenente circa 26 ore di registrazioni video e 23.000 parole trascritte;
- il database RU di adulti sordi, contenente i risultati di un compito di scrittura in formato ScriptLog.

Ulteriori informazioni su queste collezioni sono reperibili alla pagina dedicata al progetto VALID[26], che contiene anche il link all'identificatore persistente dei database curati presso il TLA[27].

[24] https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_8C315BC1_AD5E_4348_9A79_A41FE3DE1150.

[25] https://archive.mpi.nl/forums/c/tla/archiving-info/9.

[26] https://validdata.org/clarin-project/datasets.

[27] https://hdl.handle.net/1839/00-8C315BC1-AD5E-4348-9A79-A41FE3DE1150.

Un altro progetto è il set di dati P-MoLL[28], accessibile a utenti registrati presso il TLA. Il progetto P-MoLL (*Modalität von Lernervarietäten im Längsschnitt*), condotto da Norbert Dittmar alla Libera Università di Berlino tra il 1987 e il 1992, è uno studio longitudinale che documenta l'acquisizione della modalità in tedesco da parte di apprendenti adulti non istruiti di madrelingua italiana e polacca. I dati coprono due anni e mezzo del processo di acquisizione linguistica dei soggetti coinvolti e consistono nelle loro produzioni orali a seguito di elicitazione e sottoforma di conversazioni spontanee con parlanti nativi, per un totale d circa 100 ore di audioregistrazioni, 16 ore di videoregistrazioni e 520.000 parole trascritte (Dittmar *et al.* 1990).

Un ulteriore esempio ben documentato di set di dati di apprendenti di una L2 è il corpus LESLLA (*Literacy Education and Second Language Learning for Adults*)[29], che contiene il parlato di 15 apprendenti donne di olandese come lingua seconda scarsamente istruite, delle quali otto sono turche e sette marocchine (in rappresentanza dei due maggiori gruppi di immigrati nei Paesi Bassi). Alle partecipanti, che durante le registrazioni avevano tra i 22 e i 45 anni, è stato chiesto di completare cinque compiti, tutti collegati alla produzione orale sebbene con livelli diversi di pianificazione (dal parlato rigidamente controllato a quello semispontaneo). Il corpus, disponibile sul TLA[30], consiste in complessive 30 ore circa di audioregistrazioni corrispondenti a circa 180.000 parole trascritte. Una descrizione dettagliata del corpus viene fornita in Sanders *et al.* (2014).

Il corpus LeaP (*Learning Prosody in a Foreign Language*)[31] (Gut 2012) è stato compilato con l'obiettivo di indagare l'acquisizione della prosodia da parte di parlanti non nativi di tedesco e inglese. Le sezioni tedesca e inglese del corpus contengono le registrazioni rispettivamente di 62 e 50 parlanti di diverse L1. Si tratta di 12 ore di audio trascritte e annotate manualmente per un totale di circa 72.000 parole, con etichettatura per parti del discorso (*part-of-speech tagging*) e lemmatizzazione eseguite automaticamente e una descrizione dettagliata del corpus nel manuale allegato.

Il Dutch Bilingual Database[32] (Muysken 2008) è un'altra importante raccolta di dati che rientra nel raggio di azione del Centro ACE ed è ospitata nel TLA. Consiste nei risultati di numerosi progetti e programmi di ricerca volti a studiare il multilinguismo e include dati di parlanti di olandese, sranan, hindustani del Suriname, papiamento, arabo, berbero e turco, per un totale di oltre 500 ore di audioregistrazioni, 10 ore di videoregistrazioni e circa 615.000 parole trascritte, con accesso consentito agli utenti accademici.

[28] https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A.
[29] https://www.leslla.org.
[30] https://hdl.handle.net/1839/00-37EBCC6D-04A5-4598-88E2-E0F390D5FCE1.
[31] https://hdl.handle.net/1839/00-0000-0000-000A-3D5E-1.
[32] https://hdl.handle.net/1839/00-0000-0000-0001-4AF0-7.

Il TLA ospita anche una serie di corpora di lingue segnate, molti dei quali sono stati annotati utilizzando ELAN[33]. Il Corpus NGT (di Lingua dei Segni Olandese;[34,35] cfr. Crasborn & Zwitserlood 2008; Crasborn *et al.* 2008) è una raccolta altamente sistematizzata di 72 ore e 92 segnanti di lingua dei segni olandese videoregistrati da diverse angolazioni e alle prese con diversi compiti e diversi generi testuali. Una porzione significativa dei video è stata annotata manualmente grazie a ELAN, per un totale di circa 200.000 etichette (*annotation tokens*) nella versione più recente. La maggior parte del corpus è liberamente accessibile.

Va precisato che molte delle raccolte di dati che rientrano nella sfera di attività del Centro ACE non rappresentano in realtà sistemi di comunicazione atipici. Le lingue segnate, ad esempio, non sono atipiche, ma sono sistemi linguistici maturi e complessi che si sono evoluti spontaneamente all'interno delle comunità di sordi, analogamente a quanto successo alle lingue parlate nelle comunità di udenti. Tuttavia, come già menzionato, poiché si tratta di lingue che pongono questioni spesso non gestibili da metodi e strumenti standard per la raccolta, l'analisi e l'archiviazione dei dati, è stata adottata una denominazione che fa riferimento alle risorse dedicate dal centro ACE a chi si occupa specificamente di queste lingue.

5. *Lavori in corso*

Questo paragrafo si occupa di corpora accessibili grazie alle collaborazioni attivate nell'ambito del Centro ACE. Nel paragrafo 2 abbiamo già fatto riferimento alla collaborazione con la TalkBank della Carnagie Mellon University. Come esempio di set di dati con registrazione presso la TalkBank e archiviazione dei soli dati primari presso il TLA, abbiamo curato il Corpus di Cued Speech polacco di bambini audiolesi. Si tratta di dati *legacy* di 20 bambini audiolesi di età compresa tra gli 8 e i 12 anni (11 bambine e 8 bambini) messi a disposizione da Anita Trochimyuk-Lorenc e Katarzyna Klessa dell'*Institute of Applied Polish Studies* dell'Università di Varsavia (per una descrizione si vedano Trochymiuk 2003; 2005). L'intervento su questo set di dati ha comportato la creazione di un registro per la metadatazione CMDI e di uno script per la normalizzazione dei nomi dei file e per la loro conversione in formato CHAT – ivi comprese le necessarie *head* che contengono i metadati parzialmente derivabili dai nomi dei file. Per questa raccolta è stata creata una pagina di atterraggio presso la TalkBank[36], le trascrizioni CHAT sono state aggiunte al database TalkBank ed è stato inserito nella pagina di atterraggio l'identificatore persistente (*handle*) per i file audio disponibili presso il TLA[37], in modo da consentire agli utenti di scaricarli direttamente lì. In tal modo è stato possibile rendere il corpus reperibile tramite TalkBank (archivio noto a ricercatrici e ricercatori che

[33] https://tla.mpi.nl/tools/tla-tools/elan.
[34] https://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6.
[35] https://www.ru.nl/corpusngtuk.
[36] https://phonbank.talkbank.org/access/Clinical/PCSC.html.
[37] https://hdl.handle.net/1839/77ea572d-f4c4-48d8-b67b-956f946b59c5.

si occupano di acquisizione della L2 e di disabilità linguistiche) e allo stesso tempo archiviare i dati audio sensibili su server europei protetti da adeguate misure e contratti di licenza. Viste le differenze a livello di struttura e sistemi tra TalkBank e TLA, è stato poi creato uno script per estrarre determinati tipi di file dalle raccolte contenute nel sistema di archiviazione Fedora Commons del TLA e inserirli in una struttura che possa essere facilmente inglobata nella TalkBank. Lo script trasforma inoltre i metadati del TLA in metadati della TalkBank, operazione piuttosto semplice dato che entrambi si basano sullo schema IMDI[38].

Un altro esempio riguarda l'archiviazione di materiali collegati all'ex ospedale neuropsichiatrico di Arezzo. Si tratta delle registrazioni e delle relative trascrizioni di interviste condotte negli anni 70 dalla storica Anna Maria Bruzzone con pazienti dell'ospedale, e del diario di un paziente schizofrenico del medesimo ospedale. Molte delle interviste sono state pubblicate (Bruzzone 2021); tuttavia, le relative audioregistrazioni non sono attualmente accessibili tramite un archivio. Sebbene molti dei pazienti siano deceduti e non rientrino dunque tecnicamente nella giurisdizione del GDPR, è necessario trattare le registrazioni con la massima cura, anche in considerazione dei familiari. L'archiviazione è ancora nella fase iniziale in cui le ricercatrici dell'Università di Siena, alla quale è stata ceduta la raccolta, stanno stabilendo quali dati possono essere condivisi in forma anonima e quali richiedono invece accesso limitato (Calamai *et al.* 2021). È previsto un processo dinamico in base al quale il materiale codificato come non accessibile potrà essere diffuso una volta acquisito il consenso, e il gruppo di lavoro che fa capo a Calamai sta predisponendo per le registrazioni un'accurata metadatazione. Come nel caso del Corpus di Cued Speech polacco di bambini audiolesi, sarà creata una pagina di atterraggio presso la TalkBank, che ospiterà i dati secondari come le trascrizioni, mentre le audioregistrazioni originarie saranno conservate sui server del TLA.

6. *Formazione e sensibilizzazione*

I destinatari delle attività del Centro ACE sono tutti coloro i quali lavorano con set di dati che pongono specifiche sfide in termini di ricerca sulla lingua e sulla comunicazione. Vi rientrano dunque linguisti, psicologi, neuroscienziati, informatici, logopedisti e pedagogisti. Il Centro offre risorse online tramite il relativo sito, un servizio di helpdesk per richieste specifiche e un servizio di consulenza personalizzata per ricercatrici e ricercatori che necessitano di indicazioni su misura.

Il fulcro del programma di assistenza e formazione del Centro ACE è proprio il sito https://ace.ruhosting.nl/, sul quale sono disponibili tutte le informazioni, ivi compresi i link a risorse utili accessibili tramite altri siti, come il TLA e la TalkBank. I servizi del Centro sono inoltre promossi su numerosi altri canali. Per esempio, la sua inaugurazione nel 2019 è stata annunciata in un comunicato stampa pubblicato sui siti della Radboud University e del Max Planck Institute e il personale

[38] https://tla.mpi.nl/imdi-metadata.

del Centro propone iniziative di formazione e sensibilizzazione come conferenze e seminari, ma anche webinar e *screencast* disponibili sul sito (cfr. Draxler *et al.* 2022).

Un primo seminario è stato organizzato in modalità webinar sotto l'egida del progetto SSHOC, in virtù di un'attività legata all'accesso sicuro a dati sensibili nell'ambito del progetto stesso. Il webinar, intitolato *Sharing Datasets of Pathological Speech*[39], si è tenuto il 14 ottobre 2020 e ha toccata i seguenti argomenti:

- progressi dell'iniziativa DELAD nella condivisione di corpora di disturbi del linguaggio (CSD) e ruolo del Centro ACE;
- GDPR e aspetti etici rilevanti per la compilazione e la diffusione di CSD;
- archiviazione e condivisione di CSD in conformità al GDPR presso il TLA e collaborazione con la TalkBank;
- condizioni stabilite dall'infrastruttura per un accesso da remoto sicuro a dati sensibili con specificità di carattere legale (per esempio condizioni di servizio dei social media), etico (per esempio informanti minorenni) e tecnico (audio e video) e valutazione delle piattaforme esistenti;
- progetto CAVA di comunicazione umana audiovisiva – un archivio video digitale a sostegno della comunità internazionale di ricerca sulla comunicazione umana che faciliti la reperibilità e il riutilizzo di contenuti video realizzati a costi elevati;
- cura e diffusione di corpora di patologie del parlato: come reperire CSD attraverso un'organizzazione e renderli accessibili attraverso un'altra; con dimostrazione sul Corpus di Cued Speech polacco di bambini audiolesi (cfr. § 5).

Il webinar è stato registrato e pubblicato su YouTube[40] e le relative diapositive sono disponibili su Zenodo[41]. Un resoconto sottoforma di appunti è disponibile attraverso l'open cloud SSHOC[42].

Tra il 27 e il 28 gennaio 2021 DELAD ha organizzato un seminario dal titolo *How to Share Your Data in a GDPR-Compliant Way*, durante il quale alcune ricercatrici e alcuni ricercatori hanno illustrato i corpora da loro compilati e le analisi condotte sugli stessi, concentrandosi su come poter condividere i dati con altre ricercatrici e altri ricercatori. In ulteriori seminari e presentazioni si è discusso dei seguenti temi:

- possibilità del Centro ACE di ospitare CSD prodotti dai membri dell'iniziativa DELAD;
- scambio di conoscenze ed esperienze in materia di valutazione d'impatto della protezione dei dati (DPIA), anche tramite role-play;
- conversione della voce come strumento di anonimizzazione del parlato.

[39] https://www.sshopencloud.eu/sshoc-webinar-sharing-datasets-pathological-speech.
[40] https://www.youtube.com/watch?v=qjTJ4ZxzfvI.
[41] https://zenodo.org/record/4081602#.X42YC9Azba8.
[42] https://www.sshopencloud.eu/news/webinar-notes-sharing-datasets-pathological-speech.

La DPIA e il role-play sono stati condotti da un membro del Comitato che per CLARIN segue questioni legali e legate alla proprietà intellettuale (CLIC)[43], CLARIN[44] ha pubblicato un resoconto del seminario e tutti i materiali collegati sono stati resi disponibili su Zenodo[45]. È stata inoltre predisposta una versione didattica del role-play sulla DPIA, che è stata pubblicata e presentata al convegno annuale di CLARIN nel 2021[46].

Il Centro ACE ha infine partecipato alla Giornata sullo sviluppo del linguaggio nei bambini (il congresso annuale sul tema che raggruppa ricercatori e logopedisti dei Paesi Bassi e del Belgio, *TOK day*), tenutasi a Nijmegen nel dicembre 2021. Materiale informativo cartaceo come poster, volantini e un documento sintetico di presentazione sono disponibili a partire dalla ripresa delle iniziative in presenza dopo la pandemia da Covid-19.

Bibliografia

Broersma, Paul & Weenink, David. 2021. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.41. http://www.praat.org/

Bruzzone, Anna Maria. 2021. *Ci chiamavano matti. Voci dal manicomio (1968-1977)*. Milano: Il Saggiatore.

Calamai, Silvia & Nodari, Rosalba & van den Heuvel, Henk. 2021. Less is more when FAIR. The minimum level of description in pathological oral and written data. In Monachini, Monica & Eskevich, Maria, (eds.), *CLARIN Annual Conference Proceedings, 2021,* 166-171. Virtual edition.

Crasborn, Onno 2015. Transcription and notation methods. In Orfanidiou, Eleni & Woll, Bencie & Morgan, Gary (eds.), *Research methods in sign language studies: A practical guide*, 74-88. Chichester: John Wiley & Sons.

Crasborn, Onno & Zwitserlood, Inge. 2008. The Corpus NGT: An online corpus for professionals and laymen. In Crasborn, Onno & Hanke, Thomas & Efthimiou, Eleni & Zwitserlood, Inge & Thoutenhoofd, Ernst (eds.), *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and exploitation of sign language corpora,* 44-49. Paris: ELRA.

Crasborn, Onno & Zwitserlood, Inge & Ros, Johan. 2008. The Corpus NGT. A digital open access corpus of movies and annotations of sign language of the Netherlands. Centre for Language Studies, Radboud Universiteit Nijmegen. ISLRN175-346-174-413-3.https://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6

De Jong, Franciska. 2019. CLARIN: Infrastructural support for impact through the study of language as social and cultural data. In Maegaard, Bente & Pozzo, Riccardo & Melloni, Alberto & Woollard, Matthew (eds.), *Stay tuned to the future: Impact of the research infra-*

[43] The roleplay can be found at https://sites.google.com/rug.nl/privacy-in-research/cases.
[44] https://www.clarin.eu/blog/outcomes-fifth-delad-workshop.
[45] https://zenodo.org/record/4560478#.YEeAEJ1Ki71.
[46] Tutti i materiali sono reperibili al link: https://delad.ruhosting.nl/wordpress/dpia-role-play-with-video.

structures for social sciences and humanities (Lessico intellettuale Europeo 128), 121-129. Roma: Leo Olschki.

De Jong, Franciska & Maegaard, Bente & De Smedt, Koenraad & Fišer, Darja & van Uytvanck, Dieter. 2018. CLARIN: Towards FAIR and responsible data science using language resources. *International Conference on Language Resources and Evaluation (LREC)* 11. 3259-3264.

Dittmar, Norbert & Reich, Astrid & Skiba, Romuald & Schumacher, Magdalena & Terborg, Heiner. 1990. Die Erlernung modaler Konzepte des Deutschen durch erwachsene polnische Migranten: Eine empirische Längsschnittstudie. *Informationen Deutsch als Fremdsprache: Info DaF* 17(2). 125-172.

Dittmar, Norbert & Reich, Astrid & Skiba, Romuald & Schumacher, Magdalena & Terborg, Heiner. 2002. The P-MoLL Corpus. https://hdl.handle.net/1839/00-0000-0000-0000-4EAB-A.

Draxler, Christoph & Geyken, Alexander & Hinrichs, Erhard & Klosa-Kückelhaus, Annette, & Teich, Elke & Trippel, Thorsten. 2022. How to connect language resources and infrastructures, and communities. In Fišer, Darja & Witt, Andreas (eds.), *CLARIN. The infrastructure for language resources*, 275-306. Berlin: de Gruyter.

ELAN (Version 6.0) [Computer software]. 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

Gut, Ulrike. 2009. LeaP Corpus. https://hdl.handle.net/1839/00-0000-0000-000A-3D5E-1.

Gut, Ulrike. 2012. The LeaP corpus. A multilingual corpus of spoken learner German and learner English. In Schmidt, Thomas & Wörner, Kai (eds.), *Multilingual corpora and multilingual corpus analysis*, 3-23. Amsterdam: John Benjamins.

Hinrichs, Erhard & Krauwer, Steven. 2014. The CLARIN research infrastructure: Resources and tools for ehumanities scholars. *International Conference on Language Resources and Evaluation (LREC)* 9. 1525-1531.

Kamocki, Paweł & Kelli, Aleksei & Lindén, Krister. 2022. The CLARIN Committee for Legal and Ethical Issues and the Normative Layer of the CLARIN infrastructure. In Fišer, Darja & Witt, Andreas (eds.), *CLARIN. The infrastructure for language resources*, 457-479. Berlin: de Gruyter.

Klatter, Jetske & van Hout, Roeland & van den Heuvel, Henk & Fikkert, Paula & Baker, Anne & de Jong, Jan & Wijnen, Frank & Sanders, Eric & Trilsbeek, Paul. 2014. Vulnerability in acquisition, language impairments in Dutch: Creating a VALID data archive. *International Conference on Language Resources and Evaluation (LREC)* 9. 356-364.

Kolen, Esther. 2014. Bilingual Deaf Children RU-Kentalis Database. ISLRN 941-351-623-486-4. https://hdl.handle.net/1839/00-F6BC06C4-B2AD-4ED8-8527-AB81F4E-F4E8F.

Krauwer, Steven & Maegaard, Bente. 2022. CLARIN – how it started. In Fišer, Darja & Witt, Andreas (eds.), *CLARIN. The infrastructure for language resources*, 3-29. Berlin: de Gruyter.

Lee, Alice & Bessell, Nicola & van den Heuvel, Henk & Saalasti, Satu & Klessa, Katarzyna & Müller, Nicole & Ball, Martin J. 2021. The latest development of the DELAD project for sharing corpora of disordered speech. *Clinical Linguistics & Phonetics*. https://doi.org/10.1080/02699206.2021.1913514.

Lenardič, Jakob & Fišer, Darja. 2022. The CLARIN resource and tool families. In Fišer, Darja & Witt, Andreas (eds.), *CLARIN. The infrastructure for language resources*, 343-372. Berlin: de Gruyter.

Lorenc, Anita. 2019. Polish Cued Speech Corpus of Hearing-Impaired Children. https://hdl.handle.net/1839/77ea572d-f4c4-48d8-b67b-956f946b59c5.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Muysken, Pieter. 2008. Dutch Bilingual Database. https://hdl.handle.net/1839/00-0000-0000-0001-4AF0-7.

Parigger, Esther. 2014. ADHD and SLI Corpus UvA database. ISLRN 456-360-189-350-0. https://hdl.handle.net/1839/00-2766F32F-4305-4F13-A02C-F4A8F5216425.

Sanders, Eric & van de Craats, Ineke & de Lint, Vanja. 2014. The curated Dutch LESLLA corpus. The Dutch LESLLA corpus. *International Conference on Language Resources and Evaluation (LREC)* 9, 2715-2718.

Trochymiuk, Anita. 2003. Voiced realisations of plosives in word initial position by hearing impaired children: Acoustic phonetics analysis. In Böttger, Katharina & Dönninghaus, Sabine & Marzari, Robert (eds.), *Die Welt der Slaven*. Vol. 16 (Beiträge der Europäischen Slavistischen Linguistic 6), 111-123. Munich: Sagner.

Trochymiuk, Anita. 2005. Realization of the voiced-voiceless contrast by hearing impaired children. *Studia Phonetica Posnaniensia* 7. 75-96.

Van den Heuvel, Henk & Oostdijk, Nelleke & Rowland, Caroline & Trilsbeek, Paul. 2020a. The CLARIN knowledge centre for atypical communication expertise. *International Conference on Language Resources and Evaluation (LREC)* 12. 3312-3316.

Van den Heuvel, Henk & Kelli, Aleksei & Klessa, Katarzyna & Salaasti, Satu. 2020b. Corpora of disordered speech in the light of the GDPR: Two use cases from the DELAD initiative. *International Conference on Language Resources and Evaluation (LREC)* 12. 3317-3321.

Van Emmerik. Joanne. 2014. Deaf Adults RU Database. ISLRN 944-022-313-325-3. https://hdl.handle.net/1839/00-97AF29EA-877D-422A-BAF7-25FA269351A6.

Van der Made, Annika. 2014. SLI RU-Kentalis Database. ISLRN 541-534-411-504-6. https://hdl.handle.net/1839/00-712802F3-C245-4EF0-BE9D-D09714DEDE67.

Windhouwer, Menzo & Goosen, Twan. 2022. Component Metadata Infrastructure. In Fišer, Darja & Witt, Andreas (eds.), *CLARIN. The infrastructure for language resources*, 191-222. Berlin: de Gruyter.