Literature about data-sharing in corpus-based research on talk-in-interaction: a collection of references

CHORD-Talk-in-interaction

January 2025

Table of Contents

Introduction 1
1. Corpus Design & Corpus Structure 1
1.1. General considerations 1
1.2. Reference corpora 2
2. Recording & Processing Spoken Interaction 3
3. Transcription & Annotation 4
4. Data Storage 6
5. Data Use 7
6. Data Citation 8
7. Law, Ethics & Principles 9
8. Corpus Hosting Platforms 10

Introduction

Concepts, norms, practices and tools related to data-sharing in linguistics and in the social sciences form a complex and rapidly changing socio-technical reality within the wider context of academic research and Open Science policy making. The scientific, and normative, debate on these practices takes place in an interdisciplinary field, with contributions from the interested disciplines, computational sciences, law, ethics, social and political science. The document at hand has been prepared with the intention of providing some guidance for bibliographical research in this wide field from the point of view of interactional linguistics, with special attention to corpus-based approaches. It lists references to works that appeared in the last 10-20 years and can be associated to eight topics: corpus design and corpus structure; recording and processing spoken interaction; transcription and annotation; data storage; data use; data citation; law, ethics and principles; corpus hosting.

The list of references has been compiled within the project *Data-sharing skills in corpus-based research on talk-in-interaction* (CHORD-Talk-in-interaction), an exploratory investigation funded by the swissuniversities program on open research data, which was conducted from April 2023 to September 2024 and involved partners from USI Università della Svizzera italiana (leading house) and the universities of Basel, Neuchâtel and Lausanne (cf. https://www.chord-talk-in-interaction.usi.ch/cti/team). The document contains references in English, German and French. It reflects discussions that took place during a series of workshops organized within the CHORD-Talk-in-interaction project. While it can only sketch a temporary, and fragmentary, map of the vast areas that are potentially relevant, the authors nevertheless hope to provide some useful starting points for bibliographical research and to stimulate further reading about open research data in interactional linguistics and in the neighbouring disciplines.

1. Corpus Design & Corpus Structure

1.1. General considerations

Dimitriadis, A., & Musgrave, S. (2009). Designing linguistic databases: A primer for linguists. In M. Everaert, S. Musgrave & A. Dimitriadis (Eds.), *The Use of Databases in Cross-Linguistic Studies* (pp. 13–75). De Gruyter Mouton. https://doi.org/10.1515/9783110198744.13

Empfehlungen des DFG-Fachkollegiums 104 “Sprachwissenschaften". (2019). *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*

Ferger, A., Krause, A. F., & Pitsch, K. (2023). Workflows and Methods for Creating Structured Corpora of Multimodal Interaction. Paper presented at the *10th International Conference on CMC and Social Media Corpora for the Humanities,* Mannheim, Germany. 73–77. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/12095

Robert, S. (2013). *Bilan intermédiaire du Consortium Corpus Oraux et Multimodaux de la TGIR Huma-Num.* https://hal.science/hal-01494688

Säily, T., & Tyrkkö, J. (2021). Challenges of combining structured and unstructured data in corpus development. *Research in Corpus Linguistics (RiCL), 9*(1), i–viii. https://doi.org/10.32714/ricl.09.01.01

Schmidt, T. (2018). Gesprächskorpora. Aktuelle Herausforderungen für einen besonderen Korpustyp. In M. Kupietz, & T. Schmidt (Eds.), *Korpuslinguistik* (pp. 209–230). De Gruyter. https://doi.org/10.1515/9783110538649-010

Schmidt, T. (2022). Daten und Metadaten. In M. Beißwenger, L. Lemnitzer & C. Meyer-Spitzer (Eds.), *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium* (pp. 249–258). Wilhelm Fink.

1.2. Reference corpora

Avanzi, M., Béguelin, M.-J., Corminboeuf, G., †Diémoz, F. & Johnsen, L.A. (2012-2023). *OFROM – corpus oral de français de Suisse romande*, Université de Neuchâtel, ofrom.unine.ch

Anderson, Jean, David Beavan and Christian Kay (2007) 'SCOTS: Scottish Corpus of Texts and Speech', in Joan Beal, Karen Corrigan, Hermann Moisl, eds, *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, Basingstoke: Palgrave Macmillan

Ballarè, S., E. Goria & C. Mauri. 2022. *Spoken Italian and linguistic variation. Theory and practice in the construction of the KIParla corpus*. Bologna: Pàtron Editore.

Boves, L. & N. Oostdijk. Spontane Sprache im Spoken Dutch Corpus. In *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*. 14.-16. April 2003. Tokio, Japan .

Čermák, F. 2009, The Case of the Czech National Corpus: Its Design, Development and Current State, Katsumasa Yagi, Takaaki Kanzaki, eds., 2009, *Phraseology, Corpus Linguistics and Lexicography: Papers from Phraseology in Japan*, Kwansei Gakuin University Press, Nishinomiya, (rozšířená a přepracovaná verze, v tisku pro Proceedings of PALC 2008) 39-76.

Coleman, J., Baghai-Ravary, L., Pybus, J. & Grau, S. (2012). *Audio BNC: the audio edition of the Spoken British National Corpus.* http://www.phon.ox.ac.uk/AudioBNC.

Deppermann, A., & Hartung, M. (2011). Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In E. Felder, M. Müller & F. Vogel (Eds.), *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen* (pp. 414–450). De Gruyter. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/2985

Haugh, Michael, Kate Burridge, Jean Mulder and Pam Peters (eds.) (2009). *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*. Cascadill Proceedings Project, Sommerville, MA.

Hoffmann, S., Evert, S., Smith, N., David, Y. W. L., & Berglund, Y. (2009). *Corpus Linguistics with «BNCweb» – a Practical Guide*. Peter Lang Verlag.

Ide, N. (2008). The American National Corpus: Then, Now, and Tomorrow. In Michael Haugh, Kate Burridge, Jean Mulder and Pam Peters (eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, Cascadilla Proceedings Project, Sommerville, MA. http://www.lingref.com/cpp/ausnc/2008/paper2293.pdf

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., & Vangsnes, Ø A. (2009). The Nordic Dialect Corpus — an advanced research tool. Paper presented at the *17th Nordic Conference of Computational Linguistics NODALIDA,* 73–80. http://hdl.handle.net/10062/9730

Kaiser, J. (2019). Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion,* 19 (2018), 515–552. http://www.gespraechsforschung-online.de/2018.html

Küttner, U., Kornfeld, L., Mack, C., Mondada, L., Rogowska, J., Rossi, G., ... Zinken, J. (2024). Introducing the "Parallel European Corpus of Informal Interaction" (PECII). *New Perspectives in Interactional Linguistic Research* (pp. 132–160) John Benjamins Publishing Company. http://www.degruyter.com/doi/10.1075/slsi.36.05kut

Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.) (2012). *Narodowy Korpus Języka Polskiego (National Corpus of Polish Language)*. Warsaw: Wydawnictwo Naukowe PWN.

Савчук С. О., Архангельский Т. А., Бонч-Осмоловская А. А., Донина О. В., Кузнецова Ю. Н., Ляшевская О. Н., Орехов Б. В., Подрядчикова М. В. Национальный корпус русского языка 2.0: новые возможности и перспективы развития. Вопросы языкознания, 2024, 2: 7–34. https://ruscorpora.ru/en

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation, 47*(4), 1031–1048. https://doi.org/10.1007/s10579-013-9216-5

Voghera, M., Iacobini, C., Savy, R., Cutugno, F., De Rosa, A., & Alfano, I. (2014). VoLIP: a searchable Italian spoken corpus. In L. Veselovská, & M. Janebová (Eds.), *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium: Language Use and Linguistic Structure* (pp. 628–640). Olomouc: Palacký University.

2. Recording & Processing Spoken Interaction

Garfinkel, S. (2015). *De-identification of personal information.* (No. 8053). NIST National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8053

Hoey, E. M., & Raymond, C. W. (2022). Managing Conversation Analysis Data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management* (pp. 257–266). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0025

Hukkelås, H., & Lindseth, F. (2023). DeepPrivacy2: Towards realistic full-body anonymization. Paper presented at the *Winter Conference on Applications of Computer Vision (WACV),* Waikoloa, Hawaii. 1329–1338. https://doi.org/10.1109/WACV56688.2023.00138

Jouin-Chardon, E., Mondada, L., Niccolai, G. P., & Traverso, V. (2010). Contraintes technologiques sur les enregistrements de corpus et analyse des cadres de participation. *Pratiques, 147-148*, 53–81. https://doi.org/10.4000/pratiques.1606

Kleiner, B. & Heers, M. (2024). *Quantitative data anonymisation: practical guidance for anonymising sensitive social science data*. FORS Guide, 23, Version 1.0, 1-17. https://doi:10.24449/FG-2024-00023

[moca] - multimodal oral corpus administration. http://moca.phil2.uni-freiburg.de/web/index.html

Mondada, L. (2005). Constitution de corpus de parole-en-interaction et respect de la vie privée des enquêtés : une démarche réflexive. *Rapport sur le projet « Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité » financé par le Programme Société de l'Information / Archivage et patrimoine documentaire.* (pp. 1–43)

Mondada, L. (2012). The Conversation Analytic Approach to Data Collection. In J. Sidnell, & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 32–56). John Wiley & Sons. https://doi.org/10.1002/9781118325001.ch3

Mondada, L. (2014). Ethics in action: Anonymization as a participant's concern and a participant's practice. *Human Studies, 37*, 179–209. https://doi.org/10.1007/s10746-013-9286-9

Owoyele, B. A., Schilling, M., Sawahn, R., Kaemer, N., Zherebenkov, P., Verma, B., ... de Melo, G. (2024). In Vogel D. R., Gewald H., Sapsomboon A., Cheung C. M. K., Laumer S. and Thatcher J.(Eds.), *MaskAnyone Toolkit : Offering Strategies for Minimizing Privacy Risks and Maximizing Utility in Audio-Visual Data Archiving*. Association for Information Systems. https://aisel.aisnet.org/icis2024/adv_theory/adv_theory/1

Pätzold, H. (2005). Secondary analysis of audio data. Technical procedures for virtual anonymization and pseudonymization. *Forum: Qualitative Social Research, 6*(1).

Põldvere, N., Frid, J., Johansson, V., & Paradis, C. (2021). Challenges of releasing audio material for spoken data: The case of the London-Lund Corpus 2. *Research in Corpus Linguistics, 9*(1), 35–62. https://doi.org/10.32714/ricl.09.01.04

Potter, J., & Humă, B. (2025). Ensuring Quality: The Power and Potential of Naturally Occurring Data in the Social Sciences. *The Sage Handbook of Qualitative Research Quality* (pp. 185–200). 55 City Road London: Sage Publications Ltd.10.4135/9781529674354.n12. https://sk.sagepub.com/reference/the-sage-handbook-of-qualitative-research-quality/i1506.xml

Schmidt, T. (2014). (More) common ground for processing spoken language corpora? In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Corpora in Linguistic Research* (pp. 249–265). Cambridge Scholars Publishing.

Stam, A, Diaz, P. (2023). *Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts.* FORS Guide No. 20, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. https://doi.org/10.24449/FG-2023-00020

Tomashenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J., Paul-Gauthier, N., & Todisco, M. (2020). Introducing the VoicePrivacy Initiative. *Proc. Interspeech 2020* (pp. 16393–1997). https://doi.org/10.21437/Interspeech.2020-1333

Traverso, V. (2022). Anonymisation, pseudonymisation, consentement. Réflexions à partir d'expériences de collectes de données vidéo sur le terrain. *Sonorités, 48*, 26–51.

Wiles, R., Coffey, A., Robinson, J., & Heath, S. (2012). Anonymisation and visual images: issues of respect, 'voice' and protection. *International Journal of Social Research Methodology, 15*(1), 41–53. https://doi.org/10.1080/13645579.2011.564423

3. Transcription & Annotation

Barceló, G., Bonu, B., Détrie, C., Fauré, L., Lefeuvre, F., Leroy, S., ... Traverso, V. (2002). Transcrire l'interaction. *Cahiers de praxématique,* (39). https://doi.org/10.4000/praxematique.452

Bergounioux, G., Jacobson, M., & Pietrandrea, P. (2018). Annotating oral corpora. In W. Ayres-Bennett, & J. Carruthers (Eds.), *Manual of Romance Sociolinguistics* (pp. 27–58). De Gruyter. https://doi.org/10.1515/9783110365955-002

BigSoftVideo. *DOTE - A New Type of Transcription Software for Social Conduct and Multimodal Interaction.* Retrieved from https://www.dote.aau.dk/

Brugman, H., & Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. Paper presented at the *4th International Conference on Language Resources and Evaluation (LREC'04),* 2065–2068. https://aclanthology.org/L04-1285/

Brunner, M., & Diemer, S. (2021). Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription. *Research in Corpus Linguistics, 9*(1), 63–88. https://doi.org/10.32714/ricl.09.01.05

Cassidy, S., & Schmidt, T. (2017). Tools for multimodal annotation. In N. Ide, & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 209–227). Springer. https://doi.org/10.1007/978-94-024-0881-2_7

Centre for Language and Speech Technology, Radboud University. *FoLiA Format for Linguistic Annotation.* http://proycon.github.io/folia/

Couper-Kuhlen, E., & Barth-Weingarten, D. (2011). A system for transcribing talk-in-interaction: GAT 2 (translation and adaption). *Gesprächsforschung, 12*, 1–51. http://www.gespraechsforschung-online.de/heft2011/heft2011.html

Dister, A. & Simon, A. (2007). La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé. *Arena Romanistica, 1*(1), 54. https://dial.uclouvain.be/pr/boreal/object/boreal:83571

Finlayson, M. A., & Erjavec, T. (2017). Overview of annotation creation: Processes and tools. In N. Ide, & J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 167–191). Springer. https://doi.org/10.1007/978-94-024-0881-2_5

Hepburn, A., & Bolden, G. B. (2012). The conversation analytic approach to transcription. In J. Sidnell, & T. Stivers (Eds.), *The Handbook of Conversation Analysis* (pp. 57–76). Wiley-Blackwell. https://doi.org/10.1002/9781118325001.ch4

Hepburn, A., & Bolden, G. B. (2017). *Transcribing for Social Research*. SAGE Publications Ltd. https://doi.org/10.4135/9781473920460

Ide, & J. Pustejovsky (Eds.), *Handbook of linguistic annotation*. Springer.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation Analysis: Studies from the First Generation* (pp. 13–31). John Benjamins. https://doi.org/10.1075/pbns.125.02jef

Klie, J., Bugert, M., Boullosa, B., Castilho, R. E. d., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations,* Santa Fe, USA. 5–9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Laurier, E. (2014). The Graphic Transcript: Poaching Comic Book Grammar for Inscribing the Visual, Spatial and Temporal Aspects of Action. *Geography Compass, 8*(4), 235–248.https://api.istex.fr/ark:/67375/WNG-F8W9V4JD-4/fulltext.pdf

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion,* (11), 154–173. http://www.gespraechsforschung-online.de/heft2010/heft2010.html

Mondada, L. (2008a). Documenter l'articulation des ressources multimodales dans le temps : la transcription d'enregistrements vidéos d'interactions. In M. Bilger (Ed.), *Données orales, les enjeux de la transcription,* (pp. 127–155). Presses Universitaires de Perpignan.

Mondada, L. (2008b). La transcription dans la perspective de la linguistique interactionnelle. In M. Bilger (Ed.), *Données orales, les enjeux de la transcription,* (pp. 78–109). Presses Universitaires de Perpignan.

Mondada, L. (2018). Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Social Interaction, 51*(1), 85–106. https://doi.org/10.1080/08351813.2018.1413878

Reimer, E., Ullrich, A. V., Trevisan, B., & Jakobs, E. (2017). Mehrebenenannotation multimodaler Daten. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion, 18*, 91–116. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2017/px-reimer.pdf

Rühlemann, C., & Gee, M. (2017). Conversation analysis and the XML method. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion,* (18), 274–296. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2017/px-ruehlemann.pdf

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative,* (1) https://doi.org/10.4000/jtei.142

Schmidt, T., Schwendemann, M., & Wallner, F. (2023). ZuViel: Transkriptvisualisierung und arbeiten mit Transkripten. *Korpora Deutsch Als Fremdsprache, 3*(1), 72–91. https://doi.org/10.48694/kordaf.3723

Selting, M., Schütte, W., Stukenbrock, A., Uhmann, S., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Deppermann, A., Gilles, P., Günthner, S., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Auer, P., ... Bergmann, J. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion, 10*, 353–402.

Sloetjes, H., & Seibert, O. (2016). Measuring by marking; the multimedia annotation tool ELAN. In A. J. Spink, G. Riedel, L. Zhou, L. Teekens, R. Albatal & C. Gurrin (Eds.), *Measuring Behavior 2016, 10th International Conference on Methods and Techniques in Behavioral Research* (pp. 492–495). Dublin City University.

TEI Consortium. (2023, 16.11.). *8 Transcriptions of Speech.* Retrieved from https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

The Language Archive (2017). *ELAN Annotation Format EAF*. Schema version 3.0. Max Planck Institute for Psycholinguistics. https://www.mpi.nl/tools/elan/EAF_Annotation_Format_3.0_and_ELAN.pdf

Umair, M., Mertens, J. B., Albert, S., & De Ruiter, J. P. (2022). GailBot: An automatic transcription system for Conversation Analysis. *Dialogue and Discourse, 13*(1), 63–95. https://doi.org/10.5210/dad.2022.103

van Gompel, M., & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal, 3*, 63–81. https://clinjournal.org/clinj/article/view/26/22

Walker, S. (2019). *Post-processing speech-to-text transcripts. State-of-the-art survey*. ZHAW InIT.

WebLicht. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

Weisser, M. (2016). DART – The dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory, 12*(2), 355–388. https://doi.org/10.1515/cllt-2014-0051

Westpfahl, S., Schmidt, T., Jonietz, J., & Borlinghaus, A. (2017). *STTS 2.0. Guidelines für die Annotation von POS -Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS).* Unpublished manuscript. Retrieved from https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063

4. Data Storage

Andreassen, H. N. (2022). Archiving Research Data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management* (pp. 89–100). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0011

CLARIN ERIC. *CLARIN Zotero Library.* https://www.zotero.org/groups/562080/clarin/library

CMDI and Metadata Curation task forces of the Standing Committee on CLARIN Technical Centres (2017). *CMDI Best Practices Guide*, v. 1.2.0. https://www.clarin.eu/content/cmdi-best-practices-guide

Drude, S., Trilsbeek, P., Sloetjes, H., & Broeder, D. (2014). Best practices in the creation, archiving and dissemination of speech corpora at the Language Archive. In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Corpora in Linguistic Research* (pp. 183–207). Cambridge Scholars Publishing.

Ferger, A., & Hedeland, H. (2020). Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation, 15*(1), 1–13. https://doi.org/10.2218/ijdc.v15i1.601

Fischer, P. M., & Witt, A. (2014). Best practices on long-term archiving of spoken language data. In S. Ruhi, M. Haugh, T. Schmidt & K. Wörner (Eds.), *Best Practices for Spoken Corpora in Linguistic Research* (pp. 162–182). Cambridge Scholars Publishing.

Hedeland, H., & Schmidt, T. (2021). The TEI-based ISO standard “Transcription of Spoken Language” as an exchange format within CLARIN and beyond. Paper presented at the *CLARIN Annual Conference 2021,* Online. 100–104. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/10717

Hinrichs, E., & Vogel, I. (Eds.). (2010). *Interoperability and Standards*. Common Language Resources and Technology Infrastructure CLARIN.

ISO 24624:2016 Language resource management - Transcription of spoken language. (2016). Retrieved from https://www.iso.org/standard/37338.html

Liégeois, L. (2013). De l'analyse au partage des données, quel(s) format(s) choisir ? L'exemple d'un corpus d'interactions parents-enfant. In M. Damiani, K. Dolar, C. Florez-Pulido, R. Loth, J. Magnier & A. Pegaz (Eds.), *Traitement de corpus (Actes de Coldoc 2012)* (pp. 128– 142). Paris, France: Modyco. Retrieved from https://hal.science/hal-00850172

Mattern, E. (2022). The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management. In A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management* (pp. 61–71). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0009

Parisse, C., & Morgenstern, A. (2010). A multi-software integration platform and support for multimedia transcripts of language. Paper presented at the *2010 Language Resources and Evaluation Conference Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality,* La Valette, Malta. 106–110.

Rosenthaler, L., Fornaro, P., & Clivaz, C. (2015). DASCH: Data and Service Center for the Humanities. *Digital Scholarship in the Humanities, 30*(1), i43–i49. https://doi.org/10.1093/llc/fqv051

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., & Sloetjes, H. (2008). An exchange format for multimodal annotations. Paper presented at the *6th International Conference on Language Resources and Evaluation (LREC'08),* Marrakech, Morocco. https://aclanthology.org/L08-1032/

Schmidt, T., Wörner, K., Hedeland, H., & Lehmberg, T. (2013). Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache., 1–22. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/1331

Weigel, T., Schwardmann, U., Klump, J., Bendoukha, S., & Quick, R. (2020). Making data and workflows findable for machines. *Data Intelligence, 2*(1-2), 40–46. https://doi.org/10.1162/dint_a_00026

5. Data Use

André, V., Benzitoun, C., Canut, E., jeanne-marie, D., Gaiffe, B., & Jacquey, E. (2010). Traitement informatique de données orales: quels outils pour quelles analyses? *Recherches Qualitatives – Hors Série «Logiciels Pour L'analyse Qualitative : Innovations Techniques Et Sociales,* (9), 131–150.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research, 30*(2), 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

Banski, P., Frick, E., & Witt, A. (2016). Corpus Query Lingua Franca (CQLF). Paper presented at the *Tenth International Conference on Language Resources and Evaluation (LREC 2016),* 2804–2809. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/5040/file/Banksi_Frick_Witt_Corpus_Query_Lingua_Franca_2016.pdf

Batinic, J., Frick, E., & Schmidt, T. (2021). Accessing spoken language corpora: An overview of current approaches. *Corpora, 16*(3), 417–445. https://doi.org/10.3366/cor.2021.0229

Biliotti, F., Calamai, S., & Ginouvès, V. (2018). Les archives sonores entre demande sociale et usages scientifiques. Quelles modalités pour réutiliser les sources enregistrées? In V. Ginouvès, & I. Gras (Eds.), *La diffusion numérique des données en SHS. Guide des bonnes pratiques éthiques et juridiques* (pp. 169–194). Presses universitaires de Provence.

Charaudeau, P. (2009). Dis-moi quel est ton corpus, je te dirai quelle est ta problématique. *Corpus,* (8), 37–66. https://doi.org/10.4000/corpus.1674

Debras, C. (2018). Petits et grands corpus en analyse linguistique des gestes. *Corpus,* (18), 1–21. https://doi.org/10.4000/corpus.3287

Druskat, S., Gast, V., Krause, T., & Zipser, F. (2016). corpus-tools.org: An interoperable generic software tool set for multi-layer linguistic corpora. Paper presented at the *10th International Conference on Language Resources and Evaluation (LREC'16),* Portorož, Slovenia. 4492–4499. https://aclanthology.org/L16-1711

Ehmer, O. (2023). Arbeiten mit zeitalignierten multimodalen Korpora in R. Vorstellung des Aligned Corpus Toolkit (act). *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion, 24*, 67–126. http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2023/px-ehmer.pdf

Fandrych, C., Portmann, A., Schmidt, T., Schwendemann, M., Wallner, F., Wörner, K., Frick, E., Kaiser, J., & Meißner, C. M. (2022). ZuMult: Neue Zugangswege zu Korpora gesprochener Sprache. In H. Kämper, & A. Plewnia (Eds.), *Sprache in Politik und Gesellschaft: Perspektiven und Zugänge* (pp. 305–312). De Gruyter. https://doi.org/10.1515/9783110774306-018

Fandrych, C., Frick, E., Hedeland, H., Iliash, A., Jettka, D., Meißner, C., Schmidt, T., Wallner, F., Weigert, K., & Westpfahl, S. (2016). User, who art thou? User profiling for oral corpus platforms. Paper presented at the *10th International Conference on Language Resources and Evaluation (LREC'16),* Portorož, Slovenia. 280–287. https://aclanthology.org/L16-1043

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har

ISO 24623-1:2018 Language resource management - Corpus query lingua franca (CQLF) - Part 1: Metamodel. Retrieved https://www.iso.org/standard/37337.html

MacWhinney, B. (2017). TalkBank and CLARIN. Paper presented at the *CLARIN Annual Conference 2016,* Aix-en-Provence, France. 76–89. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=136&Article_No=6

Merkel, S., & Schmidt, T. (2009). Korpora gesprochener Sprache im Netz - eine Umschau. *Online-Zeitschrift Zur Verbalen Interaktion, 10*, 70–93. http://www.gespraechsforschung-ozs.de/fileadmin/dateien/heft2009/px-merkel.pdf

Pezik, P. (2014). Spokes – a search and exploration service for conversational corpus data. Paper presented at the *3rd CLARIN Annual Conference,* Soesterberg, Netherlands. 99–109. https://ep.liu.se/ecp/116/009/ecp15116009.pdf

Schmidt, T., Hedeland, H., & Jettka, D. (2017). Conversion and annotation web services for spoken language data in CLARIN. Paper presented at the *CLARIN Annual Conference 2016,* Aix-en-Provence, France. 113–130.

Traverso, V. (2011). Analyser un corpus de langue parlée en interaction: questions méthodologiques. *Verbum, 4*, 313–329. https://shs.hal.science/halshs-00658840

Uytvanck, &., Dieter Van, Olsson, &., Leif-Jöran, Schonefeld, O., Eckart, &., ... Illig, E. M. (2023). *CLARIN Federated Content Search (CLARIN-FCS) - Core 2.0, v. 1.0.* https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf

6. Data Citation

Bornatici, C. & Fedrigo, N. (2023). *Data Citation: How and Why Citing (Your Own) Data*. FORS Guide No. 19, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. https://doi.org/10.24449/FG-2023-00019

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. FORCE11. https://doi.org/10.25490/a97f-egyk

Parisse, C., & Sadoun, D. (2022, December 12). Projet CITATION: CORPUCIT. Fournir des outils pour faciliter la citation de corpus ou d'extraits de corpus. Slides presented at the *Journée D'Information Et De Réflexion Sur L'Éthique De La Recherche, Atelier "Ethique Dans Les Humanités Numériques : La Citation Des Données De Recherche (Pratiques, Formats, Outils),* https://hal.science/hal-03947375

7. Law, Ethics & Principles

Badin, F., Cance, C., Dugua, C., Kanaan-Caillol, L., Minard, A., & Ploog, K. (2022). Les données orales en linguistique. *Le bulletin de l'AFAS,* (48), 158–181. https://doi.org/10.4000/afas.7496

Baude, O., Blanche-Benveniste, C., Calas, M., Cappeau, P., Cordereix, P., Goury, L., Jacobson, M., de Lamberterie, I., Marchello-Nizia, C., & Mondada, L. (2006). *Corpus oraux. Guide des bonnes pratiques*. Presses universitaires d'Orléans.

Branche, R., Descamps, F., Saffroy, F., & Vaisse, M. (2018). La parole et le droit: Recommandations pour la collecte, le traitement et l'exploitation des témoignages oraux. In V. Ginouvès, & I. Gras (Eds.), *La diffusion numérique des données en SHS. Guide de bonnes pratiques éthiques et juridiques* (pp. 103–127). Presses universitaires de Provence.

Collister, L. B. (2022). Copyright and Sharing Linguistic Data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller & L. B. Collister (Eds.), *The Open Handbook of Linguistic Data Management* (pp. 117–128). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0013

D'Arcy, A., & Bender, E. M. (2023). Ethics in Linguistics. *Annual Review of Linguistics, 9*, 49–69. https://doi.org/10.1146/annurev-linguistics-031120-015324

Federal Act on Data Protection, (2020). https://www.fedlex.admin.ch/eli/cc/2022/491/en

General Data Protection Regulation (GDPR). https://gdpr.eu/tag/gdpr/

Ginouvès, V., & Gras, I. (2018). *La diffusion numérique des données en SHS - Guide des bonnes pratiques éthiques et juridiques*. Presses universitaires de Provence.

Ginouvès, V., & Traverso, V. (2022). Le droit et l'éthique: qu'est-ce qui change dans les pratiques de terrain? – Introduction. *Le bulletin de l'AFAS,* (48), 8–24. https://doi.org/10.4000/afas.7470

GO FAIR initiative (2022). *FAIR principles*. https://www.go-fair.org/fair-principles/

Huyghe, M., Cailly, L., & Oppenchaim, N. (2018). Ouverture de données qualitatives à caractère personnel Approche éthique, juridique et déontologique. In V. Ginouvès, & I. Gras (Eds.), *La diffusion numérique des données en SHS Guide des bonnes pratiques éthiques et juridiques* (pp. 159–168). Presses universitaires de Provence.

Jouin, É, & Ticca, A. C. (2022). Le dilemme de la recherche. *Le Bulletin De L'AFAS,* (48), 52–66. https://doi.org/10.4000/afas.7475

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019). The GDPR & speech data: Reflections of legal and technology communities, first steps towards a common understanding. Paper presented at the *20th Annual Conference of the International Speech Communication Association,* Graz, Austria. 3695–3699. https://doi.org/10.48550/arXiv.1907.03458

Polonetsky, J., Tene, O., & Finch, K. (2016). Shades of gray: Seeing the full spectrum of practical data de-identification. *Santa Clara Law Review, 56*(3), 593–629. https://digitalcommons.law.scu.edu/lawreview/vol56/iss3/3/

Speer, S. A. (2014). Reflecting on the Ethics and Politics of Collecting Interactional Data: Implications for Training and Practice. *Human Studies [Special Issue], 37*(2), 279–286. https://www.jstor.org/stable/24022226

Stam, A., & Kleiner, B. (2020). *Data anonymization: legal, ethical, and strategic considerations*. FORS Guide No. 11, Version 1.1 (last update January 2022) Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. https://doi.org/10.24449/FG-2020-00011

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

8. Corpus Hosting Platforms

Alcón, D. *moca3.* http://moca.phil2.uni-freiburg.de/moca3_v3/index.php.

Baldauf-Quilliatre, H., Carvajal, I. C. d., Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., & Traverso, V. (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus,*15 https://doi.org/10.4000/corpus.2991

*Bavarian Archive for Speech Signals (BAS) CLARIN Repository.*https://clarin.phonetik.uni-muenchen.de/BASRepository/index.php.

CLAPI Corpus de LAngue Parlée en Interaction. http://clapi.ish-lyon.cnrs.fr/V3_Accueil.php?interface_langue=EN.

CLARIN ERIC. *CLARIN certified B-centres.* https://www.clarin.eu/content/certified-b-centres.

CLARIN ERIC. *CLARIN Virtual Language Observatory.* https://vlo.clarin.eu/?2.

CoCoON COllections de COrpus Oraux Numériques. https://cocoon.huma-num.fr/exist/crdo?lang=en. https://cocoon.huma-num.fr/exist/crdo?lang=en

DaSCH - Swiss National Data and Service Center for the Humanities. https://www.dasch.swiss/

Forschung am KFM | Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit. https://centre-plurilinguisme.ch/de/kfm_forschung?keyword_id%5B%5D=241&people=&year_from=0&year_to=0&op=Suche.

French Learner Language Oral Corpora (flloc). http://www.flloc.soton.ac.uk/index.html.

Hamburger Zentrum für Sprachkorpora. https://www.fdr.uni-hamburg.de/communities/hzsk/search?page=1&size=20&q=&type=dataset&type=video.

IRCOM Consortium Corpus Oraux & Multimodaux. http://ircom.huma-num.fr/site/corpus.php?discipline=Interaction.

Kielipankki - The Language Bank of Finland. https://www.kielipankki.fi/.

Language Infrastructure made Accessible (LIA). https://www.hf.uio.no/iln/english/research/projects/language-infrastructure-made-accessible/index.html.

Language repository of Switzerland (LaRS) @ SWISSUbase. https://www.swissubase.ch/en/catalogue/search?q=&p=0&ps=10&sn=ref-number&sd=desc.

Leibniz-Institut für Deutsche Sprache. *Archiv für Gesprochenes Deutsch (AGD).* https://agd.ids-mannheim.de/index.shtml. https://agd.ids-mannheim.de/index.shtml

LiRI corpus platform. https://www.liri.uzh.ch/en/services/LiRI-Corpus-Platform-LCP.html

META-SHARE. https://metashare.ut.ee/.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods, 51*(4), 1919–1927. https://doi.org/10.3758/s13428-018-1174-9

ORTOLANG – Plate-forme d'outils et de ressources linguistiques pour un traitement optimisé de la langue française. https://www.ortolang.fr/fr/accueil/.

PFC – Phonologie du Français Contemporain. https://www.projet-pfc.net/. https://www.projet-pfc.net/