Databases for the Study of Aphasia

Nichol Castro, Department of Communicative Disorders and Sciences, University at Buffalo, Buffalo, NY, United States

Key Points

- Linguistic databases of aphasia
- Recovery databases of aphasia

Abstract

This article describes databases used in the study of aphasia, a language disorder disrupting production and/or comprehension. Examples of two types of databases are provided: linguistic and recovery databases. Linguistic databases focus (predominantly) on spoken language of people with aphasia in the pursuit of characterizing the variety of language impairments that emerge with aphasia. Recovery databases focus on how aphasia changes over time and may or may not include active treatment delivery and monitoring. Brief examples of how these databases have been used are provided, along with a closing commentary of benefits and limitations.

Introduction

Aphasia is a language disorder that disrupts the ability to communicate with others. Language impairments due to aphasia can emerge in a variety of ways, including difficulty with comprehension and/or production of language in spoken, written, and/or signed modalities. This results in aphasia being a heterogeneous disorder. The impact of aphasia on language affects many aspects of a person's life beyond just their language knowledge and skills, including their mental health, participation in desired activities, and maintenance of social relationships. The understanding of aphasia and its impact across a person's life requires the analysis of a wide variety of qualitative and quantitative data.

Databases provide one way in which researchers and clinicians can better assess and understand the language impairments of aphasia (Faroqi-Shah, 2016). This article will provide a review of several commonly used databases, with examples of how they have been used to advance research and clinical practice. Additional databases relevant to neurorehabilitation are described in Faroqi-Shah (2016), including with relevance to Alzheimer's disease. This article will include discussion of two types of databases: linguistic databases and recovery databases. The linguistic databases predominantly focus on spoken language production at the word, sentence, and discourse levels of people with aphasia. The recovery databases focus on how aphasia changes over time, which may or may not include active treatment. Finally, this article will close with a brief commentary on the benefits and limitations of current aphasia databases.

Linguistic Databases

AphasiaBank

One of the most commonly known databases for aphasia research is AphasiaBank (MacWhinney et al., 2011; MacWhinney & Fromm, 2016), which began as a way to gather discourse samples of people with aphasia and healthy control participants. Today, AphasiaBank is a rich research and clinical resource (Forbes et al., 2012), with a detailed protocol for several language tasks (e.g., multiple discourse prompts, picture naming, repetition) and a variety of languages, including English, Croatian, French, Italian, Mandarin, Romanian, and Spanish. There is even a Cantonese AphasiaBank that follows a similar protocol as AphasiaBank, but with careful consideration of cultural and linguistic differences (Kong & Law, 2018). Beyond discourse analysis, the rich data within AphasiaBank also allows for many other applications of its use. Two examples are examining video recordings of discourse production to understand gesture use by people with aphasia (de Beer et al., 2017; Stark & Oeding, 2023) and to understand how clinicians make treatment decisions (Hinckley & Sanchez, 2023). Please refer to section 21001 for more details on AphasiaBank.

Moss Aphasia Psycholinguistics Project Database (MAPPD)

MAPPD contains data from people with aphasia and healthy controls on the Philadelphia Naming Test (PNT), which is a 175-item confrontation naming test (Mirman et al., 2010). There is also additional data on repetition of PNT items, as well as standardized aphasia batteries (i.e., Western Aphasia Battery; Kertesz, 2007, or Boston Aphasia Diagnostic Examination; Goodglass et al., 2001) and tests of spoken word recognition, semantics, verbal short-term memory, and sentence comprehension. All the PNT data are available at the individual, trial-level allowing for analysis of item specific variables, like lexical frequency, imageability, and neighborhood density.

The MAPPD database has been widely used. To explore some of the research stemming from the MAPPD database, a brief literature review was conducted. First, journal articles and conference proceedings citing the original MAPPD paper (Mirman et al., 2010) in PubMed Central and Google Scholar were collated together on January 2, 2025. **Fig. 1** provides the summary results of the 76 publications identified. A co-authorship network is presented in **Fig. 2** demonstrating the frequency of publications by author (i.e., author name size) and the frequency of co-authorship on publications (i.e., links between author names, with thickness of link representing frequency weighting). As seen in **Fig. 2**, there are several co-author groups that published a single paper (i.e., all author names and links are of a similar small size), with a few clusters of co-author groups tied together by one or more frequent authors (e.g., the Mirman cluster).

Top 10 Publication Venues

Publication	#
Journal of Speech, Language, and Hearing Research	10
Cognitive Neuropsychology	7
Aphasiology	5
Neuropsychologia	4
American Journal of Speech-Language Pathology	3
Cortex	3
Frontiers in Psychology	3
Journal of Experimental Psychology: Learning, Memory, and Cognition	3
Psychonomic Bulletin & Review	3
Seminars in Speech and Language	3

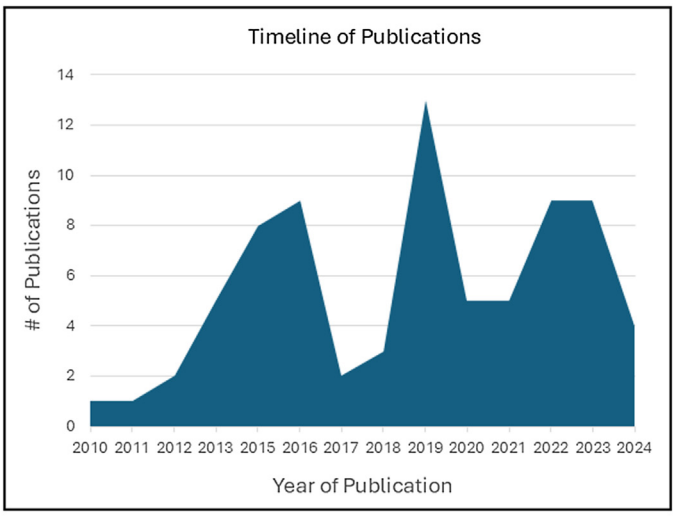

Fig. 1 Infographic of publications using the Moss Aphasia Psycholinguistic Project Database. The data presented was generated based on publications citing Mirman et al. (2010), the Moss Aphasia Psycholinguistics Project Database. The top left box shows the total number of publications identified (n = 76) using PubMed Central and Google Scholar search engines. The bottom left box shows the top 10 most frequent publication venues, with the most frequent venue being the *Journal of Speech, Language, and Hearing Research*. Finally, the right box shows a timeline of publications with the number of publications (y-axis) by each publication year (x-axis).

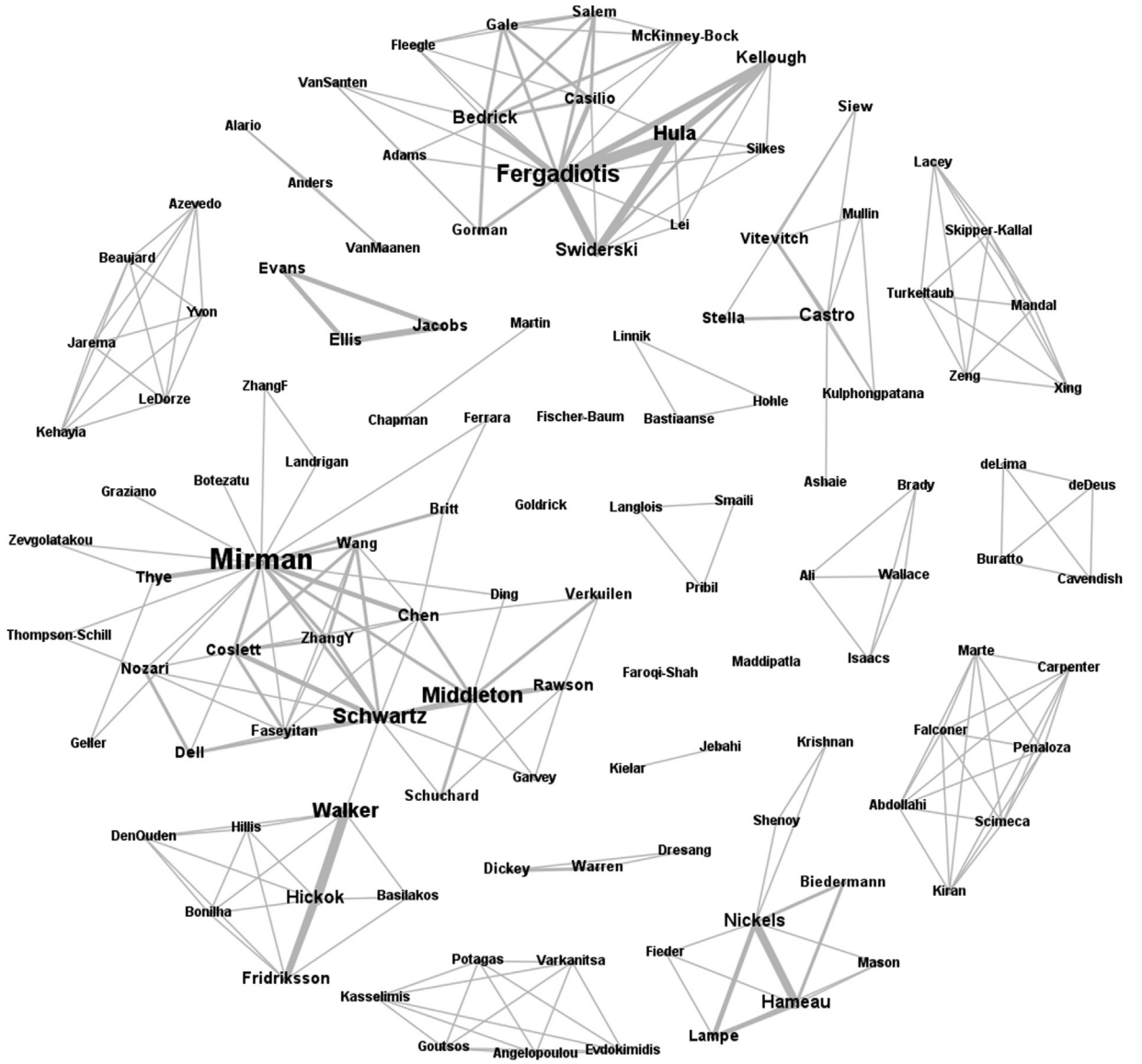

Fig. 2 Co-author network of publications using the Moss Aphasia Psycholinguistics Project Database. The co-author network was created based on publications citing Mirman et al. (2010), the Moss Aphasia Psycholinguistics Project Database. Nodes in the co-author network are represented as author last names. The size of the nodes (author names) represents the number of publications citing Mirman et al. (2010) by that author. Nodes are connected to each other with a gray link if the two authors were co-authors on the same paper. The weight of the links, represented by thickness, reflects the number of papers the co-authors were on together. For example, Mirman has the most publications citing Mirman et al. (2010), with a large co-author network. Note there are several clusters of co-authors, reflecting a single paper citing Mirman et al. (2010) by that author group, noted by equally small nodes and thin links, as well as a few single author papers (i.e., no links).

Several major themes of research were identified in reviewing the titles and abstracts of publications citing the original MAPPD paper. The most common theme was developing and testing models of word retrieval, including through lesion-symptom mapping (Chen et al., 2019; Dell et al., 2013; Landrigan et al., 2021), lexical networks (Castro et al., 2020; Mirman, 2010; Vitevitch et al., 2023), and cognitive psychometrics (Walker et al., 2018). Another common theme was developing short-form naming tests (Walker & Schwartz, 2012), including computer-adaptive (Hula et al., 2015, 2020) and severity calibrated tests (Walker et al., 2022). There were also papers examining each of the following themes: (1) cognitive-linguistic skills in aphasia (Ashaie & Castro, 2021; Jebahi & Kielar, 2024), (2) naming error classification algorithms (Casilio et al., 2023), and (3) the influence of social determinants of health on aphasia (Jacobs et al., 2023).

Language Experience in Bilinguals With and Without Aphasia Dataset (LEX-BADAT)

LEX-BADAT contains data from people with aphasia and healthy controls who are bilingual speakers of Spanish and English (Marte et al., 2022). Participants completed a language use questionnaire for both languages, with the bilinguals with aphasia reporting on

their language experience prior to and after aphasia onset (Peñaloza et al., 2020). The language use questionnaire includes measures of age of acquisition, language ability rating, daily use, family proficiency, educational history, lifetime exposure and lifetime confidence. There is also data on participants with aphasia about their lesion and aphasia severity and sub-type based on the Western Aphasia Battery-Revised (Kertesz, 2007).

Recovery Databases

Rehabilitation and Recovery of People With Aphasia After Stroke (RELEASE)

The Collaboration of Aphasia Trialists (CATs) is an international network of aphasiology researchers and clinicians with several working groups advancing research and practice for aphasia. One of the working groups is dedicated to the compilation and management of completed aphasia treatment research trials, called RELEASE (Williams et al., 2022). To be included, datasets shared with the working group must contain a minimum set of information: individual patient data, aphasia etiology, time post-onset of aphasia, language impairment at baseline, and repeated measurement of at least one language assessment using a validated tool at follow-up. All the data shared with the working group is anonymized and can be released to interested researchers with appropriate approvals for novel analyses. The CATs working group has published papers based on these data to examine predictors of treatment (The RELEASE Collaborators, 2021, 2022).

Aphasia Recovery Cohort

The Aphasia Recovery Cohort contains data from people with chronic aphasia, including neuroimaging, collected from multiple treatment studies (Gibson et al., 2024). Aphasia diagnostic information using the Western Aphasia Battery (Kertesz, 2007) is available on these individuals. There are some individuals within the dataset who have been observed at multiple time points allowing for further modeling of recovery trajectories.

Aphasia Recovery Project

The Aphasia Recovery Project contains data from people with aphasia, starting in acute care (e.g., as early as 5 days post-stroke) with follow-up evaluations at one-month, three-months, and 1-year timepoints (Wilson et al., 2023). Data includes an aphasia evaluation using the Quick Aphasia Battery (Wilson et al., 2018) and neuroimaging. The data is currently available in the supplementary material of Wilson et al. (2023).

Conclusion

There has been an increasing recognition of the importance of accessible databases to study aphasia and the recovery of aphasia (Faroqi-Shah, 2016). There are many benefits to having large datasets available, including the ability to study aphasia with larger sample sizes than what is typically obtained in any individual lab. This is particularly useful for testing novel ideas, particularly generating preliminary data for grant funding. There are limitations though (Faroqi-Shah, 2016), including having adequate resources to develop and maintain databases, ensuring protection of human subjects (Girolamo et al., 2023), keeping to a consistent protocol for data collection (Wallace et al., 2023), and developing thorough documentation of variables. Users are limited by what data was collected, as the use of these databases are "retrospective" in design, which will naturally limit the kinds of questions that can be asked with the data available. For example, many of the databases do not include functional communication or quality of life measures. The existing databases are also limited in the participants included, focusing predominantly on monolingual English speakers with acquired aphasia, neglecting important language diversity and progressive aphasia. Despite these limitations, these aphasia databases provided a valuable resource for continued efforts to understand the nature of language impairment in aphasia and serve as a model for continued efforts to develop and collate new databases.

References

Ashaie, S., & Castro, N. (2021). Exploring the complexity of aphasia with network analysis. *Journal of Speech, Language, and Hearing Research, 64*(10), 3928–3941. https://doi.org/10.1044/2021_JSLHR-21-00157

Casilio, M., Fergadiotis, G., Salem, A. C., Gale, R. C., McKinney-Bock, K., & Bedrick, S. (2023). ParAlg: A paraphasia algorithm for multinomial classification of picture naming errors. *Journal of Speech, Language and Hearing Research, 66*(3), 966–986. https://doi.org/10.1044/2022_JSLHR-22-00255

Castro, N., Stella, M., & Siew, C. S. Q. (2020). Quantifying the interplay of semantics and phonology during failures of word retrieval by people with aphasia using a multiplex lexical network. *Cognitive Science, 44*(9). e12881. https://doi.org/10.1111/cogs.12881.

Chen, Q., Middleton, E., & Mirman, D. (2019). Words fail: Lesion-symptom mapping of errors of omission in post-stroke aphasia. *Journal of Neuropsychology, 13*(2). https://doi.org/10.1111/jnp.12148

de Beer, C., Carragher, M., van Nispen, K., Hogrefe, K., de Ruiter, J. P., & Rose, M. L. (2017). How much information do people with aphasia convey via gesture? *American Journal of Speech - Language Pathology, 26*(2), 483–497. https://doi.org/10.1044/2016_AJSLP-15-002

Dell, G. S., Schwartz, M. F., Nozaro, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition, 128*(3), 380–396. https://doi.org/10.1016/j.cognition.2013.05.007

Faroqi-Shah, Y. (2016). The rise of big data in neurorehabilitation. *Seminars in Speech and Language, 37*(1), 3–9. https://doi.org/10.1055/s-0036-1572385

Forbes, M. M., Fromm, D., & MacWhinney, B. (2012). AphasiaBank: A resource for clinicians. *Seminars in Speech and Language, 33*(3), 217—222. https://doi.org/10.1055/s-0032-1320041

Gibson, M., Newman-Norlund, R., Bonilha, L., Fridrikkson, J., Hickok, G., Hillis, A. E., den Ouden, D. B., & Rorden, C. (2024). The Aphasia Recovery Cohort, an open-source chronic stroke repository. *Nature, 11*, 981. https://doi.org/10.1038/s41597-024-03819-7

Girolamo, T., Castro, N., Hendricks, A. E., Ghali, S., & Eigsti, I. M. (2023). Implementation of open science practices in communication sciences and disorders research with Black, Indigenous, and people of color. *Journal of Speech, Language and Hearing Research, 66*(6), 2010—2017. https://doi.org/10.1044/2022_JSLHR-22-00272

Goodglass, H., Kaplan, E., & Weintraub, S. (2001). *BDAE: The Boston Diagnostic aphasia examination*. Lippinott Williams & Wilkins.

Hinckley, J., & Sanchez, L. (2023). Treatment time and treatment selection in aphasia: A preliminary study using vignettes. *American Journal of Speech - Language Pathology, 32*(5S), 2430—2443. https://doi.org/10.1044/2023_AJSLP-22-00294

Hula, W. D., Fergadiotis, G., Swiderski, A. M., Silkes, J. P., & Kellough, S. (2020). Empirical evaluation of computer-adaptive alternate short forms for the assessment of anomia severity. *Journal of Speech, Language and Hearing Research, 63*(1), 163—172. https://doi.org/10.1044/2019_JSLHR-L-19-0213

Hula, W. D., Kellough, S., & Fergadiotis, G. (2015). Development and simulation testing of a comptuerized adaptive version of the Philadelphia Naming Test. *Journal of Speech, Language and Hearing Research, 58*(3), 878—890. https://doi.org/10.1044/2015_JSLHR-L-14-0297

Jacobs, M., Evans, E., & Ellis, C. (2023). Intersectional sociodemographic and neurological relationships in the naming ability of persons with post-stroke aphasia. *Journal of Communication Disorders, 105*, 106352. https://doi.org/10.1016/j.jcomdis.2023.106352.

Jebahi, F., & Kielar, A. (2024). The relationship between semantics, phonology, and naming performance in aphasia: A structural equation modeling approach. *Cognitive Neuropsychology, 41*(3—4), 113—128. https://doi.org/10.1080/02643294.2024.2373842

Kertesz, A. (2007). *Western aphasia Battery revised*. Pearson.

Kong, A. P. H., & Law, S. P. (2018). Cantonese AphasiaBank: An annotated database of spoken discourse and co-verbal gestures by healthy and language-impaired native Cantonese speakers. *Behavior Research Methods, 51*, 1131—1144. https://doi.org/10.3758/s13428-018-1043-6

Landrigan, J., Zhang, F., & Mirman, D. (2021). A data-driven approach to post-stroke aphasia classification and lesion-based prediction. *Brain, 144*(5), 1372—1383. https://doi.org/10.1093/brain/awab010

MacWhinney, B., & Fromm, D. (2016). AphasiaBank as BigData. *Seminars in Speech and Language, 37*(1), 10—22. https://doi.org/10.1055/s-0036-1571357

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*(11), 1286—1307. https://doi.org/10.1080/02687038.2011.589893

Marte, M. J., Carpenter, E., Falconer, I. B., Scimeca, M., Abdollahi, F., Peñaloza, C., & Kiran, S. (2022). LEX-BADAT: Language EXperience in bilingual with and without aphasia DATaset. *Frontiers in Psychoology, 13*, 875928. https://doi.org/10.3389/fpsyg.2022.875928.

Mirman, D. (2010). Effects of near and distant semantic neighbors on word production. *Cognitive, Affective, & Behavioral Neuroscience, 11*, 32—43. https://doi.org/10.3758/s13415-010-0009-7

Mirman, D., Strauss, T., Brecher, A., Walker, G., Sobel, P., Dell, G., & Schwartz, M. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology, 27*(6), 495—504. https://doi.org/10.1080/02643294.2011.574112

Peñaloza, C., Barrett, K., & Kiran, S. (2020). The influence of prestroke proficiency on poststroke lexical-semantic performance in bilingual aphasia. *Aphasiology, 34*(10), 1223—1240. https://doi.org/10.1080/02687038.2019.1666082

Stark, B. C., & Oeding, G. (2023). Demographic, neuropsychological, and speech variables that impact iconic and supplementary-to-speech gesturing in aphasia. *Gesture, 22*(1), 62—93. https://doi.org/10.1075/gest.23019.sta

The RELEASE Collaborators. (2021). Predictors of poststroke aphasia recovery: A systematic review-informed individual participant data meta-analysis. *Stroke, 52*(5), 1778—1787. https://doi.org/10.1161/STROKEAHA.120.031162

The RELEASE Collaborators. (2022). Dosage, intensity, and frequency of language therapy for aphasia: A systematic review-based, individual participant data network meta-analysis. *Stroke, 53*(3), 956—967. https://doi.org/10.1161/STROKEAHA.121.035216

Vitevitch, M. S., Castro, N., Mullin, G. J. D., & Kulphongpatana, Z. (2023). The resilience of the phonological network may have implications for developmental and acquired disorders. *Brain Sciences, 13*(2), 188. https://doi.org/10.3390/brainsci13020188

Walker, G. M., Fridriksson, J., Hillis, A. E., den Ouden, D. B., Bonilha, L., & Hickok, G. (2022). The severity-calibrated aphasia naming test. *American Journal of Speech - Language Pathology, 31*(6), 2722—2740. https://doi.org/10.1044/2022_AJSLP-22-00071

Walker, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological Assessment, 30*(6), 809—826. https://psycnet.apa.org/record/2018-11628-001.

Walker, G. M., & Schwartz, M. F. (2012). Short form Philadelphia naming test: Rationale and empirical evaluation. *American Journal of Speech-Language Pathology, 21*(2), S140—S153. https://doi.org/10.1044/1058-0360(2012/11-0089

Wallace, S. J., Isaacs, M., Ali, M., & Brady, M. C. (2023). Establishing reporting standards for participant characteristics in post-stroke aphasia research: An international e-Delphi exercise and consensus meeting. *Clinical Rehabilitation, 37*(2), 199—214. https://doi.org/10.1177/02692155221131241

Williams, L. R., Ali, M., VandenBerg, K., Williams, L. J., Abo, M., Becker, F., Bowen, A., Brandenburg, C., Breitenstein, C., … Wright, H. H., Brady, M. C., & The RELEASE Collaborators. (2022). Utilising a systematic review-based approach to create a database of individual participant data for meta- and network meta-analyses: The RELEASE database of aphasia after stroke. *Aphasiology, 36*(4), 513—533. https://doi.org/10.1080/02687038.2021.1897081

Wilson, S. M., Entrup, J. L., Schneck, S. M., Onuscheck, C. F., Levy, D. F., Rahman, M., Willey, E., Casilio, M., Yen, M., Brito, A. C., Kam, W., Davis, T., de Riesthal, M., & Krishner, H. S. (2023). Recovery from aphasia in the first year after stroke. *Brain, 146*(3), 1021—1039. https://doi.org/10.1093/brain/awac129

Wilson, S. M., Eriksson, D. K., Schneck, S., & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS ONE, 13*(6). e0199469. https://doi.org/10.1371/journal.pone.0192773.

Relevant Websites

Linguistic Databases

AphasiaBank https://aphasia.talkbank.org/.
MAPPD https://www.mappd.org/about.html.
LEX-BADAT https://osf.io/jaxsg/?view_only=11d61617d5cf49d39ec9aa446f21c4f7.

Recovery Databases

RELEASE https://www.aphasiatrials.org/aphasia-dataset/.
Aphasia Recovery Cohort https://openneuro.org/datasets/ds004884/versions/1.0.1.
Aphasia Recovery Project https://langneurosci.org/recovery/.