Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 181–186
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Integrating Auslan Resources into the Language Data Commons of Australia

River Tae Smith, Louisa Willoughby, Trevor Johnston
Monash University
Melbourne, Australia
{river.smith, louisa.willoughby, trevor.johnston}@monash.edu

Abstract

This paper describes a project to secure Auslan (Australian Sign Language) resources within a national language data network called the Language Data Commons of Australia (LDaCA). The resources are Auslan Signbank, a web-based multi-media dictionary, and the Auslan Corpus, a collection of video recordings of the language being used in various contexts with time-aligned ELAN annotation files. We aim to make these resources accessible to the language community, encourage community participation in the curation of the data, and facilitate and extend their uses in language teaching and linguistic research. The software platforms of both resources will be made compatible with other LDaCA resources; and the two will also be aggregated and linked so that (i) users of the dictionary can view attested corpus examples for an entry; and (ii) users of the corpus can instantly view the dictionary entry for an already glossed sign to check phonological, lexical and grammatical information about it, and/or to ensure that the correct annotation gloss (aka 'ID-gloss') for a sign token has been chosen. This will enhance additions to annotations in the Auslan Corpus, entries in Auslan Signbank and the integrity of research based on both.

Keywords: Auslan, signed language dictionaries, signed language corpora, data repositories, language documentation, linking corpora and lexicons

1. Introduction

This paper describes a project to secure Auslan resources within a national digital network called the Language Data Commons of Australia (LDaCA). Auslan is the signed language of the deaf community in Australia. The aim is to make these resources readily accessible to the language community, encourage community participation in the on-going curation of them, and facilitate and extend their uses in Auslan language teaching and linguistic research.

We begin by describing the purpose and principles behind the LDaCA initiative. This is followed by an overview of the two Auslan resources: Auslan Signbank[1] and the Auslan Corpus[2]. We then continue with more specific information about securing, aggregating and linking these two resources, and the desired outcomes of the project.

We then describe how we intend to achieve these goals in terms of data management (including community participation) and software development. We summarise previous work to develop software for related signed language data storage and research. These are compared with software development which is already underway or is planned as part of the restructure of the Auslan language resources so that they are seamlessly integrated into the LDaCA network.

2. The Language Data Commons of Australia (LDaCA)[3]

Australia has amassed significant collections of language data concerning Australian Indigenous languages, the signed language of the **deaf** community, languages of the Pacific region, and of Australian English. Many of these collections remain under-utilised or at risk and are difficult to access for researchers and communities. These collections need to be hosted on durable infrastructure that ensures that those collections are given perennity and security. LDaCA will address this by working with key groups to capitalise on existing infrastructure to secure vulnerable and dispersed language collections of written, spoken, multimodal and **signed** text. Moreover, LDaCA will link these collections with improved analysis environments for new research outcomes.

LDaCA aims to establish a sustainable long-term repository for ingesting and curating language data collections of national significance: to democratise access to Australia's rich linguistic heritage through enabling those collections to become more FAIR while following the CARE principles[4]; and to demonstrate how to balance research needs with preserving community rights. It also aims to develop the computational capa-

[1] https://auslan.org.au/

[2] https://www.elararchive.org/dk0001/

[3] The Language Data Commons of Australia (LDaCA) is a sub-part of the Humanities and Social Sciences Research Data Commons (HASS RDC) which is itself part of the large Australian Research Data Commons (ARDC) initiative

[4] Further information about the CARE Principles for Indigenous Data Governance can be found on the GIDA website (https://www.gida-global.org/care)

bilities, technical infrastructure and support services to analyse language collections at scale.

The overall LDaCA project will build connections to other research data collections in the humanities and social sciences in Australia by developing APIs and text analytics tools that can be applied to any of these collections; by facilitating text analysis of aggregated administrative data collections; and by developing a community-driven approach for governance of Indigenous language collections as well as the signed language of the Australian deaf community (Auslan).

3. The Auslan Resources

The two major Auslan language resources that exist are Auslan Signbank (Johnston, 2004) and the Auslan Corpus (Johnston, 2008). They are perfect examples of the type of language resources which are the focus of the LDaCA initiative.

3.1. Auslan Signbank

Auslan Signbank is an online dictionary of Auslan. Entries consist of a form-based sequence of Auslan headsigns as videos with accompanying definitions in Auslan and English. It is thus a true dictionary of a sign language not simply an English word list with each word equated with a sign.

A lexical database of Auslan was begun in 1984 by one of the authors of this paper (Johnston, 2001) and it slowly migrated over various programs and platforms. In 2004, the first online iteration of the dictionary was released, Auslan Signbank.

Currently there are c. 7,500 sign entries in Auslan Signbank of which c. 5,000 are viewable by all visitors to the site. The remaining c. 2,500 are in development and can only be viewed by registered users with special access. Auslan Signbank is thus not a static repository of lexical information. The web-based format was designed to enable on-going change to the content by the editor/lexicographer without any need for discrete editions. For example, sign entries can be deleted, corrected, or created (including showing or hiding entries in the public portal). This includes phonological, semantic, and grammatical information about each sign. These changes are informed by on-going linguistic research, input from Auslan teachers or interpreters, and deaf community feedback.

Since its creation Auslan Signbank has had no official long-term institutional home and hosting site. Thus, its future was insecure until an appropriate repository could be established.

3.2. Auslan Corpus

The Auslan Corpus is a collection of digital video recordings of deaf users of Auslan using the language in various contexts. Many of the recordings have time-aligned annotation files; thus, the Auslan Corpus is not just a collection of video recordings.

The Auslan Corpus was one of the first digital video archives of a signed language that was collected with the express purpose of creating a machine-readable linguistic corpus in the modern sense. It was collected as part of a three-year Endangered Languages Documentation Program project (2004-2006) which resulted in the deposit of the Auslan Corpus in the Endangered Languages Archive (ELAR) in 2008. One hundred participants from Australia's five largest cities were filmed. This yielded approximately 150 hours of edited clips. At the time of deposit <50 clips had been annotated for glosses and literal and free translations.

Over the fourteen years since its initial deposit, hundreds of thousands of annotations have been added to the corpus during a number of research projects and are not part of the original or current ELAR resource (Examples include Ferrara & Johnston 2014; Gray, 2013; Hodge & Ferrara 2014; Hodge & Johnston 2014; Johnston 2012, 2013, 2018, 2019; Johnston et al. 2015, 2016.). Today about 200 clips have been annotated in detail, with roughly 200 more annotated sporadically. In the current working corpus, there are approximately 105,000 sign token gloss annotations, but 385,000 linguistic annotations in total. These annotations include part-of-speech tagging, morphemic tagging, clause tagging and grammatical role tagging.

4. Auslan/LDaCA project

The Auslan/LDaCA project will significantly increase research opportunities in Auslan language and linguistics using these two aggregated and linked resources. The project has been briefly summarised above. It has three phases which focus on (i) **securing language data collections**, (ii) **aggregating and linking collections**, and (iii) **enhancing their research potential and facilitating new research, especially text data analysis environments**. The LDaCA project has several streams and hubs based on the language resources and expertise associated with each of its partner organisations and associated language repositories[5]. At the Monash University hub of LDaCA, the focus is on Auslan and other signed languages of Australia and its region, and on co-speech gesture and multimodal language research.

4.1. Phase 1: Securing Language Data Collections

The first phase is particularly relevant to Auslan. Until this project, Auslan Signbank did not have any long-term institutional home or hosting site and its future was not secure. Its ownership and hosting had migrated with the chief investigator (Trevor Johnston) over many

[5] The project partners are: The University of Queensland, Australian National University, Monash University, The University of Melbourne, The University of Sydney, AARNet, First Languages Australia (FLA), Australian Institute for Aboriginal and Torres Strait Islander Studies (AIATSIS), PARADISEC,ARC Centre of Excellence for the Dynamics of Language (CoEDL), Digital Observatory (Queensland University of Technology), CLARIN.

separately funded research projects across multiple institutions. Without a long-term institutional home for Signbank, it is difficult not only to simply secure the data but also to support editing, and improvements and additions to enhance its usability and accessibility for learners and teachers of Auslan, Auslan/English interpreters, deaf community members, and language researchers; as well as to prepare its underlying data structure for linking with the Auslan Corpus. Tweaking of the software and the design of the dictionary webpages will be made to include and then distinguish each type of sign form that are possible, such as (i) conventional lexical signs of Auslan, (ii) gestures used by both deaf Auslan-users and hearing Australian English speakers; (iii) conventional signs of the sign languages of Indigenous Australia; and (iv) conventional signs found in any deaf community sign language anywhere around the world.

With respect to Phase 1 the Auslan Corpus as it currently exists (with enriched annotation files that have been created since 2008) also had no institutional home. The reason is that these newer annotation files are not part of the ELAR deposit from 2008. Partly because there has been no easy way to manage version control as various research projects add or change annotations to existing files, these enhanced or project-specific annotations files have instead been saved and backed up privately by Trevor Johnston and other Auslan researchers.

The first phase of the Auslan/LDaCA project started in early 2022 when the long term hosting of both resources at Monash University was secured. Work has begun on re-configuring the resources to conform to the LDaCA data protocols.

4.2. Phase 2: Aggregating and Linking Language Collections

The second phase is again of particular relevance with regard to the Auslan resources. Since its creation the Auslan Corpus has been completely separate from Auslan Signbank, and vice versa. The two need to be linked both for the benefit of the language community itself, and to maximise the utility of the data contained within each.

With respect to the language community, deaf and hearing teachers of Auslan, deaf students and their teachers in schools, and hearing learners of Auslan (school children, adults, and parents of deaf children), and trainee Auslan interpreters, have all on multiple occasions asked to access the Auslan Corpus to simply get more exposure to the language or to have more examples of particular constructions to use in teaching. Auslan Signbank also needs to be linked to the Auslan Corpus to unlock the full potential of the latter as a resource for the teaching and learning of Auslan. Linking will make it possible for teachers and learners to jump from an entry in Auslan Signbank to attested examples from the Auslan Corpus. This is particularly useful for students to appreciate the ways the actual production of a sign in context can vary from its citation form due to (i) phonological processes found in continuous signing (which are not unlike those found in continuous speech), and (ii) morphological processes that change the shape of signs in systematic ways to express various meanings.

We envisage that Auslan textbooks and classroom resources will link to Signbank and the Auslan Corpus to provide vocabulary lists and resources for students, and that teachers and students will access the dictionary and the corpus both for explicit teaching in class and private study.

With respect to language researchers, linking will have practical advantages. A unique identifying gloss (known as the 'ID-gloss') is used for each sign form entry in Signbank. It was quickly realised when the first annotations of the corpus in ELAN were being made that the Signbank site could be used by annotators to view ID-glosses and thus ensure consistency in corpus glossing. Linking the two resources by exploiting the common data point (the ID-gloss) so that (i) users of the dictionary can view attested corpus examples for an entry; and (ii) users of the corpus can instantly view the dictionary entry for an already glossed sign to check phonological, lexical and grammatical information about it, and/or to ensure the correct ID-gloss for a sign token is chosen. This will enhance additions to the annotations in the Auslan Corpus, entries in Auslan Signbank and the integrity of research based on both.

4.3. Phase 3: Enhancing and Facilitating the Research Potential of Language Resources

The aggregation of the data in Auslan Signbank and the Auslan Corpus will improve the research potential of both datasets, as well as enable new and accelerated research into unexplored aspects of the lexicon and grammar of the language. In this phase we will develop specialist tools for text analytics and extend the text analytics workbench to enable large-scale computational analysis of written, spoken, multimodal and signed language data, and to share those workflows with other researchers. Thus, once the annotation environment for the Auslan Corpus has been streamlined, researchers should be able to take advantage of LDaCA text analytics tools to enhance their research.

These developments will also serve to enhance the accuracy of phonology description the dictionary and corpus, and facilitate the management of other multimedia resources, such as co-speech gesture data, Australian Indigenous signed language data, and, ultimately, data from potentially any signed language. It will also enhance the research potential of these resources.

4.4. Software Development

4.4.1. Previous Work

Over its lifetime, the user interface and functionality of Auslan Signbank has been consistently improved and updated. For example, in 2008 a major overhaul of the Signbank site created new data fields to accommodate import of data from Prof. Johnston's existing FileMaker lexicon. These new data fields were visible to logged-in researchers and lexicographers who could then edit them. In 2014, the ability to add video definitions in Auslan was added, meaning that headsigns could be given definitions in both English and Auslan. In other words, after ten years the web-based dictionary has started to evolve into the first true monolingual signed language dictionary—Auslan signs can now be defined and explained in Auslan itself.

Previous attempts have been made to enable online access and annotation of ELAN annotation files (EAFs). One such attempt using the Auslan Corpus was Cassidy and Johnston (2009), which described the approach of converting EAF files into an RDF-based format which could more easily be communicated over HTTP. The new LDaCA corpus storage solution described, in part, in this paper would make this kind of intermediate format unnecessary. While online collaborative corpus editing is not planned for the Auslan/LDaCA project, it is an avenue ripe for later projects in this area.

4.4.2. This Project

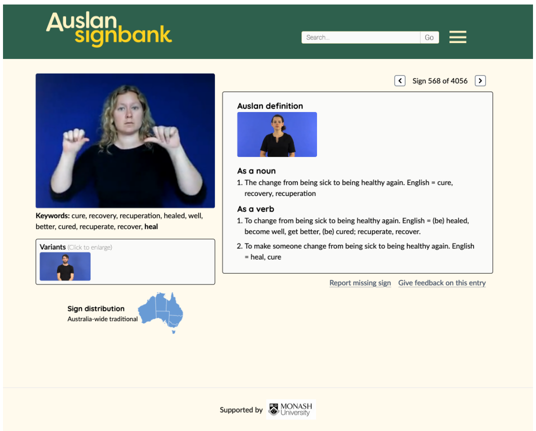
Nationally significant collections of sign language data of Australia and its region will be secured as preservable digital objects using Arkisto Platforms standards, a combination of the Oxford Common File Layout (OCFL) and RO-Crate, and access protocols for Australian researchers and communities will be developed. This will involve using Signbank – the online dictionary of Auslan – and the Auslan Corpus, and working with sign language and gesture researchers, to deposit annotated multimodal video data and dictionary resources. Deliverables include migration of selected sign language data collections into RO-Crate/OCFL formats, the development of access protocols and resources for Australian researchers and communities, including Auslan teachers and deaf Australians. This work package is led by Monash University.

4.4.3. Current State of Auslan Signbank

The existing Signbank, both Auslan and subsequent forks for other languages, is written in Python, using the Django web framework. Django is an easy-to-use web framework, which includes a built-in object-relation mapping library (ORM). While Django's ORM has served the needs of Auslan Signbank till now, it necessitates high coupling between the dictionary and website content. The ORM stores all data in a single PostgreSQL database, the architecture of which is partially described in Cassidy et al. (2018). Prior to March 2022, Auslan Signbank was hosted by A/Prof. Cassidy, who developed the original Signbank, at Macquarie University. In February 2022, it was moved to temporary hosting, funded by Monash/LDaCA. This temporary hosting is planned to end in August 2022, when a new version of Auslan Signbank will become available to the public. The final hosting solution has not been confirmed, but it is expected to use Monash-managed AWS cloud resources. The auslan.org.au URL will remain in use, and existing links will be redirected where possible.

4.4.4. Signbank Next

The Auslan/LDaCA project has begun a ground-up redesign of Signbank, Signbank Next, which will incorporate modern web development best-practices, including WCAG accessible design standards[6] while delivering on existing functionality described in Cassidy et al. (2018). Signbank Next will be built with Javascript and NextJS, an open-source web development framework.


Auslan signbank
Search...
Go
Sign 568 of 4056
Auslan definition
Variants (Click to enlarge)
Report missing sign
Give feedback on this entry
Sign distribution
Australia-wide traditional
Supported by MONASH University

Keywords: cure, recovery, recuperation, healed, well, better, cured, recuperate, recover, **heal**

As a noun

1. The change from being sick to being healthy again. English = cure, recovery, recuperation

As a verb

1. To change from being sick to being healthy again. English = (be) healed, become well, get better, (be) cured; recuperate, recover.
2. To make someone change from being sick to being healthy again. English = heal, cure

Figure 1: Signbank Next dictionary entry.

In contrast to the Django version of Signbank, and with some similarity to the FileMaker one, Signbank Next will use document storage (i.e., MongoDB). Document storage is well suited to dictionaries, as the amount of relational information is minimal in such applications. Each headsign, usually a video or picture, will be stored with all related information, eliminating the need for complex joins to construct a dictionary page. Document storage has high efficiencies of scale and will easily maintain speed as Signbank grows.

After the architectural redesign, improvements will be made to existing Signbank functionality. Headsigns have historically been sorted according to a manually curated 'sign number', which was used to order signs by their phonology. With each sign added to Signbank, it becomes increasing likely that sign numbers must be reassigned in bulk to make space for new entries. Sorting based on existing phonology fields will be implemented, and the 'sign number' field removed. This has the added benefit of opening the door to custom sorting

[6] https://www.accessibility.org.au/guides/what-is-the-wcag-standard/

orders, which would be invaluable as non-Auslan signs are added to Signbank.

More advanced searching and filtering will also be implemented. Currently, a manual tagging system is used to track problematic entries and missing or low-quality videos. This system is error prone and incorrectly tagged entries can be easily missed. Text search capabilities are similarly lacking; regular expressions are not supported, and queries only attempt to match from the start of a translation. These issues will be addressed with a combination of more comprehensive filters and the addition of an advanced query syntax.

4.4.5. Unified LDaCA Search Portal

The LDaCA corpora will share a corpus search portal, Oni. Information about the Auslan Corpus will be provided on the Auslan Signbank website, but users will directed to access it through Oni. The Oni search portal will provide a user interface for browsing and filtering its corpora and will enable text search across multiple corpora simultaneously, limited by the text available in the search index. This project will include scripts to add the corpus' English free translation tiers, where they exist, into the Oni search index.

To conform with other LDaCA resources, the Auslan Corpus will be transformed to fit the Arkisto Platform standards. These standards focus on sustainable and scalable data management. This will also allow the Auslan Corpus to leverage Arkisto-compatible tools, including data description software and web portals to accept new data from the public.

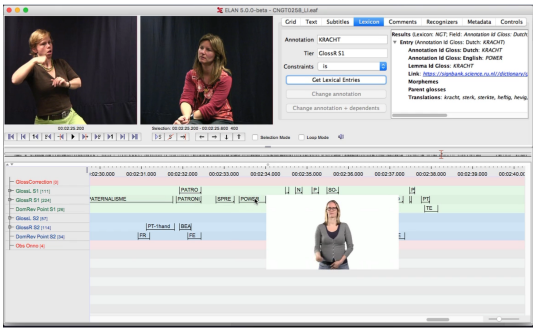
4.4.6. Dictionary Lookup Within ELAN

Global Signbank (Crasborn et al., 2018) has created two methods to aid the annotation of sign language against an instance of Signbank. Global Signbank can generate an ELAN controlled vocabulary from entries, requiring manual updates from the user as the Signbank is updated. The more recent innovation in this area is ELAN's 'lexicon service'[7], which allows ELAN to log into Signbank with a user's login details and download a local cache of compressed headsign videos. Once this cache has been created, one can search Signbank from within ELAN. This kind of interoperability is a great quality-of-life improvement for corpus annotators, who can quickly confirm glossing accuracy and improve data consistency, and researchers, who can quickly compare citation forms with real usage in their data.

Lexicon service compatibility will be added after the first release of Signbank Next, with possibility of augmenting the service with live API calls in the future, without needing to download all entries ahead-of-time.

4.4.7. Corpus Lookup From Signbank

Going the other direction, finding attested examples of Signbank entries in the corpus, will be achieved by constructing a search index over the corpus' ELAN glossing tiers. This index will map from ID glosses in the annotation onto their timestamps. To ensure proper access control checks can be made, users will be required to link their Signbank and Oni accounts. This work will benefit any current, or future, corpora that include ELAN annotation files by allowing gloss searches from within the regular Oni search portal.



Annotation: KRACHT
Tier: GlossR S1
Constraints: is

Selection: 00:02:25.200 - 00:02:25.600 400

Results (Lexicon: NGT; Field: Annotation Id Gloss: Dutch;
Entry (Annotation Id Gloss: Dutch: KRACHT)
Annotation Id Gloss: Dutch: KRACHT
Annotation Id Gloss: English: POWER
Lemma Id Gloss: KRACHT
Link: https://signbank.science.ru.nl//dictionary/g
Morphemes
Parent glosses
Translations: kracht, sterk, sterkte, heftig, hevig,

Figure 2: The ELAN lexicon service in use.

5. Conclusion

A new version of Auslan Signbank will be released, hosted by Monash University. The Auslan Corpus will also be made available to both researchers and the general public via Oni, a repository for digital multimodal language data created in partnership between five Australian universities and seven institutions concerned with language preservation and research. It will aid the development of a portal and associated backend infrastructure for multilingual corpus search and access, which will enable sign language and gesture researchers to deposit annotated multimodal video data and dictionary resources, focusing on Australia and its region. It will use Signbank—the online dictionary of Auslan—and the Auslan Corpus as testbed collections. This project involves an interdisciplinary team including data scientists, linguists, and educators. Importantly, it also includes the establishment of a formal deaf community based mechanism to advise on the curation and augmentation of these Auslan resources into the future.

The online, interlinked Signbank and Auslan Corpus will be a boon for Auslan teachers and learners at all levels across Australia. These are extremely valuable resources for language learners to see how words/signs are used in context and to understand how grammatical features of sign languages—such as facial expression or the movement path of the sign—can modify the meaning of a particular sign. It also allows both teachers and learners to better understand subtle shades of meaning between two signs that may have similar meanings in English by seeing the different environments in which they are used.

[7] Available since version 5.0.0b, documented in the ELAN manual

6. Acknowledgements

This research is being supported by LDaCA and Monash University. The original Auslan Signbank was created, and maintained for over a decade, by A/Prof. Steve Cassidy.

7. Bibliographical References

Cassidy, S. and Johnston, T. (2009). Ingesting the Auslan corpus into the DADA annotation store. In *Proceedings of the Third Linguistic Annotation Workshop on - ACL-IJCNLP '09*, pages 154–157, Suntec, Singapore. Association for Computational Linguistics.

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., and Johnston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language.

Crasborn, O., Zwitserlood, I., Kooij, E. V. D., and Schüller, A. (2018). Global SignBank manual. Radboud University, Centre for Language Studies.

Ferrara, L. and Johnston, T. (2014). Elaborating Who's What: A Study of Constructed Action and Clause Structure in Auslan (Australian Sign Language). *Australian Journal of Linguistics*, 34(2):193–215, April.

Gray, M. (2013). Aspect marking in Australian Sign Language: a process of gestural verb modification. Macquarie University.

Hodge, G. and Ferrara, L. (2013). Showing the story: enactment as performance in Auslan narratives. University of Melbourne.

Hodge, G. and Johnston, T. (2014). Points, Depictions, Gestures and Enactment: Partly Lexical and Non-Lexical Signs as Core Elements of Single Clause-Like Units in Auslan (Australian Sign Language). *Australian Journal of Linguistics*, 34(2):262–291, April.

Johnston, T., Cresdee, D., Schembri, A., and Woll, B. (2015). FINISH variation and grammaticalization in a signed language: How far down this well-trodden pathway is Auslan (Australian Sign Language)? *Language Variation and Change*, 27(1):117–155, March.

Johnston, T., van Roekel, J., and Schembri, A. (2016). On the Conventionalization of Mouth Actions in Australian Sign Language. *Language and Speech*, 59(1):3–42, March.

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics*, 4(1-2):145–169, December.

Johnston, T. (2012). Lexical Frequency in Sign Languages. *Journal of Deaf Studies and Deaf Education*, 17(2):163–193, April.

Johnston, T. (2013). Formational and functional characteristics of pointing signs in a corpus of Auslan (Australian sign language): are the data sufficient to posit a grammatical class of 'pronouns' in Auslan? *Corpus Linguistics and Linguistic Theory*, 9(1):109–159, May. De Gruyter Mouton.

Johnston, T. (2018). A corpus-based study of the role of headshaking in negation in Auslan (Australian Sign Language): Implications for signed language typology. *Linguistic Typology*, 22(2):185–231, August.

Johnston, T. A. (2019). Clause constituents, arguments and the question of grammatical relations in Auslan (Australian Sign Language): A corpus-based study. *Studies in Language*, 43(4):941–996, December.

8. Language Resource References

Johnston, T. (2004). *Auslan Signbank*. Royal Institute for Deaf and Blind Children & Catalyst Training Systems.

Johnston, T. (2008). *Auslan Corpus*. London: SOAS, Endangered Languages Archive.