Licence Agreement for the Ethical and Sustainable Development of South African Sign Language Technology and Research

Report of the project, Advancing SASL for 4IR Technological Development using Place Names

Funded by the Department of Sport, Arts and Culture (South Africa)

April 2022 – June 2025

HMVE COMBRINK | P KECHE

sport, arts & culture
Department: Sport, Arts and Culture
REPUBLIC OF SOUTH AFRICA

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS INTERDISCIPLINARY CENTRE FOR DIGITAL FUTURES

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS THE HUMANITIES
SOUTH AFRICAN SIGN LANGUAGE AND DEAF STUDIES

FOREWORD

This research is the product of a project, Advancing SASL for 4IR Technological Development using Place Names, funded by the Department of Sport, Arts and Culture (South Africa), April 2022 – June 2025. It is a collaborative project between the Interdisciplinary Centre for Digital Futures and the Department of South African Sign Language and Deaf Studies, both at the University of the Free State (South Africa).

The following people made notable contributions to the sociolinguistic component of the project: Dr Chrismi Loth, Dr Patrick Sibanda, Dr Sara Siyavoshi, Ms Jani de Lange, Prof Annalene van Staden, Ms Emily Matabane, Ms Susan Lombaard, Prof Theodorus du Plessis, Mr Donovan Wright, Ms Lucia Mamotete Mapeshoane, Ms Kirsten de Villiers, Ms Gloria Motshoeneng, Mr Nhlanhla Simelane, Ms Anele Kotoyi and Ms Annemarie le Roux.

The following people made notable contributions to the computational component of the project: Dr Herkulaas Combrink, Mr Nkateko Nkuna, Ms Priscilla Keche, Mr Taylon Colbert, Mr Aviwe Matoti, Mr Jarryd Trip, Ms Valusha Oelofse and Mr Molemo Dibe.

A **Licence** is essential for the ethical and sustainable development of **South African Sign Language (SASL) technology and research**. Unlike spoken languages, signed languages are **visually complex**, requiring datasets that capture **hand movements, facial expressions, and spatial grammar**. These datasets are **scarce and highly valuable**, making them vulnerable to **exploitation, commercial lock-in, and restricted access** if not protected under an appropriate licensing model. Without structured safeguards, there is a real risk that **corporations and private entities** will extract, enhance, and monetise SASL datasets, creating **proprietary versions that limit accessibility and exclude the very communities these advancements are meant to serve**. We justify why we need a **CC BY-NC-SA** for **South African Sign Language (SASL) research and technology development.**

The **Non-Commercial (NC) clause** is critical in preventing **unauthorised commercialisation** while still allowing SASL research to thrive in **academic, educational, and community-driven spaces**. SASL, like many minority and low-resource languages, is **often overlooked in mainstream AI development** due to its smaller speaker base compared to dominant spoken languages. This makes open-access datasets particularly valuable for **sign language recognition models, translation systems, and educational tools**. However, without restrictions, large companies could **monopolise these resources,** incorporating them into **paid applications or closed-source AI models** while failing to **compensate or credit the researchers, linguists, and Deaf communities who contributed to the dataset's creation**. The NC clause ensures that SASL remains **a public good rather than a private commodity**, allowing free use in **schools, universities, and non-profit initiatives**, where the priority is **accessibility and community empowerment rather than profit.**

Attribution (BY) is equally vital in a linguistic space where **sign languages have historically been marginalised**. Unlike spoken languages with established written forms, SASL relies on **visual data and annotated datasets** that require **significant effort to compile, document, and refine**. Failing to credit contributors would **erase the work of Deaf educators, interpreters, and researchers,** reinforcing historical patterns where **signed languages were undervalued or disregarded in formal linguistic and technological advancements**. By requiring clear attribution, this licence ensures that **SASL dataset creators receive recognition**, establishing **academic legitimacy** and creating **incentives for further research and development**. In AI and machine learning, where models are frequently built on multiple datasets, attribution also **enhances transparency,** allowing developers to understand **the origins and biases** within the data they use.

The **Share-Alike (SA) clause** is particularly crucial for SASL because it ensures that **all future modifications, enhancements, and derivative datasets remain open.** As SASL datasets grow and improve, new data may include **better annotations, expanded vocabularies, regional variations, and increased accessibility for diverse signers**. Without Share-Alike protections, private entities could **modify these datasets, apply proprietary algorithms, and release closed-source versions** that restrict access, preventing further community-driven improvements. This would create **an uneven playing field**, where commercial players hold **monopolies on the best sign language AI models,** while educators, researchers, and Deaf communities are left with outdated or incomplete datasets. The SA clause mandates that **anyone building upon existing work must share their improvements under the same open licence**, ensuring that **knowledge and technological advancements are not hoarded but continuously circulated for the collective benefit of all users.**

For SASL, this licensing framework is not just about **data governance**, it is about **linguistic rights, digital inclusion, and the ethical advancement of AI in a way that serves, rather than exploits, the Deaf community.** Without it, the risk of **data extraction without reciprocation, commercial appropriation without recognition, and restricted accessibility without alternatives** would increase, reinforcing **inequities in technological access.** This licence provides a **sustainable model** that allows **research and development to progress while ensuring that SASL resources remain open, collaborative, and beneficial to all, rather than controlled by a select few.**

DR HMVE COMBRINK

CONTENTS

RECOMMENDED LICENSING MODEL: CC BY-NC-SA LICENSE

To legally enforce these principles while maintaining the openness of the project, we propose adopting the **Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) Licence** for all work within **South African Sign Language (SASL) research and technology development**. It ensures that linguistic, cultural, and technological advancements **remain accessible, transparent, and fairly distributed**. At the same time, it protects against **unauthorised commercialisation and exploitation**. Unlike spoken languages, **SASL and other signed languages** face significant challenges in **data collection, processing, and representation**. This is due to **limited resources, underrepresentation in AI and NLP models**, and the need for **specialised datasets** that accurately capture **hand movements, facial expressions, and non-manual markers.**

AI-driven tools are increasingly integrating **gesture recognition, avatar-based translation, and sign language processing**. However, there is a **high risk** that corporations, private developers, and proprietary platforms will **appropriate publicly available datasets**. With private funding, they may **enhance these datasets and restrict access to improved versions**. This could create **knowledge silos**, benefiting commercial entities while excluding the **Deaf and hard-of-hearing communities** these technologies are meant to support.

The **Non-Commercial (NC) restriction** prevents such appropriation. It ensures that **research institutions, non-profits, and educational organisations** can freely **use and improve** SASL datasets without **corporate lock-in**. However, it still allows for **structured commercial licensing**, ensuring **fair compensation** for continued dataset **refinement and expansion**.

The **Attribution (BY) requirement** guarantees that the **contributions of researchers, linguists, and sign language experts** are **properly credited**. This prevents the **erasure of community-driven linguistic work**, which forms the foundation of SASL datasets. It also **promotes accountability in AI development**, ensuring that **users know which datasets power specific sign language recognition models**.

The **Share-Alike (SA) clause** is particularly crucial for SASL. Many improvements to sign language datasets involve **adding new signs, refining annotation systems, or improving accessibility for diverse signers.** Without a **mechanism enforcing the release of modifications under the same open framework**, these advancements could be **locked away in proprietary datasets**. This would **block community-driven iteration, hinder linguistic preservation, and slow inclusive technological progress**.

SASL, like many **low-resource languages**, is vulnerable to **marginalisation in AI systems**. Large companies often **prioritise commercial viability over linguistic inclusivity**. Without **strong licensing protections**, corporations could **extract high-quality SASL data for private use** without contributing back to the **research ecosystem**. This would **stall collective progress**, restricting access to improvements and innovation.

By enforcing a **monitoring and compliance system**, this framework ensures that **users uphold the ethical and legal obligations of its licensing terms**. This **safeguards against misuse** while allowing for **responsible AI development** aligned with **accessibility, transparency, and fairness.** Without such protections, SASL technology risks becoming **dominated by commercial interests**, limiting access for **Deaf individuals, researchers, and educators** who rely on **open resources to advance linguistic and technological inclusion**.

This framework is not just about **data governance**. It is a **safeguard for linguistic rights, ethical AI development, and the empowerment of the SASL community**. It ensures that technological advancements in **sign language processing** serve their intended purpose — **enhancing communication, education, and accessibility for all**.

A CC BY-NC-SA LICENSE BREAKDOWN

To maintain **openness, fairness, and collaboration** while ensuring that contributions remain **accessible and protected**, this dataset and its derivatives are governed by the **Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) License**.

Licensing Conditions

1. Attribution (BY):

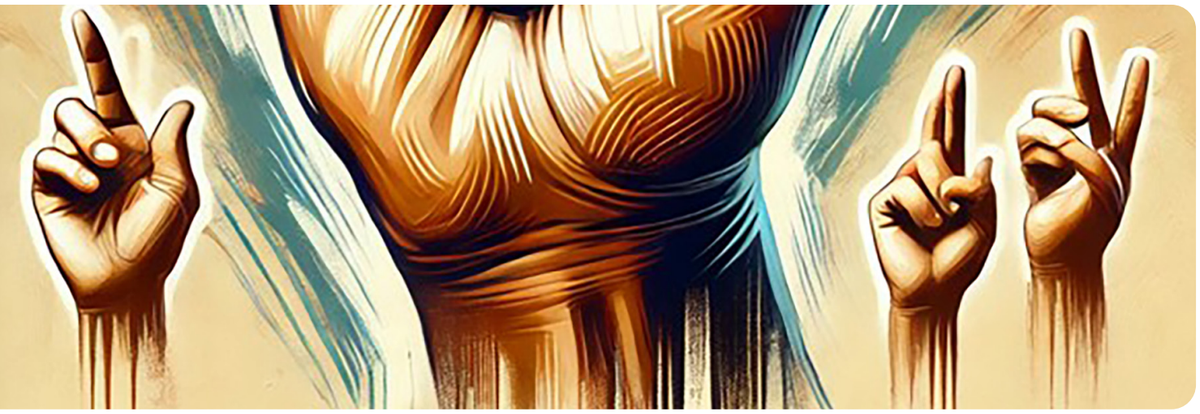
- Anyone using this dataset, its models, or applications **must provide clear credit** to the original project.
- Proper attribution must be included in **all publications, software, reports, and online platforms** that reference or incorporate this dataset.
- Citations should follow best practices to **acknowledge the project's contributions and maintain transparency**.

2. Non-Commercial Use (NC):

- This dataset and any derived works **can be freely used for research, education, and personal projects**.
- **Commercial use is prohibited** unless explicit **permission is granted** through a formal **commercial licensing agreement**.
- This ensures that the dataset remains a **public good**, supporting knowledge-sharing without unauthorized monetization.

3. Share-Alike (SA):

- Any modifications, enhancements, or derivative works **must be released under the same CC BY-NC-SA license**.
- This requirement **prevents privatization** of improvements, ensuring that all advancements remain **freely available** to the community.
- The principle of **open collaboration** is upheld by requiring that any adaptations **benefit the broader ecosystem**.



Commercial Licensing Requirement

*For organisations or developers seeking to **commercialise an application, product, or service** that integrates this dataset, **the following conditions apply:***

1. Obtain a Commercial License:

- A **formal agreement** must be established with the project entity.
- This ensures a **fair financial contribution** toward the dataset's ongoing development and maintenance.

2. Provide Proper Recognition:

- Any commercialized product **must visibly acknowledge** the SASL Data Sharing Framework as its foundational data source.
- The original dataset's contributions **must not be omitted or misrepresented**.

3. Ensure Transparency:

- The commercialized product **must disclose how the dataset and its derivatives have been used**.
- This helps maintain **accountability** while ensuring that **commercial adaptations remain traceable**.

Implementation & Enforcement

To uphold these principles and ensure compliance:

Clear Licensing Agreements:

- All distributed datasets, models, and applications **will include an explicit licensing statement** detailing these terms.

Usage Monitoring:

- A **tracking system** will be employed to monitor dataset usage and **verify adherence to attribution and licensing conditions**.

Legal Protections & Compliance Measures:

- **Violations of the licensing terms** may result in **restricted access** to the dataset.
- **Legal action may be pursued** to safeguard the project's integrity against unauthorized commercial exploitation.

Commitment to Ethical and Open Innovation

*By adopting the **CC BY-NC-SA License,** this framework ensures that **SASL research and technology:***

- Remains **accessible, inclusive, and beneficial** to the wider community.
- Encourages **collaboration and ethical development** without private monopolization.
- Protects against **uncompensated commercial exploitation**, fostering a fair and sustainable model for open innovation.

This structured approach **balances openness with responsible usage**, ensuring that **contributions to the SASL ecosystem continue to support global research, education, and technology development.**

ATTRIBUTION (BY) – GIVING CREDIT

The **Attribution (BY) requirement** ensures that **anyone using your dataset must acknowledge the original creators.** ***This includes:***

- Citing the **original dataset** in research papers, projects, and applications.
- Including a **visible attribution statement** in applications, software, or online platforms using the data.
- Linking back to the **original license or project webpage**, so users can see the full licensing terms.

Example Attribution Statement:

"This dataset is provided under the Creative Commons BY-NC-SA License. Original source: University of the Free State Department of Sport, Arts, and Culture SASL Data Sharing Framework."

The **Attribution (BY) requirement** is a fundamental principle of Creative Commons licenses, ensuring that the original creators receive proper credit whenever their dataset, research, or work is used, modified, or shared. This attribution mechanism fosters **transparency, ethical reuse, and academic integrity** while allowing open access to knowledge.

Why is Attribution Important?

Attribution serves multiple purposes beyond just giving credit:

1. **RECOGNITION:**
   Ensures that the dataset's original creators are acknowledged for their contributions.
2. **ACADEMIC & RESEARCH INTEGRITY:**
   Helps maintain ethical standards by ensuring that research is properly cited.
3. **TRANSPARENCY & TRACEABILITY:**
   Allows users to track the origin of data, ensuring credibility and reproducibility.
4. **ENCOURAGES FURTHER CONTRIBUTION:**
   When creators are credited, it incentivizes them to continue developing open resources.

By requiring attribution, the CC BY license **ensures that users cannot claim the dataset as their own**, reinforcing the principles of **open science, fair use, and collaborative knowledge-building.**

How Should Attribution Be Given?

A. Citation in Research Papers, Projects, and Applications

Whenever the dataset is used in an academic paper, research study, technical report, or any project, users **must provide a proper citation.** ***This means:***

- **Including a formal reference** in the bibliography, footnotes, or endnotes.
- **Mentioning the dataset's name, creators, and the licensing terms** within the document.
- **If applicable, specifying the dataset version or update date** to maintain reproducibility.

Example of Attribution in a Research Paper:

"This study used the University of the Free State Interdisciplinary Centre for Digital Futures Department of Sport, Arts, and Culture SASL Dataset, provided under the Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) License. Citation: Combrink, HMVE, P Keche, C Loth, K de Villiers, P Sibanda, S Siyavoshi, [Title of Dataset], [Year], available at [URL]."

Visible Attribution in Applications, Software, or Online Platforms

When the dataset is integrated into **a software application, machine learning model, or an online service, it must be visibly acknowledged.** ***This can be done in several ways:***

- **Displaying a credit statement within the software's UI** (e.g., an "About" section or splash screen).
- **Adding a visible acknowledgment** in the application documentation or README file.
- **Ensuring that derivative datasets or AI models also reference the original creators.**

Example of Attribution in Software Documentation:

"This application was developed using the University of the Free State Interdisciplinary Centre for Digital Futures Department of Sport, Arts, and Culture SASL Dataset, provided under the Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) licensed under CC BY-NC-SA. Original dataset: [URL]."

In AI and machine learning applications, when a dataset is used to train a model, the attribution must be **included in model descriptions, publications, and software repositories.**

Linking Back to the Original License or Project Webpage

A **core requirement** of CC BY attribution is that **users must provide a clear link** to the original dataset and licensing terms. ***This ensures that:***

- Future users **can access the dataset directly** from the original source.
- The **correct licensing conditions are always available,** preventing confusion or misuse.
- The dataset's **creators remain connected to their work**, maintaining visibility and credit.

The link to the **original dataset repository** (e.g., GitHub, Zenodo, institutional archive) ***should be included in:***

- **Online platforms where the dataset is hosted or discussed.**
- **Web applications using the dataset as a data source.**
- **Supplementary materials accompanying AI models, research papers, and datasets.**

Example of Proper License Linking:

"Data sourced from the University of the Free State Interdisciplinary Centre for Digital Futures Department of Sport, Arts, and Culture SASL Dataset, provided under the Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) License, licensed under CC BY-NC-SA. Full details and access at: [Original Dataset URL]."

Common Mistakes in Attribution & How to Avoid Them

Even though CC BY is simple, improper attribution is common. ***Some typical mistakes include:***

MISTAKE	WHY IT'S WRONG	CORRECT PRACTICE
Not providing any credit	Users fail to acknowledge the dataset, violating CC terms.	Always state the original dataset name and creator(s).
Vague or missing citation	Saying "Dataset from an online source" without specifics.	Include full citation with dataset title, authors, year, and URL.
Incorrect or missing license reference	Users cite the dataset but don't mention the CC BY-NC-SA license.	Always include "Licensed under CC BY-NC-SA" to clarify usage terms.
Failing to link to the original source	Users don't provide access to the dataset's webpage.	Add a direct, clickable URL to the dataset repository or licensing page.
Claiming ownership	Using the dataset without mentioning the original creators.	Clearly state that the dataset is from third-party creators and not self-produced.

By avoiding these mistakes, users **stay compliant** while ensuring that **the original dataset's integrity remains protected.**

How Attribution Strengthens Open Data & Ethical AI

Proper attribution is ***not just a legal requirement, it promotes:***

1. **TRUST & CREDIBILITY:**
   Users can verify **where data comes from**, increasing confidence in research and AI models.
2. **FAIR RECOGNITION:**
   Original creators get **credit for their contributions**, which can **help career progression** in research and academia.
3. **ENCOURAGES OPEN SCIENCE:**
   When datasets are properly cited, it inspires more collaboration and shared **knowledge**.

Final Thoughts: Why Attribution is Non-Negotiable

Attribution is the **simplest yet most powerful rule** of Creative Commons licensing. **Giving credit costs nothing,** but it **protects integrity, fosters collaboration, and ensures that valuable datasets remain open and accessible.**

By **citing properly, displaying credit visibly, and linking to the original source**, users **respect the work of creators while ensuring that knowledge remains freely available** for future innovation.

NON-COMMERCIAL (NC) – NO UNAUTHORISED COMMERCIAL USE

The **Non-Commercial (NC) clause** restricts people and companies from **profiting from the dataset** without a **commercial licensing agreement**. ***This means:***

- **ALLOWED:** Research, educational projects, personal use.
- **NOT ALLOWED:** Selling, licensing, or integrating the dataset into commercial applications.

If a **company, developer, or organisation** wants to **sell or monetize** a product using your dataset, they **must obtain a separate commercial license.**

Why is this important?

- Prevents **big corporations** from taking an open-source dataset and selling products without **contributing back.**
- Ensures **ethical innovation** by keeping **research free and public** while still allowing monetization **through a fair licensing system.**

The **Non-Commercial (NC) clause** in the **CC BY-NC-SA license** ensures that **your dataset can be freely used for research, education, and personal projects**, but it **cannot be used for commercial gain** without **explicit permission and a separate commercial license**. This restriction is **critical for maintaining open access** while ensuring that organizations profiting from the dataset **contribute fairly** to its development.

What Does Non-Commercial (NC) Actually Mean?

Under **CC BY-NC-SA**, the term **"Non-Commercial"** (NC) means that **any use of the dataset must not be primarily intended for or directed toward commercial advantage or monetary compensation.**

This means:

Allowed:

- **ACADEMIC RESEARCH:**
  Universities, researchers, and students can freely use the dataset for studies, papers, and non-profit research.
- **EDUCATIONAL PURPOSES:**
  Schools, online courses, and training programs can integrate the dataset into **teaching materials** and **curricula.**
- **PERSONAL PROJECTS:**
  Individuals can use the dataset for **experiments, hobby projects, and creative endeavours**, as long as they do not make money from it.

Not Allowed Without a Commercial License:

- **SELLING THE DATASET:**
  The dataset cannot be packaged and sold as a product or included in a paid service.
- **LICENSING FOR COMMERCIAL USE:**
  No one can **license or sublicense** the dataset for financial gain.
- **COMMERCIAL AI MODELS & APPLICATIONS:**
  If the dataset is used to **train AI models, develop software, or integrate into an app** that is later sold, this violates the NC clause **unless a separate agreement is made.**

The Commercial Licensing Requirement: When Do Users Need Permission?

While **Non-Commercial uses are freely permitted**, businesses, startups, and developers **who intend to make money** from the dataset **must obtain a commercial license.**

Who Needs a Commercial License?

1 TECH COMPANIES & STARTUPS

- If a company uses the dataset to develop a **commercial AI model, app, or software**, they **must** negotiate a licensing agreement.

2 PUBLISHERS & MEDIA COMPANIES

- If a media company wants to **use dataset-derived insights in commercial reports** or **paywalled content**, they need permission.

3 AI & MACHINE LEARNING SERVICES

- If a business trains an **AI system** using the dataset and **sells access to the AI**, this constitutes **commercial use**.

4 CONSULTING & ENTERPRISE SOLUTIONS

- Any firm offering **data analysis, trend predictions, or business insights** using the dataset must **obtain a license**

Why is the Non-Commercial Clause Important?

The **NC restriction** is a **critical safeguard** against **corporate exploitation and unethical commercialisation.**

A PREVENTS BIG CORPORATIONS FROM PROFITING WITHOUT CONTRIBUTING

- Without the **NC clause**, large companies could freely **use and modify** the dataset, then **sell AI-powered services or software** without giving back.
- This leads to a situation where **open-source research benefits only a handful of corporations**, rather than the **broader community**.
- The NC clause **forces businesses to compensate** for dataset usage, **ensuring fairness**.

B PROTECTS ETHICAL OPEN RESEARCH & PUBLIC ACCESS

- **Academic research and educational use remain unrestricted**, ensuring that scientists, students, and nonprofits can **innovate freely**.
- If the dataset were fully commercialized, **access could become restricted**, limiting knowledge-sharing.
- The NC clause **strikes a balance**:
  - Open for **nonprofit & public innovation.**
  - Restricted for **corporate profit-making** without fair compensation.

C ENCOURAGES FAIR MONETIZATION THROUGH LICENSING

- The NC clause **does not ban commercialization completely** — it just requires a **fair licensing system**.
  - ***This allows project creators to:***
    - Generate **sustainable funding** for continued dataset development.
    - Ensure that **businesses pay their fair share** if they generate revenue.
    - Support **community-driven governance**, where funds from licensing help improve the dataset.

Common Misunderstandings About Non-Commercial (NC)

The **NC restriction** sometimes leads to confusion. ***Here are some common misconceptions and clarifications:***

MISCONCEPTION	WHY IT'S WRONG	CORRECT UNDERSTANDING
"Non-Commercial means I can't use it at all."	NC only restricts commercial profit-making, nonprofit use is allowed.	You can use the dataset for research, education, or personal projects without payment.
"If I don't charge for access, my use is automatically non-commercial."	Even if a service is free, it may still be commercial if it supports a business or generates revenue.	If the dataset helps a business make money indirectly, it still requires a commercial license.
"I could sell my AI model if I trained it on an NC-licensed dataset."	Any commercial product built from the dataset violates NC unless separately licensed.	Businesses must negotiate commercial terms to monetize AI models trained on the dataset.
"As long as I don't sell the dataset itself, I can use it in a paid service."	Indirect commercial use (e.g., integrating it into paid apps, consulting services, or AI tools) also requires licensing.	Any profit-driven use of the dataset requires explicit permission.
"A non-profit can use the dataset for anything, even commercial services."	Some non-profits generate revenue (e.g., consulting, enterprise solutions, paid reports) that would still count as commercial use.	If a non-profit organization profits from the dataset, a commercial license may still be required.

How the NC Clause Strengthens Open Data & AI Ethics

By **enforcing the Non-Commercial condition**, datasets remain **widely available for learning and research** while **ensuring ethical AI development**.

The NC clause supports:

1 TRANSPARENCY & ACCOUNTABILITY:

Forces companies **to disclose when they use the dataset** and **how they integrate it.**

2 FAIR COMPENSATION MODELS:

Protects against **corporate free riding**, ensuring **financial contributions to the dataset's future development**.

3 SUSTAINABILITY:

Helps fund **ongoing dataset updates, maintenance, and expansion** through **commercial licensing revenue**.

SHARE-ALIKE (SA) – KEEPING THE COMMUNITY OPEN

The **Share-Alike (SA) clause** ensures that **all future modifications and improvements remain open**. ***This means:***

- If someone **modifies or improves** the dataset or models, they **must** release their version under the **same CC BY-NC-SA license.**
- They **cannot** turn a **public dataset into a private, proprietary product.**

Why does this matter?

- Prevents **companies or individuals** from **locking away improvements.**
- Encourages **collaboration and transparency** in the **SASL research community.**
- Ensures that **everyone benefits from advancements** rather than just **a few commercial players.**

The **Share-Alike (SA) clause** in the **CC BY-NC-SA license** ensures that **any modifications, improvements, or derivative works** based on the original dataset must also remain **open and freely available under the same license**. This means that **if someone enhances the dataset, develops new models using it, or creates a modified version**, they **cannot restrict access or claim exclusive ownership.**

By enforcing **Share-Alike**, the dataset stays **accessible, transparent, and beneficial to the entire research community**, rather than becoming **locked behind paywalls or proprietary licenses.**

What Does Share-Alike (SA) Actually Mean?

The SA clause ensures that anyone who modifies or builds upon the dataset must:

- **Release their modifications under the same CC BY-NC-SA license.**
- **Allow others to use, share, and build upon their improvements.**
- **Acknowledge that their new version is based on the original dataset.**

They cannot:

- **Re-license the modified dataset under a more restrictive or proprietary license.**
- **Claim ownership of the modified dataset as a private resource.**
- **Charge fees for access to their modified version while restricting others from using it.**

How Share-Alike Works in Practice

A. If Someone Modifies or Improves the Dataset or Models:

- If a researcher **cleans, structures, or expands the dataset**, they **must share the improved version** under the same CC BY-NC-SA license.
- If a developer **trains a new AI model using the dataset**, they **must allow others to use their model freely** under the same terms.
- If a team **adds annotations, translations, or extra metadata**, these enhancements **must be available for others to use and improve upon.**

Example Scenario:

- A university researcher uses your dataset to **create a new, more accurate version** with additional data points.
- Instead of keeping it private, they must **publish their improved version under the same CC BY-NC-SA license.**
- Now, other researchers and developers can **continue building upon it**, making further improvements.

Preventing Public Data from Becoming Private or Proprietary

A major risk with open data is that **companies or individuals could take publicly available resources, modify them slightly, and then restrict access**. The SA clause **prevents this from happening**.

Allowed:

- Using the dataset for research, education, and Non-Commercial projects.
- Modifying and enhancing the dataset **as long as the new version remains open.**

Not Allowed:

- Taking an **open dataset**, adding minor changes, and **claiming exclusive ownership.**
- Turning a **community dataset into a closed-source, pay-to-access product.**
- **Restricting others** from using improved versions by placing them under a different license.

Example Violation:

A company takes an **open dataset**, cleans it, and **sells access to the modified version** without sharing the cleaned dataset back with the community.

- They claim their dataset is **"proprietary"**, blocking others from using their version.
- This **violates the SA clause**, because **any modified version must also remain open.**

Why Does Share-Alike (SA) Matter?

A. Prevents Companies or Individuals from Locking Away Improvements:

- Without Share-Alike, **private companies could take open datasets, improve them, and then restrict access.**
- This would mean that **the original creators and the broader community would never benefit from these improvements.**
- SA ensures that **no matter who improves the dataset, the advancements remain freely available.**

Real-World Example:

- Wikipedia is licensed under **CC BY-SA**—this means that **if someone copies Wikipedia, they cannot put their version behind a paywall.**
- This ensures that **Wikipedia remains an open knowledge source for everyone.**

Similarly, with a **Share-Alike dataset**, no company can **take it private, profit off it, and block others from using the enhancements.**

Encourages Collaboration and Transparency in the SASL Research Community

- Share-Alike creates a **collaborative ecosystem** where researchers, developers, and institutions can **build upon each other's work.**
- It ensures **continuous improvement** of the dataset without barriers.
- Researchers can **combine datasets, improve AI models, and refine methodologies**, knowing that **everyone in the community benefits.**

Example of a Collaborative Cycle with Share-Alike :

- A university team **cleans and enhances** the dataset.
- A machine learning lab **uses the improved dataset** to train an AI model.
- A developer **fine-tunes the AI model** and releases it for educational use.
- A non-profit **applies the AI model to a real-world problem** (e.g., improving accessibility for Deaf communities).
- The entire research community benefits from **a continuously evolving, high-quality dataset.**
- Without Share-Alike, improvements **would remain isolated**, and the entire ecosystem **would suffer from fragmentation.**

Ensures That Everyone Benefits from Advancements, Not Just a Few Players

- **Public data should benefit everyone, not just those who can afford to monetize it.**
- If Share-Alike wasn't in place, **large tech companies could take open datasets, refine them, and sell them as proprietary services.**
- This would mean that **only those who pay would get access to the best improvements**, creating **an unfair knowledge gap.**

Example of Share-Alike Preventing Corporate Monopoly:

- A tech startup uses an **open speech dataset** to develop an **advanced speech recognition AI.**
- If they **were allowed to re-license, it under a restrictive license**, only their company would benefit.
- Because of **SA**, they must **share their AI model improvements back with the community**, ensuring **equal access.**
- The **SA clause keeps innovation fair, ethical, and open.**

Common Misunderstandings About Share-Alike (SA)

MISCONCEPTION	WHY IT'S WRONG	CORRECT UNDERSTANDING
"Share-Alike means I can't modify the dataset."	Share-Alike allows modifications — it just requires that modified versions remain open.	You can modify, improve, and redistribute the dataset as long as you use the same CC BY-NC-SA license.
"I can change the dataset and license it under a more restrictive license."	SA prevents restricting access to modified versions.	If you modify the dataset, you must share it under the same license, keeping it open for others.
"A company can improve the dataset and sell access without sharing the updates."	This violates SA because improvements must be shared under the same open license.	No one can privatize or restrict improvements; all modifications must remain freely available.
"If I modify the dataset, I own the new version and can restrict access."	While you own your modifications, you must share them openly.	Ownership does not override the Share-Alike requirement, you can't block others from using your improvements.

Final Thoughts: Why Share-Alike is Essential for Ethical AI & Open Data

The SA clause ensures that:

- **Public knowledge remains open and accessible.**
- **No single company or individual can privatize community-driven work.**
- **The entire research and AI community can continue benefiting from new advancements.**

REFERENCE LIST

1. **De Meulder, M., Murray, J.J. and McKee, R.L.,** 2019. Introduction: the legal recognition of sign languages: advocacy and outcomes around the world. ***The Legal Recognition of Sign Languages. Bristol: Multilingual Matters,*** pp.1-16.
2. **Kopf, M., Schulder, M. and Hanke, T.,** 2022, June. The sign language dataset compendium: creating an overview of digital linguistic resources. In ***Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*** (pp. 102-109).
3. **Zhou, H., Zhou, W., Qi, W., Pu, J. and Li, H.,** 2021. Improving sign language translation with monolingual data by sign back-translation. In ***Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*** (pp. 1316-1325).
4. **National Academies of Sciences, Medicine, Global Affairs, Board on Research Data, Information and Committee on Toward an Open Science Enterprise, 2018. Open science by design: Realizing a vision for 21st century research.**
5. **Marschark, M. and Spencer, P.E.,** 2010. ***The Oxford handbook of deaf studies, language, and education, vol. 2.*** Oxford University Press.
6. **McMillan-Major, A.,** 2023. ***Language*** Dataset Documentation Design: Learning from Deaf and Indigenous Communities. University of Washington.
7. **Bragg, D., Koller, O., Caselli, N. and Thies, W.,** 2020, October. Exploring collection of sign language datasets: Privacy, participation, and model performance. In ***Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*** (pp. 1-14).
8. **Quer, J. and Steinbach, M.,** 2019. Handling sign language data: The impact of modality. ***Frontiers in psychology, 10,*** p.483.
9. **Bragg, D., Caselli, N., Hochgesang, J.A., Huenerfauth, M., Katz-Hernandez, L., Koller, O., Kushalnagar, R., Vogler, C. and Ladner, R.E.,** 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. ***ACM Transactions on Accessible Computing (TACCESS), 14***(2), pp.1-45.
10. **Napier, J. and Haug, T.,** 2017. Justisigns: A European overview of sign language interpreting provision in legal settings. ***Law, Social Justice & Global Development, 2016***(2).
11. **Haualand, H.,** 2009. Sign language interpreting: A human rights issue. ***International Journal of Interpreter Education, 1***(1), p.7.
12. **Pratt, V.F. and Williams, P., State Licensing of Sign Language Interpreters.**
13. **Haualand, H.,** 2023. Licence to inform: Norwegian sign language interpreters in a bureaucratic organisation. ***Interpreting and Society, 3***(1), pp.6-23.
14. **Othman, A.,** 2024. Sign language varieties around the world. In ***Sign Language Processing: From Gesture to Meaning*** (pp. 41-56). Cham: Springer Nature Switzerland.
15. **Salonen, J., Kronqvist, A. and Jantunen, T.,** 2020, May. The Corpus of Finnish Sign Language. In ***Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*** (pp. 197-202).

*Understanding **South African Sign Language (SASL) and its role in technology, research, and accessibility** is more than just an academic or legal concern, it is a **social imperative.** This report is not merely a collection of legal principles and licensing frameworks. It is a **call to action** for researchers, developers, policymakers, and advocates who believe in the power of **open knowledge, ethical AI, and linguistic rights**.*

*SASL, like all signed languages, is more than just a means of communication, it is **a cultural and linguistic identity** for the Deaf community. Yet, it remains **one of the most underrepresented languages in AI and digital technology**. As advancements in **machine learning, gesture recognition, and AI-driven translation** accelerate, there is both an **opportunity and a risk**. The opportunity lies in creating **inclusive technologies** that empower SASL users by bridging communication gaps and expanding access to information. The risk, however, is that **corporations and private entities could monopolise these resources**, extracting community-driven data while restricting access to the very people who need it most.*

*This report explores the importance of **structured licensing through the Creative Commons Attribution-Non-Commercial-Share-Alike (CC BY-NC-SA) Licence.** It explains how this framework **ensures that SASL datasets remain open for research, education, and technological innovation**, while also preventing **uncompensated commercial exploitation**. Without such protections, there is a real danger that **SASL resources could become locked behind paywalls, owned by a handful of entities rather than being a shared linguistic and cultural asset.***

*By reading this report, one gains **a deeper understanding of the ethical, legal, and technological challenges facing SASL and why this licence justification is important**. It highlights **why transparency, attribution, and equitable access are essential** in AI development. More importantly, it offers a **solution, a licensing framework that balances openness with responsibility, collaboration with recognition, and technological advancement with social justice.***

*This report is for those who **care about the intersection of language, technology, and human rights.** It is for researchers seeking **fair ways to share their work**, developers committed to **ethical AI**, and policymakers striving to **protect linguistic diversity in the digital age**. It is for those who believe that technology should serve **all communities, not just those who can afford to control it**. Reading this report means engaging in **a conversation about fairness, inclusion, and the future of SASL in an AI-driven world**. It is a step toward ensuring that **technological progress does not come at the cost of linguistic equity but rather strengthens it**.*

sport, arts & culture
Department:
Sport, Arts and Culture
REPUBLIC OF SOUTH AFRICA

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS INTERDISCIPLINARY CENTRE FOR DIGITAL FUTURES

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS THE HUMANITIES
SOUTH AFRICAN SIGN LANGUAGE AND DEAF STUDIES