Long-Term Sign Language Data Crowdsourcing Through Collaborative Lexicons

Pierre Poitier, Jérôme Fink, Ariel Basso Madjoukeng, Adélaïde Couplet, Margaux Leleu, Benoît Frénay
NADI/HuMaLearn at University of Namur
Rue Grangagnage 21, Namur, Belgium
{pierre.poitier, jerome.fink, ariel.bassomadjoukeng}@unamur.be

Abstract

While there exists a multitude of different sign languages (SLs) across the world, deaf communities often lack the digital tools required to document and process their languages. In this work, we introduce Mot-Signe (MOSI), an application designed in close collaboration with actors from the French Belgian deaf community. Our tool enables users to search for French Belgian Sign Language (LSFB) translations or to propose new ones by recording signs themselves. This crowdsourcing approach facilitates the collection of SL data in the wild, enriching the available documentation on LSFB and proposing an innovative response to the data scarcity issue inherent to sign language processing. To evaluate the sustainability of this community-driven data collection, a longitudinal user study was conducted. Following its public release, MOSI demonstrated significant real-world adoption, enabling the collection of over 3,000 distinct LSFB signs. Notably, MOSI captures highly valuable linguistic variations and specialized vocabulary often absent from traditional corpora.

Keywords: sign language, crowdsourcing, dictionary, data collection, data gathering, LSFB

1. Introduction

Sign Languages (SLs) serve as the primary mode of communication for deaf communities worldwide. Like all natural languages, SLs have evolved organically within their respective communities, resulting in a rich linguistic diversity and a multitude of distinct dialects. However, despite this prevalence and a recent surge of interest in the machine learning (ML) community (Rastgoo et al., 2021), SLs remain severely under-resourced. The deaf community still suffers from a lack of suitable digital tools, such as comprehensive dictionaries, large language models, or robust automatic translation systems.

Several structural barriers hinder the application of existing methods. First, SL data is sparse compared to textual data available on the web. The largest SL datasets contain approximately $200,000$ sign occurrences (Starner et al., 2023), whereas spoken language models benefit from datasets containing billions of word occurrences (Patel and Patel, 2020). Second, the sociopolitical status of SLs plays a role; according to the World Federation of the deaf, 59% of countries have not yet officially recognized their national sign language, leading to stigmatization and a lack of funding for documentation. Third, the technical challenge is significant: SL data consists of high-dimensional video rather than text, making it computationally expensive to process. Furthermore, annotating this data requires skilled interpreters, making the process tedious, time-consuming, and costly. Consequently, SLs remain among the least documented and studied languages in the world (Fink et al., 2023a).

Beyond data scarcity, there is a fundamental linguistic challenge: there is no direct equivalence between SLs and spoken languages. Translating SL to a written form (e.g., French or English) is non-trivial due to differences in syntax and modality. Moreover, sign languages are dynamic; they evolve rapidly, with each generation introducing neologisms and variations that static dictionaries fail to capture. To effectively document the language, a system is required that can track these new signs and dialectal specificities in real time.

In this work, we introduce *Mot-Signe* [1] (MOSI), an application designed in close collaboration with actors from the French-Belgian deaf community. MOSI addresses the data sparsity issue by creating a sustainable, community-driven ecosystem. Unlike pure data collection initiatives, MOSI creates value for the user first, with data collection acting as a beneficial side effect. This aim gives MOSI a double-sided contribution. One of those is a social contribution that focuses on the needs of the community, and the other one is a technical one that is directed towards the data collection methodology. This paper focuses on the second one. In summary, the goals of the application are:

- to establish a long-term, sustainable data collection system to enrich existing dictionaries and corpora.
- additionally, to provide a well-integrated, free,

[1] https://www.mot-signe.be/

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 419–428
16 May 2026.

and public French to French Belgian Sign Language (LSFB) lexicon;

MOSI builds upon the static, studio-recorded corpus of the LSFB-Lab (Meurant, 2015; Fink et al., 2021) to deploy a dynamic crowdsourcing approach for "in-the-wild" data collection. This strategy captures realistic recordings from diverse signers, specifically targeting domain-specific vocabulary absent from academic datasets. We validate this approach through a user study on community retention and analyze data from the public release. Results are promising: among the $3,000$ new video entries contributed, over 80% represent entirely new vocabulary not previously documented in the public lexicon.

2. Background and Related Work

This section reviews the background of Sign Language (SL) data collection and digital resources. We first present traditional data acquisition methods and their inherent limitations. Next, we examine existing online SL dictionaries and tools available to the deaf community. Finally, we explore crowdsourcing as a viable alternative, highlighting its benefits, prerequisites, and how previous works have applied it to build robust SL datasets.

2.1. Sign Language Data Acquisition

The creation of large representative SL corpora was initiated by linguists to document these languages and analyze their evolution. While these resources are now essential for leveraging machine learning (ML) algorithms, their primary design was for linguistic inquiry. The first notable corpus is the Auslan Corpus (Johnston, 2009) that contains 300 hours of Australian SL recordings. Inspired by this approach, several teams started to collect their own SL corpus, e.g., the LSFB Corpus (Meurant, 2015) and the British SL (BSL-1K) Dataset (Cormier et al., 2012). To study how the language is used by several categories of the population, those corpora often involve a relatively large number of signers to be representative. The recordings are often performed in a studio with professional cameras and controlled lighting. Consequently, while these studio-recorded datasets offer high-quality documentation, a first limitation is that they are not well suited for ML algorithms. Models trained on this clean, controlled data often struggle to generalize to the visual noise and variable conditions of real-world webcam recordings. Another limitation is that the annotation process is time-consuming and requires skilled workers. Studies show that annotating one hour of video requires up to 100 hours (Renz et al., 2021). Thus, only subsets of collected data are usually annotated and this acts as a major bottleneck for the linguistic analysis, even if the data exists. The format chosen for the annotation may also vary from one team to another, making them hard to leverage (Fink et al., 2023a).

Another source of sign language data are public lexicons developed by the deaf community. These lexicons contain videos of signs associated with their meaning in a spoken language. Such data are only relevant in the case of Isolated Sign Language Recognition (ISLR) (De Coster et al., 2023) that focuses on the recognition of sign videos containing only one sign. Usually, all signs in a lexicon are performed by the same signer and they are recorded in a controlled environment. Algorithms built upon those data are, therefore, unlikely to generalize to other signers and environments. In an effort to build larger and more diverse datasets, ML researchers started to seek data sources in the wild. The first data sources considered in the literature are TV broadcast SL translations. The Phoenix WTH dataset (Forster et al., 2012) is a collection of SL interpretations of German weather forecasts translated manually by interpreters. The BBC-Oxford British SL Dataset (BOBSL) (Albanie et al., 2021), building upon BSL-1K, introduces a method to align BBC broadcast transcription with its SL interpretation. Such datasets are larger and require less effort than collecting traditional SL corpus. Those solutions drastically increase the size of datasets, but the signers transcribing broadcasts are, usually, always the same and the recordings are still performed in a controlled environment. TV broadcasts also depict a specific vocabulary that differs from the usage of the language in the wild. The deaf community is skeptical about this approach, as interpreters are often non-native SL speakers and the translations of spoken languages are not genuine SL production (Fink et al., 2023a).

Social media constitutes another data source, e.g., Youtube-SL (Tanzer and Zhang, 2024), or video platforms where deaf people share vocabulary. The main issue with these methods lies in the fact that authors do not always give their consent for their data to be used in ML datasets or corpora (Saunders et al., 2021). Such recent datasets, e.g., ASL (Desai et al., 2024), tend to depict more realistic data captured in the wild.

Another limitation is that many existing SL datasets contain a number of biases (Atwell et al., 2024). For example, they are often produced under a controlled environment: uniform background, standard lighting conditions, etc. They may also only concern a lexical field specific to a particular theme, contain very few different signers or present a majority of non-native signers. All these biases make sign language recognition tools built upon aforementioned data sources difficult to transpose into real conditions.

2.2. Online Sign Language Tools

Several works have focused on the development of online tools or dictionaries dedicated to sign languages. Pioneer tools are Auslan and BSL SignBank, but many tools have been inspired by them, such as Elix [2], LSFB Dico [3], LifePrint [4], etc. Elix is a SL dictionary that maps the French language to the French SL (LSF). It consists of more than $27,000$ different signs and has approximately $70,000$ users per day. The LSFB Dico is a dictionary proposed for the translation from French to Belgian French SL (LSFB). It contains more than $5,900$ different signs and has the particularity of being very diversified, as many signs appear with multiple variants. These tools are specifically designed for a particular SL and do not account for multilingualism. Recently, multilingual dictionaries have emerged, covering several SLs. One of the most popular multilingual sign language tools is SpreadTheSign [5], a dictionary consisting of more than $200,000$ signs across 40 sign languages. This dictionary has achieved considerable success in the context of multilingual sign languages. However, although these tools constitute a valuable resource for deaf communities, their primary objective is to be queried. They are not designed for sign language data collection, as they do not provide signers with the ability to add new signs. We propose a new sign language tool, MOSI, that acts as a collaborative sign language dictionary where data collection is embedded as a side effect. To the best of our knowledge, sign language dictionaries have never been paired with long-term crowdsourcing solutions.

2.3. Crowdsourcing

Crowdsourcing designates the process of outsourcing tasks to crowd (Hossain and Kauranen, 2015). There are three main reasons that explain why crowdsourcing might be useful in the context of SLs (Garcia-Molina et al., 2016): first, reducing **costs** to realize annotation tasks; second, reducing **latency** by parallelizing the tasks to multiple people; third, reducing **uncertainty** by comparing the outputs of multiple people. It has to be noted that none of those three objectives can be fully met at once. Reducing uncertainty often requires getting insights from a bigger crowd, increasing the cost and the latency. Members of the crowd also may need incentives to perform the tasks. The most common one is to give money in exchange for the task. But other incentives exist, such as philanthropy or entertainment. The Wikipedia project is a great example of philanthropic crowdsourcing. Users are not paid, but willingly participate in this project of general interest (Yuen et al., 2011). This is similar in MOSI, where the general interest is the documentation of the LSFB.

Crowdsourcing the collection and labeling of LSFB signs has the potential to collect a large amount of data recorded under various conditions. Also, the deaf community is used to share vocabulary on social media platforms to compensate for the lack of up-to-date lexicon. Riemer Kankkonen et al. (2018) use data posted by the members of the Swedish SL community to gather more vocabulary for their dictionary. This approach is interesting, as it leverages a well-known platform for the community and is not invasive. However, the various posts and videos shared on the platform must be collected manually, as this process is hard to automate.

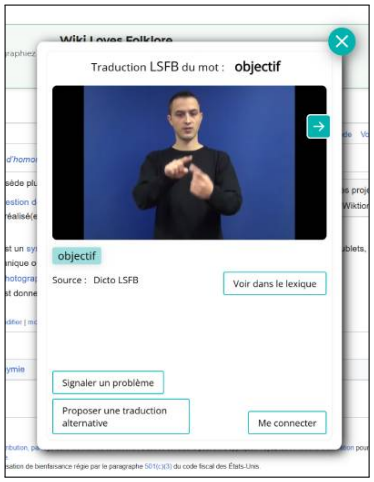
To automate data collection, researchers have developed specialized crowdsourcing platforms for sign languages. For American SL (ASL), both Bragg et al. (2022) and Starner et al. (2023) present platforms to collect and validate SL data, demonstrating that it is possible to gather high-quality data. They also report that users are willing to contribute for a reasonable compensation. Dhruvo et al. (2023) built the Bangladeshi SL dataset and collected more than $21,000$ videos using crowdsourcing. Kapitanov et al. (2023) leveraged two crowdsourcing platforms (Yandex Toloka[6] and ABC Elementary[7]) for recording and validation, creating the Russian SL (RSL) dataset with $20,000$ videos across $1,000$ different signs. This demonstrates that crowdsourcing can be an effective way to build datasets, reducing costs by eliminating the need for financial incentives, although it slightly increases the time required for data acquisition.

We observe that the primary motivation behind the existing platforms is data collection without starting from a need of the deaf community. This may result in a solution that is less relevant and not user-centered, reducing the likelihood of adoption and long-term engagement. In our study, we build on a need identified within the deaf community, using it as an incentive for application usage. Consequently, data collection and validation become secondary outcomes of the application. The expected result is improved user retention along with a long-term data collection process.

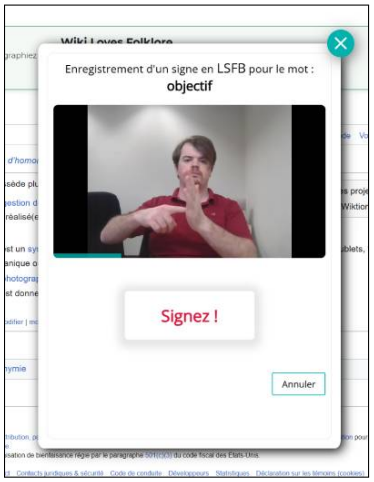
In summary, while traditional corpora and online dictionaries provide valuable linguistic resources, they remain static and do not leverage the community as an active contributor. The following section presents how MOSI bridges this gap.

[2] https://dico.elix-lsf.fr/
[3] https://dico.lsfb.be/
[4] https://www.lifeprint.com/
[5] https://www.spreadthesign.com/
[6] https://platform.toloka.ai/
[7] https://elementary.center/

(a) Selection: The user highlights a word to trigger the extension icon.



(b) Translation: The overlay displays the LSFB video from the public or private lexicons.



(c) Recording: Users can record missing signs to add to their personal lexicons.

Figure 1: Typical usage scenario of MOSI. (a) Users select a word on any webpage, (b) consult available translations from the layered lexicons, and (c) add their own signs to their lexicons.

3. Our Tool: *Mot-Signe*

This section introduces *Mot-Signe* (MOSI), our platform designed to bridge French and French Belgian Sign Language (LSFB). We detail the motivations driving the tool's evolution from a simple lexicon to a collaborative documentation platform. We then discuss the design and key features, specifically focusing on how the architecture prioritizes user privacy through private lexicons, while enabling the enrichment of public LSFB corpora.

3.1. Goals and Motivations

The development of MOSI was prompted by a direct request from the deaf community in the French-speaking part of Belgium. Our research team was contacted by stakeholders who identified a severe lack of integrated tools for signers. Citing the utility of *Elix* for French SL (LSF) and *AuslanSignbank* [8] for the Australian SL, they expressed a strong interest in a similar solution that seamlessly provides LSFB equivalents for French words directly within a web browser. The primary goal is, therefore, to foster autonomous learning and digital inclusion, by reducing reliance on external human assistance for understanding daily web content.

This collaboration presented a secondary opportunity for the research community: the large-scale documentation of the language. Existing SL datasets, including LSFB, often lack specialized vocabulary or natural variations for given concepts. Drawing inspiration from collaborative initiatives like *signdict* [9], MOSI aims to fill this gap by empowering the community to document signs that are currently absent from static lexicons. This supports two research objectives: providing signers with a richer variety of expressions for a single word, and establishing a sustainable data collection system. This system is essential for enabling linguistic research into sign variations often missing from current corpora, as well as for improving future dictionary-based AI query systems that rely on comprehensive lexical coverage.

This study builds upon a preliminary prototype developed in a master's thesis (Ye and Sow, 2023) and focuses on the creation of a sustainable data collection system for LSFB, designed to be easily transposed to other SLs in the long term.

3.2. Features Overview

MOSI operates as a unified platform accessible via a web extension (compatible with Chrome and Firefox) and a dedicated website (`mot-signe.be`). The core functionality revolves around a privacy-first architecture that distinguishes between personal usage, trusted sharing with private lexicons, and public lexicons that every user can access.

Query Signs From a Selected Word

Users can query for LSFB translations, either directly on the website or via the web extension from any external web page. Upon selecting or typing a French word (see Figure 1a), the user triggers a query that aggregates results from three distinct sources (see Figure 1b):

1. **Public Lexicon:** Standardized signs managed by linguistic experts.

[8] https://auslan.org.au/dictionary/
[9] https://signdict.org/

2. **Shared Lexicons:** Signs contained in private groups (lexicons) the user has joined via invitation codes.

3. **Personal Lexicon:** Signs recorded by the user for their own usage.

This approach ensures that users can access authoritative vocabulary while simultaneously retrieving specialized or informal signs relevant to their specific context (e.g., a biology class or a family circle) without exposing those signs to the public.

Contribution and Private Sharing

If a translation is missing or if a user wishes to document a variation, they can propose an alternative translation directly through the extension interface (see Figure 1c). This process requires the user to be logged in. The tool opens a webcam stream, allowing the user to record, review, and confirm their sign.

By default, these recordings are **private**. To share them, users must create or join a **Lexicon**. A single sign can be associated with multiple lexicons. Each lexicon is managed by **administrators** who control membership via invitation codes and moderate the signs shared within that specific group. This ensures that signs are only visible to intended peers and that the group maintains its specific focus. The companion website serves as the management hub where users can organize their lexicons, manage their profile, and visualize their contributions. In addition, users can download the signs as GIF animations to use them in external tools (e.g., slideshows in teaching environments or interactive websites).

Expert Curation and Documentation

The platform facilitates a unique workflow for language documentation. Linguistic experts and administrators have access to the pool of recorded signs to identify missing vocabulary or emerging variations. Unlike traditional crowdsourcing where user videos are simply “approved”, MOSI uses user contributions as field data. Experts analyze these contributions and, when appropriate, produce high-quality, standardized studio recordings to enrich the **Public Lexicons**. This ensures that public resources maintain high production standards while being driven by community usage.

3.3. Data Governance and Corpus Enrichment

While the primary interface serves the immediate needs of the deaf community, the underlying architecture is designed to address the scarcity of digital LSFB resources. The data collection process is strictly governed by ethical protocols and GDPR compliance, ensuring that the goal of documentation never overrides user privacy.

Integration with Existing Corpora

Current LSFB datasets are predominantly static, studio-recorded corpora created by linguistic research groups: LSFB Corpus (Meurant, 2015) and LSFB Dictionary (Sonnemans, 2026). While high in quality, they often lack the "in-the-wild" diversity of natural signing and fail to capture neologisms or regional variations rapidly. MOSI acts as a dynamic extension to these existing corpora. It does not aim to replace them but to identify gaps (missing signs) and variations (accents, synonyms) through community usage.

Privacy and Anonymization Strategy

To ensure GDPR compliance and build trust with the community, the platform integrates strict privacy protocols. Upon registration, users are presented with clear guidelines regarding data usage to ensure informed consent. Users retain full sovereignty over their contributions, including the right to withdraw their data at any time. To guarantee anonymity while maintaining research utility, we employ a two-tier strategy.

Curatorial Anonymization Public exposure of user data is prevented by design. User recordings act solely as field data for linguistic experts. Validated signs are **re-recorded** by designated signers in a controlled studio environment for the public lexicon. This intrinsically anonymizes the source, as the public never views the original contributor’s video.

Computational Anonymization For downstream machine learning tasks and private research analysis, we mitigate re-identification risks by discarding pixel data in favor of vector representations. We utilize *MediaPipe* (Lugaresi et al., 2019) to extract pose landmarks (skeletons) from the videos. This removes sensitive visual identifiers (face, background) while preserving the linguistic motion required for analysis (Saunders et al., 2021).

4. User Study

Generally, the adoption of a tool depends on several criteria (e.g., ease of use, proposed features, etc.), the amount of data captured is correlated with its adoption by final users. To understand whether our tool meets these criteria and how it fits into users’ daily practices, a longitudinal user study was

conducted. This study had two objectives: (i) to collect qualitative data on the use of the application; and (ii) to assess how the application fits into users' daily practices.

4.1. Methodology

During this study, nine users, all proficient in LSFB, were asked to fill out diary entries over a period of time, allowing us to identify their habits and follow their journey while using the application. Participants were recruited in April 2024 during the launch event of the application at a bilingual (hearing and deaf) school. An account with access to all the features was created for all attendees, and the features were presented. The study was conducted over 10 opening days, with one entry per day. It began one week after the creation of the users' credentials. This allowed them to become more familiar with the application before starting to track their habits. The study was conducted remotely; therefore, instructions on how to fill out the diary were sent by email, and a reminder was sent every day at 4 p.m. Many questions of different types (features, accomplished tasks, difficulties faced, issues, etc.) were asked of the users. These questions allowed us to track the usage of the various features and the issues encountered by the users. After the study, a debriefing was conducted in person with each user to clarify some points in their diary and gather their opinion about the application.

4.2. Usage of the Application

The entries of the diary allow identifying which features are mainly used and how users adopted the application. Table 1 summarizes the features mentioned by the users in their diary entries along with the number of occurrences of each feature in the texts.

Table 1: List of the features mentioned in the diaries along with the number of times they are mentioned.

Feature	Occurrences
Vocabulary search	35
Record vocabulary	8
Download sign GIF	4
Share signs in lexicons	4

As expected, the *vocabulary search* (either using the website or the extension) is, by far, the most popular feature as it eases the search for vocabulary in texts. It was the initial feature requested by the stakeholders. Then, the users mentioned the *record sign* feature several times. It was often mentioned as teachers prepared independent reading exercises where pupils learned to read on their own. Two users explained that they selected the text the student would be reading and they made sure that all the vocabulary was already in the lexicon by recording the missing signs. They also prepared a list of synonyms as, currently, the search in the lexicon does not handle synonym search. During the reading class, deaf students could access a computer with MOSI available to search for the vocabulary. The teachers considered it as a way of increasing their students' autonomy.

Other users shared that they used the *record sign* feature to build their own vocabulary for exercises. Once recorded, they downloaded the corresponding GIF version of their recorded signs. It enabled them to embed the signs in their course material (e.g., flashcards or slides). This also explains why the *download sign GIF* feature is often mentioned in the diary study.

Finally, some users mentioned that they browsed the *lexicon pages* of the website to see the vocabulary shared by their colleagues and to validate some signs.

4.3. Users Feedback

User feedback is divided into two main categories: disincentive and incentive factors.

Disincentive factors refer to elements that negatively impact the user experience. In this category, the most frequent problems concern the vocabulary search. Indeed, the lexicon is too limited and some words too old or too specific are not available (7 reports). The current search engine is quite simple and could be improved. Therefore, when a user searches a conjugated verb, the application is not able to find the infinitive form (3 reports). Another consequence is the fact that the synonyms or homonyms are not detected by the search engine (4 reports). Sometimes, the application returns incorrect signs either because of similar spellings, annotation mistakes or because of the specific context of the sentence (3 reports).

On the other side, incentive factors are elements that make the application appealing to the user and would lead to a better retention. First, the fact that multiple public LSFB lexicons are merged and accessible directly via the extension is appreciated and saves time (reported 3 times). The ability to share vocabulary with other people is also praised (3 reports). The ability to download GIF images of the signs eases the work of teachers (3 reports). The integration of the extension in every web page also allows users to easily discover new signs (2 reports).

5. Community Adoption

Following the initial user study and subsequent iterative refinements, MOSI was released publicly to the broader deaf community in French-speaking Belgium, i.e., the whole LSFB community. This phase marks the transition from controlled testing to "in-the-wild" usage, allowing us to observe how the tool performs with real users in diverse, uncontrolled environments.

5.1. Quantitative Analysis

Since the public release, the platform has attracted 224 registered users, in addition to a broader base of unregistered users who can query the lexicons without an account but cannot record signs. As illustrated in Figure 2, the volume of contributions has shown a steady upward trend, indicating sustained user engagement rather than a momentary spike at launch. A promotional event held in April 2025 led to a notable surge in contributions, as visible in the figure.

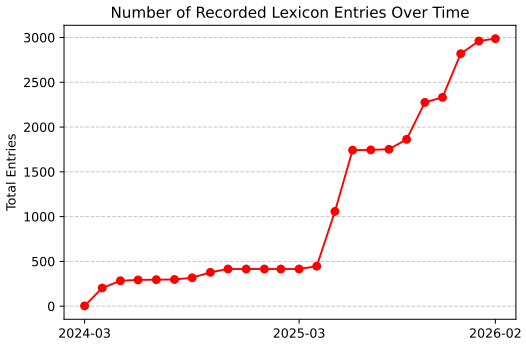


Figure 2: Cumulative evolution of the number of signs recorded on the platform since the public release. The sudden increase corresponds to a promotional event in April 2025.

Data Collection Volume To date, a total of $3,004$ distinct signs, identified through ID-glossing, have been recorded and added to private or shared lexicons. This collection volume is competitive with traditional studio-based recording sessions and guarantees a larger diversity in occurrences, validating the efficiency of the crowdsourcing approach.

Novel Entries Among the recordings, 2,501 (83.26%) correspond to concepts that were previously absent from existing LSFB dictionaries. They primarily cover specialized vocabulary (e.g., technical or academic terms) and neologisms. Such entries would be difficult to collect with traditional approaches.

Variations The remaining 503 recordings (16.74%) offer alternative signs for existing entries, capturing regional accents, generational differences, or context-specific variations that are crucial for a comprehensive understanding of the language. Additionally, users can associate multiple tags with each sign, providing metadata that helps contextualize the usage of a given variant (e.g., regional origin or thematic domain).

Expert Curation and Taxonomy Of the collected dataset, 22.7% of the signs have been reviewed by linguistic experts in order to be included in the public LSFB lexical database. This proportion reflects the time-intensive nature of the curation process, as the experts involved are also the official maintainers and annotators of the LSFB Corpus. We expect this percentage to increase substantially as the curation effort continues.

5.2. Community Impact

Beyond the raw metrics, the qualitative reception from the LSFB community has been highly positive. Feedback collected through the platform's support channels indicates a strong sense of ownership among deaf users, who value the ability to directly influence the documentation of their language.

Educational Usage LSFB learners are also well targeted by MOSI. Usage logs suggest that those users go beyond passive lookups and use the tool as a verification mechanism, comparing their own signings with the Community contributions. This dual usage, as both a reference tool and a practice platform, highlights the role of MOSI in supporting continuous learning. Due to its diverse usage, we assume that other benefits remain unknown.

6. Conclusion

This work demonstrates the viability of sustainable, long-term Sign Language (SL) data collection by prioritizing the immediate needs of the Deaf community. By designing MOSI as a collaborative tool, we successfully flipped the traditional crowdsourcing paradigm: data collection became a natural byproduct of a tool built for daily utility. This collaborative approach proved highly effective and grounded in the principle of "nothing about us without us", a Community urge that is still not respected enough (De Meulder, 2025). It yielded thousands of new SL recordings while directly supporting autonomous learning and digital inclusion for its users.

From the community's perspective, the primary contribution is a living LSFB platform where users can both consult and enrich the available vocabulary, enabling the documentation and sharing of

a greater variety of signs, including regional and specialized variations.

From a computational perspective, this process yields new training data, particularly for signs and concepts that remain undocumented in existing corpora.

Limitations Despite these promising results, this study and the current iteration of the platform have notable limitations. First, the application focuses exclusively on isolated signs. While such data is valuable for dictionary-based search engines (Fink et al., 2023b) and training Isolated Sign Language Recognition (ISLR) models, it is insufficient for advancing more complex downstream tasks. Second, expanding future evaluations to include a larger proportion of native deaf users is necessary to draw more generalized conclusions about the community-wide impact of the tool.

Future Works Moving forward, several avenues of research and development will enhance the utility of the platform and the quality of the collected data. Linguistically, user feedback highlighted a strong need for semantic features, such as the ability to explicitly link signs (e.g., synonyms, homonyms) and attach rich metadata (e.g., regionalisms, thematic categories). Technically, we plan to extend the architecture to support the recording and processing of continuous SL sentences. Finally, to ensure long-term sustainability, we open source the codebase[10], and explore its transferability to other under-documented sign languages worldwide.

7. Acknowledgments

We would like to express special thanks to St. Marie School in Namur for their participation in the user study and their precious feedback.

This work was funded by both the Walloon region with a Ph.D. grant from FRIA (F.R.S.-FNRS) and by the SPWR under grant n°2010235 - ARIAC by DIGITALWALLONIA4.AI.

8. Bibliographical References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. 2021. BBC-Oxford British Sign Language dataset.

Katherine Atwell, Danielle Bragg, and Malihe Alikhani. 2024. Studying and mitigating biases in sign language understanding models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 268–283.

Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. 2022. Exploring collection of sign language videos through crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Kearsy Cormier, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2012. From corpus to lexical database to online dictionary: Issues in annotation of the BSL corpus and the development of BSL SignBank. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 7–12, Istanbul, Turkey. European Language Resources Association (ELRA).

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Maartje De Meulder. 2025. Deaf in AI: AI language technologies and the erosion of linguistic rights. *Language and Law/Linguagem e Direito*, 12(1).

Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024. ASL Citizen: A community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.

Shahriar Elahi Dhruvo, Mohammad Akhlaqur Rahman, Manash Kumar Mandal, Md. Istiak Hossain Shihab, A. A. Noman Ansary, Kaneez Fatema Shithi, Sanjida Khanom, Rabeya Akter, Safaeid Hossain Arib, M. N. Ansary, Sazia Mehnaz, Rezwana Sultana, Sejuti Rahman, Sayma Sultana Chowdhury, Sabbir Ahmed Chowdhury, Farig Sadeque, and Asif Sushmit. 2023. Bornil: An open-source sign language data crowdsourcing platform for ai enabled dialect-agnostic communication.

Jerome Fink, Mathieu De Coster, Joni Dambre, and Benoît Frénay. 2023a. Trends and challenges for sign language recognition with machine learning. In *ESANN 2023: 31st European Symposium*

[10] https://github.com/LSFB-Team/mot-signe

on Artificial Neural Networks, Computational Intelligence and Machine Learning 2023, pages 561–570.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jérôme Fink, Pierre Poitier, Maxime André, Loup Meurice, Benoît Frénay, Anthony Cleve, Bruno Dumas, and Laurence Meurant. 2023b. Sign Language-to-Text dictionary with lightweight transformer models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. RWTH-PHOENIX-Weather: A large vocabulary sign language recognition and translation corpus. In *LREC*, volume 9, pages 3785–3789, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4):901–911.

Mokter Hossain and Ilkka Kauranen. 2015. Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal*, 8(1):2–22.

Trevor Johnston. 2009. Creating a corpus of AUSLAN within an Australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, pages 87–95. Cascadilla Proceedings Project.

Alexander Kapitanov, Kvanchiani Karina, Alexander Nagaev, and Petrova Elizaveta. 2023. Slovo: Russian Sign Language dataset. In *International Conference on Computer Vision Systems*, pages 63–73. Springer.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines.

Laurence Meurant. 2015. Corpus LSFB: Un corpus informatisé en libre accès de vidéos et d'annotations de la langue des signes de belgique francophone (LSFB).

Jay M Patel and Jay M Patel. 2020. Introduction to common crawl datasets. *Getting structured data from the internet: running web crawlers/scrapers on a big data production scale*, pages 277–324.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Katrin Renz, Nicolaj C Stache, Samuel Albanie, and Gül Varol. 2021. Sign language segmentation with temporal convolutional networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.

Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch, and Carl Börstell. 2018. Crowdsourcing for the Swedish Sign Language dictionary. In *sign-lang@ LREC 2018*, pages 171–176, Miyazaki, Japan. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Benjamin Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. ANONYSIGN: Novel human appearance synthesis for sign language video anonymisation skeleton pose original image novel human appearances fig. 1: Novel appearance human synthesis examples generated by anonysign. In *IEEE International Conference on Automatic Face and Gesture Recognition 2021*. IEEE.

Bruno Sonnemans. 2026. Dictionnaire de LSFB en ligne. `https://dico.lsfb.be/`. Accessed: 2026-03-01.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. PopSign ASL v1.0: An Isolated American Sign Language dataset collected via smartphones. In *Advances in Neural Information Processing Systems*, volume 36, pages 184–196.

Garrett Tanzer and Biao Zhang. 2024. YouTube-SL-25: A large-scale, open-domain multilingual sign language parallel corpus.

Innocent Ye and Babacar Sow. 2023. Création d'un outil d'aide à l'apprentissage du français par la

lecture pour les élèves sourds et malentendants. Mémoire de master, Université à préciser, June. Master en informatique, spécialité Data Science.

Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE.