Learning sign language from subtitles

Hannah Bull

université PARIS-SACLAY

Learning Sign Language from Subtitles

Apprentissage de la langue des signes à partir des sous-titres

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **LISN (Université Paris-Saclay, CNRS)**, sous la direction de **Michèle GOUIFFES**, Maîtresse de Conférences, et le co-encadrement d'**Annelies BRAFFORT**, Directrice de Recherche.

Thèse soutenue à Paris-Saclay, le 27 février 2023, par

Hannah BULL

Composition du jury

Membres du jury avec voix délibérative

François YVON Directeur de Recherche LISN, CNRS, Université Paris-Saclay	Président
Mounîm EL YACOUBI Professeur Télécom SudParis, Institut Mines-Telecom, Institut Polytechnique de Paris	Rapporteur & Examinateur
Thomas HUEBER Directeur de Recherche GIPSA-Lab, CNRS, Université Grenoble Alpes	Rapporteur & Examinateur
Vincent LEPETIT Directeur de Recherche LIGM, CNRS, École des Ponts ParisTech	Examinateur

THESE DE DOCTORAT
NNT : 2023UPASG013

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Apprentissage de la langue des signes à partir des sous-titres
Mots clés: langue des signes; vision par ordinateur; mouvement du corps; vidéo texte

Résumé: Les langues des signes sont un moyen de communication essentiel pour les communautés sourdes. Elles sont des langues visuo-gestuelles, qui utilisent comme modalités les mains, les expressions faciales, le regard et les mouvements du corps. Elles ont des structures grammaticales complexes et des lexiques riches qui sont considérablement différents de ceux que l'on trouve dans les langues parlées. Les spécificités des langues des signes en termes de canaux de communication, de structure et de grammaire exigent des méthodologies distinctes.

Les performances des systèmes de traduction automatique entre des langues écrites ou parlées sont actuellement suffisantes pour de nombreux cas d'utilisation quotidienne, tels que la traduction de vidéos, de sites web, d'e-mails et de documents. En revanche, les systèmes de traduction automatique pour les langues des signes n'existent pas en dehors de cas d'utilisation très spécifiques avec un vocabulaire limité. La traduction automatique de langues des signes est un défi pour deux raisons principales. Premièrement, les langues des signes sont des langues à faibles ressources avec peu de données d'entraînement disponibles. Deuxièmement, les langues des signes sont des langues visuelles et spatiales sans forme écrite, naturellement représentées sous forme de vidéo plutôt que d'audio ou de texte.

Pour relever le premier défi, nous fournissons de grands corpus de données pour l'entraînement et l'évaluation des systèmes de traduction automatique en langue des signes, avec des contenus vidéo en langue des signes interprétée et originale, ainsi que des sous-titres écrits. Alors que les données interprétées nous permettent de collecter un grand nombre d'heures de vidéos, les vidéos originalement en langue des signes sont plus représentatives de l'utilisation de la langue des signes au sein des communautés sourdes. Les sous-titres écrits peuvent être utilisés comme supervision faible pour diverses tâches de compréhension de la langue des signes.

Pour relever le deuxième défi, cette thèse propose des méthodes permettant de mieux comprendre les vidéos en langue des signes. Alors que la segmentation des phrases est généralement triviale pour les langues écrites, la segmentation des vidéos en langue des signes en phrases repose sur la détection d'indices sémantiques et prosodiques subtils dans les vidéos. Nous utilisons des indices prosodiques pour apprendre à segmenter automatiquement une vidéo en langue des signes en unités de type phrase, déterminées par les limites des sous-titres. En développant cette méthode de segmentation, nous apprenons ensuite à aligner les sous-titres du texte sur les segments de la vidéo en langue des signes en utilisant des indices sémantiques et prosodiques, afin de créer des paires au niveau de la phrase entre la vidéo en langue des signes et le texte. Cette tâche est particulièrement importante pour les données interprétées, où les sous-titres sont généralement alignés sur l'audio et non sur la langue des signes. En utilisant ces paires vidéo-texte alignées automatiquement, nous développons et améliorons plusieurs méthodes différentes pour annoter de façon dense les signes lexicaux en interrogeant des mots dans le texte des sous-titres et en recherchant des indices visuels dans la vidéo en langue des signes pour les signes correspondants.

Title: Learning Sign Language from Subtitles
Keywords: sign language; computer vision; human pose; video text

Abstract: Sign languages are an essential means of communication for deaf communities. Sign languages are visuo-gestual languages using the modalities of hand gestures, facial expressions, gaze and body movements. They possess rich grammar structures and lexicons that differ considerably from those found among spoken languages. The uniqueness of transmission medium, structure and grammar of sign languages requires distinct methodologies.

The performance of automatic translations systems between high-resource written languages or spoken languages is currently sufficient for many daily use cases, such as translating videos, websites, emails and documents. On the other hand, automatic translation systems for sign languages do not exist outside of very specific use cases with limited vocabulary. Automatic sign language translation is challenging for two main reasons. Firstly, sign languages are low-resource languages with little available training data. Secondly, sign languages are visual-spatial languages with no written form, naturally represented as video rather than audio or text.

To tackle the first challenge, we contribute large datasets for training and evaluating automatic sign language translation systems with both interpreted and original sign language video content, as well as written text subtitles. Whilst interpreted data allows us to collect large numbers of hours of videos, original sign language video is more representative of sign language usage within deaf communities. Written subtitles can be used as weak supervision for various sign language understanding tasks.

To address the second challenge, we develop methods to better understand visual cues from sign language video. Whilst sentence segmentation is mostly trivial for written languages, segmenting sign language video into sentence-like units relies on detecting subtle semantic and prosodic cues from sign language video. We use prosodic cues to learn to automatically segment sign language video into sentence-like units, determined by subtitle boundaries. Expanding upon this segmentation method, we then learn to align text subtitles to sign language video segments using both semantic and prosodic cues, in order to create sentence-level pairs between sign language video and text. This task is particularly important for interpreted TV data, where subtitles are generally aligned to the audio and not to the signing. Using these automatically aligned video-text pairs, we develop and improve multiple different methods to densely annotate lexical signs by querying words in the subtitle text and searching for visual cues in the sign language video for the corresponding signs.

Contents

Abstract

Sign languages are an essential means of communication for deaf communities. Sign languages are visuo-gestual languages using the modalities of hand gestures, facial expressions, gaze and body movements. They possess rich grammar structures and lexicons that differ considerably from those found among spoken languages. The uniqueness of transmission medium, structure and grammar of sign languages requires distinct methodologies.

The performance of automatic translations systems between high-resource written languages or spoken languages is currently sufficient for many daily use cases, such as translating videos, websites, emails and documents. On the other hand, automatic translation systems for sign languages do not exist outside of very specific use cases with limited vocabulary. Automatic sign language translation is challenging for two main reasons. Firstly, sign languages are low-resource languages with little available training data. Secondly, sign languages are visual-spatial languages with no written form, naturally represented as video rather than audio or text.

To tackle the first challenge, we contribute large datasets for training and evaluating automatic sign language translation systems with both interpreted and original sign language video content, as well as written text subtitles. Whilst interpreted data allows us to collect large numbers of hours of videos, original sign language video is more representative of sign language usage within deaf communities. Written subtitles can be used as weak supervision for various sign language understanding tasks.

To address the second challenge, we develop methods to better understand visual cues from sign language video. Whilst sentence segmentation is mostly trivial for written languages, segmenting sign language video into sentence-like units relies on detecting subtle semantic and prosodic cues from sign language video. We use prosodic cues to learn to automatically segment sign language video into sentence-like units, determined by subtitle boundaries. Expanding upon this segmentation method, we then learn to align text subtitles to sign language video segments using both semantic and prosodic cues, in order to create sentence-level pairs between sign language video and text. This task is particularly important for interpreted TV data, where subtitles are generally aligned to the audio and not to the signing. Using these automatically aligned video-text pairs, we develop and improve multiple different methods to densely annotate lexical signs by querying words in the subtitle text and searching for visual cues in the sign language video for the corresponding signs.

Our contributions are as follows: (i) a 2D-skeleton-keypoint French Sign Language dataset with written French translations, a unique corpus with original sign language content produced outside of a laboratory context

(ii) baseline results for the new task of sentence-like segmentation of sign kanguage video on the aforementioned dataset (iii) a dataset of British Sign Language interpreted videos with English subtitles, the largest corpus currently available for sign language research (iv) a state-of-the-art method for aligning subtitles to sign language video for interpreted data (v) new and improved methods for weakly-supervised dense annotation of lexical signs and (vi) an extended version of our French Sign Language dataset including RGB videos for future research perspectives.

Résumé

Les langues des signes sont un moyen de communication essentiel pour les communautés sourdes. Elles sont des langues visuo-gestuelles, qui utilisent comme modalités les mains, les expressions faciales, le regard et les mouvements du corps. Elles ont des structures grammaticales complexes et des lexiques riches qui sont considérablement différents de ceux que l'on trouve dans les langues parlées. Les spécificités des langues des signes en termes de canaux de communication, de structure et de grammaire exigent des méthodologies distinctes.

Les performances des systèmes de traduction automatique entre des langues écrites ou parlées sont actuellement suffisantes pour de nombreux cas d'utilisation quotidienne, tels que la traduction de vidéos, de sites web, d'e-mails et de documents. En revanche, les systèmes de traduction automatique pour les langues des signes n'existent pas en dehors de cas d'utilisation très spécifiques avec un vocabulaire limité. La traduction automatique de langues des signes est un défi pour deux raisons principales. Premièrement, les langues des signes sont des langues à faibles ressources avec peu de données d'entraînement disponibles. Deuxièmement, les langues des signes sont des langues visuelles et spatiales sans forme écrite, naturellement représentées sous forme de vidéo plutôt que d'audio ou de texte.

Pour relever le premier défi, nous fournissons de grands corpus de données pour l'entraînement et l'évaluation des systèmes de traduction automatique en langue des signes, avec des contenus vidéo en langue des signes interprétée et originale, ainsi que des sous-titres écrits. Alors que les données interprétées nous permettent de collecter un grand nombre d'heures de vidéos, les vidéos originalement en langue des signes sont plus représentatives de l'utilisation de la langue des signes au sein des communautés sourdes. Les sous-titres écrits peuvent être utilisés comme supervision faible pour diverses tâches de compréhension de la langue des signes.

Pour relever le deuxième défi, cette thèse propose des méthodes permettant de mieux comprendre les vidéos en langue des signes. Alors que la segmentation des phrases est généralement triviale pour les langues écrites, la segmentation des vidéos en langue des signes en phrases repose sur la détection d'indices sémantiques et prosodiques subtils dans les vidéos. Nous utilisons des indices prosodiques pour apprendre à segmenter automatiquement une vidéo en langue des signes en unités de type phrase, déterminées par les limites des sous-titres. En développant cette méthode de segmentation, nous apprenons ensuite à aligner les sous-titres du texte sur les segments de la vidéo en langue des signes en utilisant des indices sémantiques et prosodiques, afin de créer des paires au niveau de la phrase entre la vidéo en langue des signes et le texte. Cette tâche est particulièrement importante

pour les données interprétées, où les sous-titres sont généralement alignés sur l'audio et non sur la langue des signes. En utilisant ces paires vidéo-texte alignées automatiquement, nous développons et améliorons plusieurs méthodes différentes pour annoter de façon dense les signes lexicaux en interrogeant des mots dans le texte des sous-titres et en recherchant des indices visuels dans la vidéo en langue des signes pour les signes correspondants.

Nos contributions sont les suivantes : (i) un corpus de points-clés de squelettes en 2D de la langue des signes française avec des traductions écrites en français, un corpus rare avec contenu original en langue des signes produit en dehors de contexte d'un laboratoire (ii) des résultats de référence pour la nouvelle tâche de segmentation en phrases des vidéos en langue des signes sur le corpus susmentionné (iii) un corpus de vidéos interprétées en langue des signes britannique avec des sous-titres anglais, le plus grand corpus actuellement disponible pour la recherche en langue des signes (iv) une méthode de pointe pour aligner les sous-titres à la langue des signes pour les données interprétées (v) des méthodes améliorées et de nouvelles méthodes faiblement supervisées pour l'annotation dense des signes lexicaux et (vi) une version étendue de notre corpus en langue des signes française incluant des vidéos RGB pour les perspectives de recherche futures.

Acknowledgements

I would first and foremost like to thank Michèle Gouiffès and Annelies Braffort, for introducing me to the complexities of automatic sign language understanding, for the support and the freedom to pursue many different directions, and also for the opportunities to learn LSF from the fantastic teachers at Visuel LSF. I would especially like to thank Gül Varol for the lively collaborations during the quiet periods of covid lockdowns and beyond. I have had the chance to work closely with many incredible people, in particular, with Daffy Afouras, Lili Momeni, Sam Albanie, Prajwal K R, Coline Petit-Jean, Ambroise Mopendza, Théo Cheynel, Yanis Ouakrim and Andrew Zisserman. I would also like to thank the jury members Mounîm El Yacoubi, Thomas Hueber, Vincent Lepetit, François Yvon, for taking the time to provide feedback on my research. Throughout my studies, I have benefited greatly from the support of my family, even from afar. I am also grateful for all the wonderful moments spent with friends, and above all with Igor.

The work in this thesis has been partially funded by the ROSETTA project, financed by the French Public Investment Bank (Bpifrance). We thank *Média-Pi!* for the data used in Chapter 2, Chapter 3 and Chapter 7, as well as for the useful discussions contributing to these chapters. The work in Chapter 4, Chapter 5 and Chapter 6 was supported by EPSRC grant ExTol, and a Royal Society Research Professorship. The work in Chapter 6 was additionally supported by ANR project CorVis ANR-21-CE23-0003-01. The dataset in Chapter 4 was made possible due to the support of Red Bee Media Ltd. and their BSL interpreters, the assistance of Andrew Brown in preparing identity embeddings, Abhishek Dutta and his tireless support of the VIA annotation tool, the support of Ashish Thandavan, David Miguel Susano Pinto and Ivan Johnson towards the dataset release, as well as Necati Cihan Camgöz and his suggestions on dataset distribution. We thank Tom Monnier, Himel Chowdhury, Abhishek Dutta, Ashish Thandavan for helping in various ways towards Chapter 5, as well as Sagar Vaze for helping towards Chapter 6.

1 - Introduction

This thesis explores various aspects of automatic sign language understanding from sign language videos with written subtitles, including segmentation of sign language into sentence-like units using skeleton keypoint data (Part I), video-text alignment of sentences and dense annotation of lexical signs using interpreted sign language videos (Part II) and creation of a new corpus with non-interpreted sign language videos (Part III).

In Sec. 1.1, we provide a brief description of some of the characteristics of sign languages. We outline the goals of this thesis in Sec. 1.2, the motivations for these goals in Sec. 1.3 and the challenges in Sec. 1.4. We provide an overview of the history of sign language recognition and translation in Sec. 1.5 to provide context on our work on sign language understanding. Our article, software and dataset contributions are listed in Sec. 1.6 and an outline of this thesis is presented in Sec. 1.7.

1.1 . Sign Languages

Sign languages are used by millions of people around the world and are an important means of communication for deaf communities. They are visual-gestural languages, using the modalities of hand gestures, facial expressions, gaze and body movements (see Fig.1.1). The complexity of sign languages is the same as that of spoken languages [180]. However, sign languages have rich grammar structures and lexicons that differ considerably from those found among spoken languages [180]. Sign languages are oral languages, there are no standard written forms of sign languages, and the natural form of recording sign languages is through video.

Sign language is not universal; there are an estimated 144 different sign languages used globally [62]. *LSF* is the acronym for French Sign Language or la Langue des Signes Française, and *BSL* is the acroynm for British Sign Language. One universal characteristic across sign languages is the strong presence of iconicity [159]. Forms can be naturally depicted using gestures, and thus there is a strong connection between form and meaning in signed languages that is less present in vocal languages [85].

Technologies for sign languages lag behind those for written and spoken languages. Search engines, automatic translation tools and even dictionaries are at primitive stages for sign languages compared to the resources available for many common written and spoken languages. We aim to address this imbalance by contributing to more inclusive technologies.

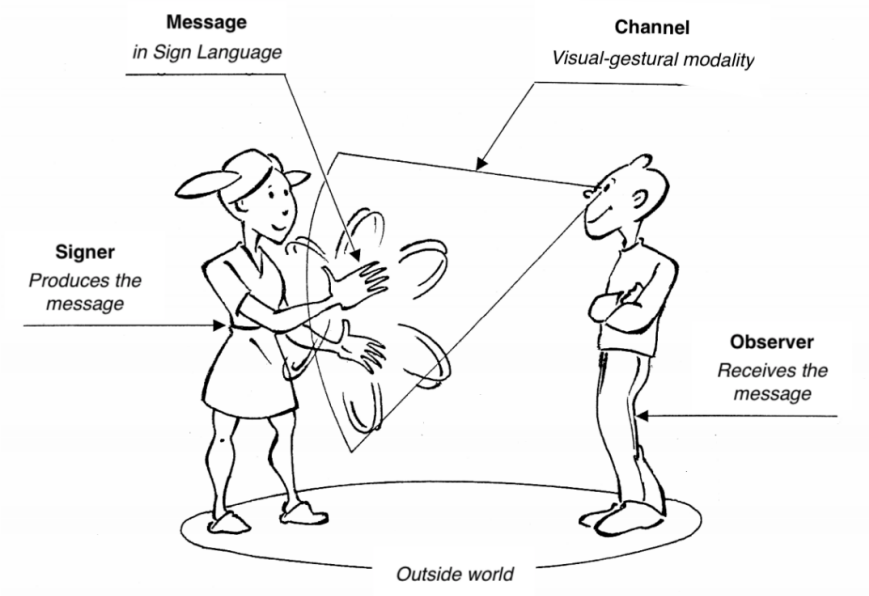


Figure 1.1: Illustration from [84] on the modalities of sign language communication

1.2 . Goals

The goal of this thesis is to create corpora with both sign language and written subtitles, and to use information from the subtitles to better understand the signing. These signing and subtitle sentence-like pairs form a parallel corpus, similar to large parallel bilingual corpora which have been successful in achieving translation systems for written languages [108, 217, 156].

We aim to create large corpora of sign language and written subtitles which can be used by academic researchers to train and evaluate automatic sign language understanding models. Although such sign language corpora will never be fully representative of how the deaf community uses sign language in daily life, we aim to release corpora with a large, open vocabulary such that models can learn a wide variety of expressions.

A long-term goal is to achieve translation systems of practical value for sign languages to written languages. There are numerous intermediate steps in order to achieve accurate translation sign language to written language, such as automatic annotation of sentence-level pairs in sign language and written language. This thesis explores three such goals. Our first goal is to understand and detect sentence-like boundaries in sign language. Secondly, we aim to learn to align written sentences to sign language segments. The third goal is to use the aligned sentence text to densely annotate lexical signs.

1.3 . Motivations

The World Federation of the Deaf estimates that there are around 70 million deaf individuals world-wide using hundreds of different sign languages.[1] Yet, technologies for sign languages lag far behind technologies for spoken and written languages. Home assistance tools such as Alexa, Siri and Google Assistant use spoken and not signed language [202]. Automatic translation tools are sufficiently accurate to allow cross-lingual communication across spoken and written languages, such as automatic subtitling and translation of videos or automatic translation of documents. Performance of automatic translation from sign language to written language is very poor and has currently has limited practical applications [110].

One of the key reasons for the lack of sign language technologies is the quantity and quality of datasets available for training [20]. A large source of sign language video comes from news sources or other journalistic content in sign language. Fig. 1.2 shows examples of TV shows with sign language interpretation. Such data sources are valuable, as they are a good way to obtain many hours of identical content in both spoken/written language and sign language. However, due to the nature of interpretation, there is source language interference in the signing and this data may not be representative of original sign language content. Fig. 1.3 shows examples of information content originally produced in sign language with a written translation in the form of subtitles. While interpretation is delivered in real time, translation allows for preparation and corrections. If there is no audio track, then the subtitles are likely to be well aligned to the signing. This is the case for *Médiapi* [2] and *The Daily Moth*. [3] These data sources create many new opportunities for sign language technologies, due to the large quantity of video content and the presence of written language subtitles.

Although automatic translation tools for sign language would facilitate production of sign language content and associated written content, this is a difficult problem which may take many more years of research to solve. Nevertheless, there are many other technologies that could already be made or improved using the work in this thesis.

For example, segmenting sign language into sentence-like units can be useful for an assistive subtitling tool, which plays back short segments to be translated into written subtitles by the user. This kind of assistive subtitling tool could be also be improved by automatic video-text alignment, where the user writes a translation, which is then automatically aligned to the signing in the video in the form of subtitles. A second application is to create bilingual written-signed corpora aligned at a sentence or phrase-like level. Such cor-

[1] http://wfdeaf.org/our-work/

[2] https://media-pi.fr/

[3] https://www.dailymoth.com/blog

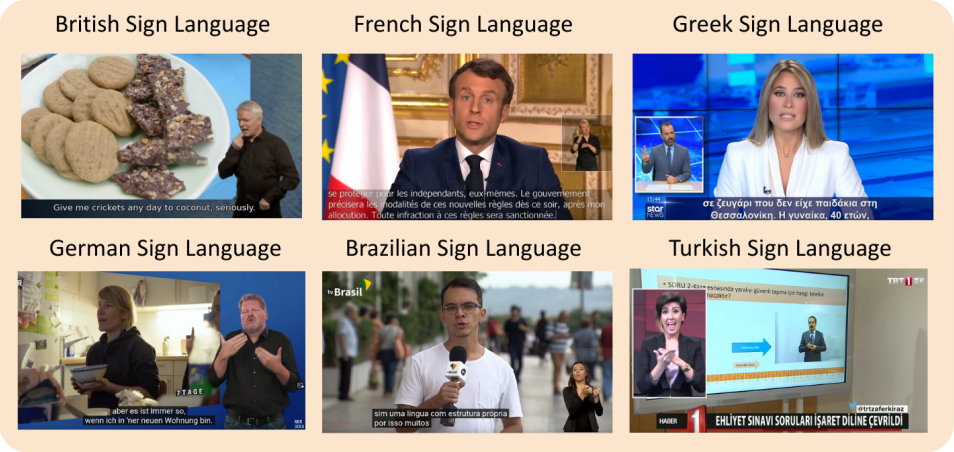


Figure 1.2: There are many sources of interpreted TV data in multiple sign languages. This data source generally comes with subtitles which are aligned to the audio track and not to the signing.

Figure 1.3: There are multiple news and information sources which are originally produced in sign language. These productions may also use audio or not, and they may contain translated subtitles in a written language to improve accessibility.

pora can be used in contextual or concordance dictionaries, useful for translation or for language learning [103]. An illustration of a bilingual concordancer for sign language is provided in Fig. 1.4.

Dense annotation can be used in multiple applications. Annotating sign language video is a time consuming task, and dense annotation methods can allow researchers to annotate more data in a shorter period of time. Search engines for sign language could use dense annotation methods to index keywords (lexical signs) or to cluster videos with similar content. Sign language learners could use automatic dense annotation methods to assist with comprehension of sign language video.

Many technologies initially created to improve accessibility for the deaf also prove useful in other applications. For example, subtitles are increasing

Figure 1.4: Illustration of a possible bilingual concordancer for French and French Sign Language. Users can search for phrases in either written or even signed language, and search results appear from the dataset.

ubiquitous and are preferred and used by a majority of young people.[4] Similarly, technologies for sign language can extend far beyond sign language applications. Improvements in sign language understanding could transfer to other areas, including human pose recognition [33], in particular the difficult problem of hand shape recognition [112, 147], action recognition on common datasets such as Kinetics [35], and automatic avatar generation, for example for video games [91, 70].

1.4 . Challenges

We discuss three main challenges for learning sign language from subtitles: particularities of sign language grammar (Sec. 1.4.1), lack of sufficient data (Sec. 1.4.2), and limitations in computer vision and natural language processing techniques (Sec. 1.4.3).

1.4.1 . Particularities of Sign Language

Research on automatic sign language understanding should take into account differences between sign languages and spoken or written languages. The grammar of sign languages differs from that of spoken or written languages [180]. Sign languages have a strong connection between form and meaning referred to as *iconicity*, they are *multilinear* and cannot be considered

[4] https://www.bbc.com/news/entertainment-arts-59259964

a linear sequence of gestural units, and they use a three-dimensional *signing space*[72, 9]. Due to these differences, methods which perform well for spoken languages will not necessarily work well for signed languages.

Signers generate or adapt signs based on context, creating infinitely many iconic structures that cannot be listed in a lexical dictionary. For example, the sign for a ball can be modified based on the size, position, texture and motion of the ball. There is no simple correspondence between 'signs' and 'words'; many signs are *non-lexical*, without a direct translation into a written language. Fully lexical signs are only part of sign language discourse. Partially lexical signs are iconic elements such as pointing signs, depicting signs and fragment buoys [9]. Partially lexical signs include signs such as spacial referents, motion, size and shape of objects [19]. These signs may be highly dependent on context. For written languages, a native speaker may use a vocabulary in the order of tens of thousands of words, and so a language model 'only' needs to understand to manipulate this list of words. For sign languages, a model must learn how signs can be created or modified in potentially infinite ways depending on context.

Signers can convey multiple concepts simultaneously by exploiting multiple articulators, including both hands, facial expressions, gaze and body movements. Sign languages cannot be considered as a linear sequences of gestural units, as information can be communicated using different modalities. For example, the non-dominant hand may represent a static object in a scene, and the dominant hand indicates the action or relation between a person and that static object. A facial expression may indicate an attitude, emotion or adjective. Fig. 1.5 shows an example of gloss annotations from the Dicta-Sign-LSF-v2 corpus [9]. There are numerous cases of multiple simultaneous gloss annotations.

Sequence-to-sequence models are commonly used to translate a linear sequence of words or tokens from one language into a linear sequence of words or tokens from another language [178, 188]. As sign languages are not a linear sequence of signs, perhaps such models are ill-adapted to sign language translation. Nevertheless, we note that sign language video is a sequence of video frames, and so at some level, it can be considered a linear sequence. However, individual frames are not semantic units, unlike words.

A discourse in sign language is not a linear sequence of signs, rather is structured in a signing space. The three-dimensionality of the signing space is also used to arrange elements of a sign language discourse including people, objects, places and temporal events [72]. Sign language video is a two-dimensional representation of a three-dimensional world. Multi-view cameras or other depth perception techniques could potentially assist in automatic sign language understanding through better perception of the three-dimensional signing space.

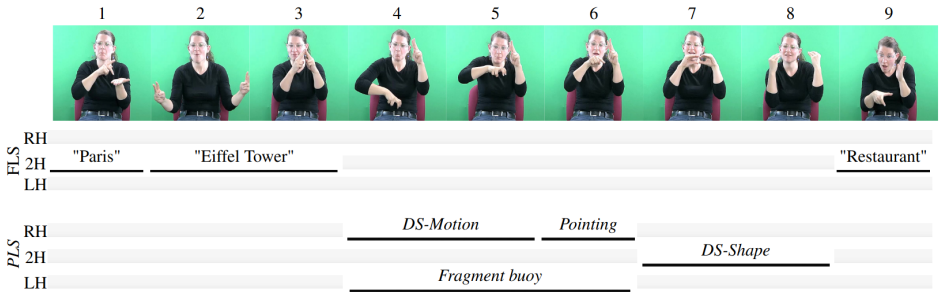


Figure 1.5: Dicta-Sign-LSF-v2 simultaneous gloss annotations [9]: There may be multiple simultaneous gloss annotations when different articulators are used to convey multiple concepts. RH, 2H and LH refer to right-handed, two-handed and left-handed signs resprectively. FLS and PLS refer to fully and partially lexical signs. Fully lexical signs can be glossed into a word in written language. Some non-manual elements have not been annotated, for example mouthings and facial expressions. Translation: *In Paris, if you climb the Eiffel Tower, you will find a square-shaped restaurant at the middle floor.*

1.4.2 . Lack of Data

Sign languages are minority languages with relatively little content in comparison to many spoken and written languages. In [20], the authors discuss the limitations of existing sign language corpora for sign language research. In addition to limited possibilities to acquire large numbers of hours of sign language content for research purposes, there is also the difficulty of acquiring relevant annotations to train models for sign language understanding.

Annotation is a time consuming task requiring fluent signers with a strong understanding of sign language grammar. In Sec. 1.4.1, we mentioned the usage of lexical signs and partially lexical signs in sign language discourse. There are many structures in sign language discourse that can be annotated in various levels of detail. The Dicta-Sign-LSF-v2 corpus [9] is annotated with lexical signs, partially lexical signs (pointing signs, depicting signs and fragment buoys), as well as non-lexical signs (fingerspelling, numbers and gestures). This choice of annotation is based on the categories proposed in [98], but there is no standard annotation method.

One way of avoiding high annotation costs is to use weak annotations from subtitle texts. Where possible, we limit the use of human annotators and use subtitles as weak annotations in this thesis.

1.4.3 . Technical Limitations

Limitations in computer vision and natural language processing techniques also limit sign language understanding. More general computer vision techniques are not yet capable of some tasks which would advance

research in sign language understanding. For example, hand pose detection methods currently do not have sufficiently high performance in order to reliably detect the subtle differences in hand shape when signing [162]. There is also a large margin of improvement for 3D human pose detection [198].

The performance of neural machine translation models is highly correlated with the amount of training data [83]. Improve techniques in natural language processing for low resource languages are vital for sign language understanding.

1.5 . A Brief History of Sign Language Recognition and Translation

The main focus of this thesis is learning sign language from subtitle texts, which builds upon an extensive body of research on sign language recognition and translation tasks. We refer to [110] for a more detailed quantitative survey of sign language recognition and translation.

The first example of an attempt to automatically recognise signs is a patent from 1983 [82] for a glove equipped with sensors and capable of recognising the single-hand hand shapes of the alphabet and numbers of American Sign Language. In an article from 1991, Murakami et al. [141] use a data glove to capture data from the alphabet in Japanese Sign Language, as well as 10 lexical signs, and train a neural network to recognise these hand gestures. In 1993, Fels and Hinton [69] train neural networks to recognise 203 gesture classes in the context of a hand-gesture to speech system also using electronic gloves. Braffort (1996) [17] emphasises the fact that sign language is much more than just hand shapes, and propose an architecture to take into account spatial information and spatial relations in order to recognise non-lexical constructs.

Nevertheless, early work in sign language recognition has a central focus on hand shape and hand movement, in particular due to the prominence of electronic and colour gloves to facilitate hand shape recognition. As a step towards recognising hand shape directly from images, colour gloves were used in a number of works between 1995 and 2005, such as Lu et al. (1997) [127] and Deng et al. (2002) [53]. Early examples of sign language recognition from images and videos are Tamura (1988) [181], who train a model to recognise signs from images, and Starner (1995) [175], who use a single colour camera to recognise signs. However, both works focus on hand tracking rather than information from the body pose and facial expressions.

Although electronic gloves and colour gloves may have been useful intermediate steps for developing methods in computer vision for sign language understanding, intrusive methods requiring the signer to wear particular equipment have limited practical purposes. The deaf community has ex-

tensively criticised the proposed everyday uses of signing-gloves.[5] Since 2005, image and video-based methods have dominated over glove-based or other intrusive methods. Cooper and Bowden [47] classify a vocabulary of 164 signs from video. Using Arabic Sign Language videos with simple phrases and limited vocabulary, Assaleh et al. [6] train a hidden Markov model for continuous sign language recognition.

Prior to 2015, most works on sign language recognition use a vocabulary of a few hundred or fewer signs. Some early attempts to recognise signs from a large 5k vocabulary include Ma et al. (2000) [128] and Wang et al. (2002) [196]. The PHOENIX14 [75] and PHOENIX14T [111] datasets became by far the most popular benchmark datasets for sign language recognition and translation after 2015, and have a vocabulary of around 1k sign glosses. These datasets contains weather reports with a very specific vocabulary and simple sentence structure.

The current recognition and translation performances on the PHOENIX14 [75] and PHOENIX14T [111] datasets are reasonbly high, albeit significantly lower in comparison to transcription and translation of high resource written and spoken languages. For example, Hu et al. (2022) [92] achieve state-of-the-art results on continuous sign language recognition and obtain a word error rate of 21 on the test set of PHOENIX14 [111]. For translation, Camgoz et al. (2020) [31] achieve a word error rate of 26 on PHOENIX14T [111]. Attempts at sign language translation on open vocabulary sign language data currently achieve catastrophically low results. In the very recent challenge WMT-SLT22 (2022) [140], the best performing model achieved a human evaluation score of 4 out of 100.

Automatic sign language understanding has largely focused on the recognition of lexical signs in continuous sign language video and isolated sign language video, with little work on recognition of non-lexical signs [10]. In Braffort (2001) [18], ethical concerns are raised regarding misunderstandings about sign language and overemphasis on lexical signs, a concern also raised in Bragg et al. (2019) [20].

The work in this thesis attempts to build upon this base of sign language recognition and translation literature. We use non-invasive methods for sign language understanding (videos) and large-vocabulary datasets with continuous sign language. We look at alternative tasks such as semantic segmentation and video-text alignment, and not just lexical sign recognition. Although I would have liked to be able to develop a sign language translation system with somewhat reasonable performance for this thesis, this remains a goal for future research.

[5] https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/

1.6 . Contributions

We list the publications (Sec. 1.6.1), software (Sec. 1.6.2) and datasets (Sec. 1.6.3) produced as part of this work. Parts of this thesis were produced at LISN (CNRS) at the Université Paris-Saclay, and other parts of this thesis were produced in collaboration with other researchers from LIGM (CNRS), Ecole des Ponts, Université Gustav Eiffel and VGG, University of Oxford.

1.6.1 . Publications

The work done during this PhD led to the following publications (* denotes equal contribution):

- Bull, H., Braffort, A. & Gouiffès, M. (2020). MEDIAPI-SKEL - A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles. In LREC. [24] (Chapter 2)

  For this publication, I ran all of the analyses and wrote the majority of the text. My supervisors Annelies Braffort and Michèle Gouiffès sourced this dataset and greatly contributed to the ideas behind this paper.

- Bull, H., Gouiffès, M. & Braffort, A. (2020). Automatic Segmentation of Sign Language into Subtitle-Units. In ECCV Workshop Proceedings (SLRTP Best Paper Award). [25] (Chapter 3)

  For this publication, I ran all of the analyses and wrote the majority of the text. My supervisors Annelies Braffort and Michèle Gouiffès greatly contributed to the ideas behind this paper.

- Albanie, S.*, Varol, G.*, Momeni, L.*, Bull, H.*, Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., Zisserman, A. (2021). BBC-Oxford British Sign Language Dataset. [4] (Chapter 4)

  The release of BOBSL is a very large project involving many people. For this publication, I contributed to the pre-processing of the BOBSL data by automatically aligning the subtitles to the audio track due to alignment errors in around one quarter of videos. I also contributed to the all of the parts of the article on alignment of subtitles to sign language video, as well as making the BOBSL challenges available online on the CodaLab platform.[6]

- Bull, H.*[7], Afouras, T*., Varol, G., Albanie, S., Momeni, L. & Zisserman, A. (2021). Aligning Subtitles in Sign Language Video. In ICCV. [26] (Chapter 5)

[6] https://codalab.lisn.upsaclay.fr/competitions/6790

[7] * denotes equal contribution

For this publication, all of the experiments were run by Triandaffolys Afouras and myself. The writing and ideas were shared by all of the listed authors.

- Momeni, L.*, Bull, H.*, Prajwal, K R*, Albanie, S., Varol, G. Zisserman, A. (2022). Automatic dense annotation of large-vocabulary sign language videos. In ECCV. [136] (Chapter 6)

  For this publication, my main contributions were on using synonyms, aligned subtitles and exemplars to increase the yield of dense annotations. The writing and ideas were shared by all of the listed authors.

- Bull, H.*, Ouakrim, Y.*, Braffort, A. & Gouiffès, M. (2022). Mediapi-RGB - A Video Database of French Sign Language With Aligned French Subtitles. Pre-print. (Chapter 7)

  For this publication, I ran most of the analyses and wrote the majority of the text. My supervisors Annelies Braffort and Michèle Gouiffès sourced this dataset. The ideas were shared by all of the listed authors.

1.6.2 . Software

The following software was developed to aid this research:

- Heuristic method to track the main signers in a video with multiple people and potentially multiple signers (e.g. signers in the background): https://github.com/hannahbull/clean_op_data_sl
- Code to reproduce results and run model used in [25] (Chapter 3): https://github.com/hannahbull/sign_language_segmentation
- Code to reproduce results and run model used in [26] (Chapter 5): https://github.com/hannahbull/subtitle_align

1.6.3 . Datasets

Figure 1.6: Logos for the three released datasets: MEDIAPI-SKEL, BOBSL and Mediapi-RGB

We release the following datasets for academic research purposes:

- **MEDIAPI-SKEL**, available at https://www.ortolang.fr/market/corpora/mediapi-skel. See Chapter 2 or [24] for more details. This dataset contains 2D OpenPose [33] keypoints for 27 hours of original journalistic content in LSF with French subtitles aligned to the signing.

- **BOBSL**, available at https://www.robots.ox.ac.uk/~vgg/data/bobsl/. See Chapter 4 or [4] for more details. The BOBSL dataset contains around 1400 hours of BSL-interpreted TV programmes from the BBC with written English subtitles. The subtitles are aligned to the audio track and not to the signing, and so there is a variable lag between the subtitles and the signing.

- **Mediapi-RGB**, available at https://www.ortolang.fr/market/corpora/mediapi-rgb. See Chapter 7 for more details. Mediapi-RGB contains 86 hours of original sign language in LSF with French subtitles aligned to the audio. The source data for this corpus is identical to that of MEDIAPI-SKEL, however, we are able to release the original RGB videos as well as derivative products (such as 2D OpenPose [33] keypoints).

1.7 . Outline

This thesis is structured in three parts, corresponding to the three different corpora we make available in this work. The common characteristic of the three corpora is that they all contain sign language video with associated written subtitles. The first corpus contains 2D-skeleton keypoints of original LSF content, the second corpus contains RGB video of interpreted BSL content and the third corpus contains RGB video of original LSF content. The differences between these three corpora make them better adapted for certain goals in sign language understanding from subtitles.

Part I presents the MEDIAPI-SKEL corpus in Chapter 2, as well as three potential challenges for this dataset: semantic segmentation, sign language sentence alignment and video-text embeddings. Baseline results for semantic segmentation on MEDIAPI-SKEL are provided in Chapter 3. Due to the limitations of MEDIAPI-SKEL, we attempt the challenge of sign language sentence alignment on another dataset in Part II.

We begin Part II by introducing the BOBSL dataset in Chapter 4, a corpus containing 1400 hours of interpreted BSL video with written English subtitles. Chapter 5 improves upon our semantic segmentation method, and we train a model using both semantic and prosodic cues to simultaneous segment sign language video and align the segments to text subtitles. The subtitles in BOBSL are not aligned to the signing, and so we can use the method developed in Chapter 5 to automatically create video-text sentence pairs.

Chapter 6 exploits this automatic alignment by querying words in the aligned subtitle text to densely annotate lexical signs in sign language video.

Finally, Part III introduces Mediapi-RGB, an improved version of MEDIAPI-SKEL, with more hours and original sign language video, not just skeleton keypoints. This corpus contains journalistic content originally in LSF, rather than interpreted content as in BOBSL, and is thus more representative of spontaneous sign language usage by native deaf signers. We discuss the characteristics and opportunities of this new corpus in Chapter 7.

Part I

Using 2D Skeleton Keypoints for Sentence-Like Segmentation of Sign Language

2 - Mediapi-Skel: A Skeleton-Based Sign Language Dataset

This chapter presents MEDIAPI-SKEL, a 2D-skeleton database of French Sign Language videos aligned with French subtitles. The corpus contains 27 hours of video of body, face and hand keypoints, aligned to subtitles with a vocabulary size of 17k tokens. In contrast to existing sign language corpora such as videos produced under laboratory conditions or interpretations of TV programs into sign language, this dataset is constructed using original sign language content largely produced by deaf journalists at the media association *Média-Pi!*.[1] Moreover, the videos are manually synchronised with French subtitles. We propose three challenges appropriate for this corpus that are related to processing units of signs in context: semantic segmentation of sign language, automatic alignment of text and video, and production of video-text embeddings for cross-modal retrieval. These challenges deviate from the classic task of identifying a limited number of lexical signs in a video stream.

The work in this chapter resulted in the publication [24]. I contributed to the majority of the writing and analyses in this chapter. The ideas behind this work come from all authors: Annelies Braffort, Michèle Gouiffès and myself. Annelies Braffort and Michèle Gouiffès sourced this dataset through an agreement with *Média-Pi!*.

2.1 . Introduction

There is a relative lack of sign language corpora in comparison to other areas of natural language processing, particularly of large, diverse sign language corpora with high quality native speakers in natural settings. Moreover, much attention in the computer vision literature has been given to automatic detection of a limited number of signs, in comparison to other sign language processing tasks [20]. In order to combat these two shortcomings, we propose a new dataset for new challenges.

We provide a new corpus available for public research called MEDIAPI-SKEL[2]. Our dataset consists of 368 videos totaling 27 hours of French Sign Language (LSF) with French subtitles, generated from the content of the bilingual LSF-French media company *Média-Pi!*. The videos are provided in the form of 2D-skeletons with 135 face, hand and body keypoints, but the original videos can be accessed through a subscription with *Média-Pi!*. The subtitles provide an accurate alignment between short segments of text and short seg-

[1] https://media-pi.fr/

[2] ortolang.fr/market/corpora/mediapi-skel

ments of sign language video. A frame of this data is shown in Figure 2.1.

This new corpus allows for challenges for sign language processing at a 'sentence' or 'phrase' level, rather than at the 'word' or 'sign' level. We propose three such machine learning challenges for MEDIAPI-SKEL.

The structure of the chapter is as follows. Firstly, we discuss differences between MEDIAPI-SKEL and other existing sign language corpora. Secondly, we justify our particular focus on 2D-skeleton data. Thirdly, we provide information relating to the production and content of the corpus. Finally, we present three data challenges appropriate for MEDIAPI-SKEL.

Figure 2.1: Frame from MEDIAPI-SKEL

2.2 . Comparison with Existing Corpora

MEDIAPI-SKEL is distinct from existing sign language corpora in multiple aspects.

Firstly, MEDIAPI-SKEL is a large sign language corpus predominantly produced by deaf journalists. The quantity and quality of original and natural content produced by deaf participants in MEDIAPI-SKEL is difficult to find outside of laboratory-produced corpora. The British Sign Language Corpus [164] is one such linguistic corpus created under laboratory conditions, that contains videos of narratives invented by the participants. The DictaSign corpus [9] contains dialogues in LSF between participants. These corpora are produced in a standard format, with consistent camera angles and uniform background conditions. Such corpora are expensive to acquire, translate and annotate; but conditions can be better controlled. On the other hand, the diversity of

scenarios and camera angles in MEDIAPI-SKEL better reflects the diversity of real-world sign language videos.

Secondly, the corpus is not produced by real-time translation of written or spoken text. This is distinct from corpora such as RWTH-PHOENIX-Weather [74] and the BBC TV corpus [151], which are acquired from sign language translations of TV programs. Sign language during real-time interpretation tends to closely follow the grammatical structure of the spoken language due to strong time constraints [116]. The spontaneous LSF used in our corpus follows a more natural grammatical structure, and it is the text in the subtitles that is adapted accordingly to align to the LSF.

Thirdly, the alignment between subtitles and video is accurate. Some laboratory-produced corpora contain aligned written translations of sign language, such as the Belgian French Sign Language corpus [129] . This is generally not the case for live interpretations, where the subtitles will be aligned to the speech and the sign language translation appears with a time lag. In the case of RWTH-PHOENIX-Weather, the subtitles are manually realigned to match video segments. In the case of the BBC TV corpus, the subtitles are not aligned to the sign language video.

Finally, we provide 2D-skeleton data for all the videos, rather than the original data. This allows us to publish data without negative impact on the economic model of *Média-Pi!*, which relies on offering exclusive content to subscribers. We include hand shapes, body pose and facial keypoints in order to best conserve the the intelligibility of the sign language. Skeleton-based representations are used in a significant number of contributions in sign language processing and offer several advantages, discussed in Section 2.3.

2.3 . Skeleton-Based Models

2D-skeleton keypoints of the face, hands and body are sufficient to maintain relatively high intelligibility in sign language videos. In [107], the authors use these keypoints to automatically translate a limited range of sentences in Korean Sign Language. In [185], signers discuss in American Sign Language using 27 hand and face keypoints.

The drastic data reduction of information by using skeleton-based models compared to the original videos, should lead to lighter and faster models, with fewer parameters to train. Moreover, external validity of models is more readily attainable, as sign language processing becomes independent of the background and appearance of the signer. Skeleton data can be normalised such that each person has the same body proportions, which removes some of the variation irrelevant to sign language processing.

Skeleton data has proved valuable in action recognition tasks. In [206], the authors demonstrate that the performance of a 2D-skeleton model is capable

of achieving a similar accuracy to models using RGB or optical flow data on action classes strongly related with body motion. The performance of skeleton models is lower for human actions in interaction with the environment. However, unlike actions such as 'playing football', sign language does not involve interaction with external objects, and so skeleton data is particularly appropriate for our case.

Finally, another key area in sign language processing is sign language production using avatars. Motion capture is highly successful in creating realistic animations. Body keypoints are captured from an actor and then transferred onto an animated figure. For example, face, body and hand keypoints can be used to animate avatars signing intelligible isolated signs [5, 190, 162]. Skeleton models can contribute to this area of research in order to create natural-looking signing avatars.

2.4 . Presentation of Corpus

To constitute this corpus, we use 368 subtitled videos totaling 27 hours of LSF footage and written French produced by *Média'Pi!*. The content is in the journalistic genre, providing information on a wide variety of events of public interest.

The mode of production varies depending on the subject matter. For example, news stories of national and international interest are generally presented by one journalist, where factual elements are assembled from written press releases. Coverage of local Deaf-related events may involve discussions and interviews with multiple people at the scene.

In a handful of videos, interviews are conducted with people who use an oral language or a foreign sign language, and these interviews are translated or interpreted into LSF. Both the interviewee and the interpreter will be shown on the screen, however the subtitles will be aligned with the LSF of the interpreter and not the original language of the interviewee. In the case where spoken French is used, the written content of the subtitles is derived from the spoken French and not from the LSF interpretation, and the audio track is removed in the final video.

The number of videos with one main signer is 295, and the number of videos with multiple signers is 73 (Table 2.1). This diversity in mode of production and mixture of monologue and dialogue makes MEDIAPI-SKEL a challenging dataset that covers a broad range of journalistic styles.

From the original videos, we extract 25 body keypoints, 2x21 hand keypoints and 70 face keypoints using OpenPose [33, 169]. We provide these 135 keypoints for every person in every frame of the 368 videos. Each keypoint includes the X and Y pixel value, as well as a confidence score between 0 and 1. Keypoints which fail to be detected or are occluded are accorded 0 values.

Note that the body keypoints of the legs and feet are essentially irrelevant for sign language processing, despite the fact that they are included in our extracted skeleton keypoints.

In addition to the 2D-skeleton video data, we provide the associated subtitles in French with their time tags. The subtitles of this corpus are accurately aligned to the 2D-skeleton video content. Each subtitle corresponds to the associated segment of sign language video. This is a particularly complex task, as the syntax of LSF is very different to the syntax of French. In LSF, contextual elements are generally provided at the start of a discourse and then later referred to, while in written French, contextual elements tend to be spread out throughout a text.

In order to maintain an accurate alignment of video segments and subtitles despite strong ordering differences in LSF and French, the subtitles produced by *Média'Pi!* are relatively long. The average length is 4.2 seconds or 11 words (Table 2.1). The French Audiovisual Council (CSA) requires subtitles to have a maximum number of 72 characters and to have a duration of at least 15 characters per second, or around 4.8 seconds for a subtitle of 72 characters.[3] This provides enough flexibility to reorder the French phrases in a natural way. The final sign of a video segment can correspond to the first word of a subtitle.

The frames at moments of transition between subtitles are semantic breaks in the LSF discourse, often characterized by a deceleration of movement. These semantic breaks are worth studying from a linguistic and machine learning perspective, as described in the challenge in Section 2.5.1

Table 2.1 provides a summary of the size and quality of MEDIAPI-SKEL. The number of signers in each video is roughly estimated by counting the number of individuals with non-occluded hands which are large enough to be easily visible in the video frame, and which also are not static. We then using facial recognition to count the number of unique individuals. We consider a video to have one signer if over 95% of the subtitle texts in that video correspond to the same signer. The vocabulary size is computed by counting the number of unique tokens, omitting punctuation.

Table 2.2 provides summary statistics for the proposed train-dev-test split for the challenges in Section 2.5

2.5 . Data Processing Challenges

Global statistics	
# subtitled videos	368
# hours	27
# frames	2.5 million
Video statistics	
Resolution	1080p (327 videos) 720p (41 videos)
Framerate	30 fps (111 videos) 25 fps (242 videos) 24 fps (15 videos)
Average length of video	4.5 minutes
# signers	>100
# videos with one main signer	295
# videos with multiple signers	73
Text statistics	
# subtitles	20 187
Average length of subtitle	4.2 seconds 10.9 words
Vocabulary size (tokens)	17 428
Vocabulary size (nouns+verbs+adjectives)	14 383

Table 2.1: Descriptive statistics

	Train	**Dev**	**Test**
# subtitled videos	278	40	50
# hours	20.9	3.0	3.5
# frames	1980k	277k	323k

Table 2.2: Train - Dev - Test split

We list the three challenges of semantic segmentation, sentence-level video-text alignment and video-text features that we intend to pursue using MEDIAPI-SKEL. These challenges are illustrated in Figure 2.2.

2.5.1 . Semantic Segmentation

[3] https://www.csa.fr/Mes-services/Foire-aux-questions/Proteger/L-accessibilite-des-programmes-aux-personnes-souffrant-de-deficience-auditive-ou-visuelle/Pourquoi-la-presentation-des-sous-titres-varie-t-elle-d-une-chaine-a-l-autre

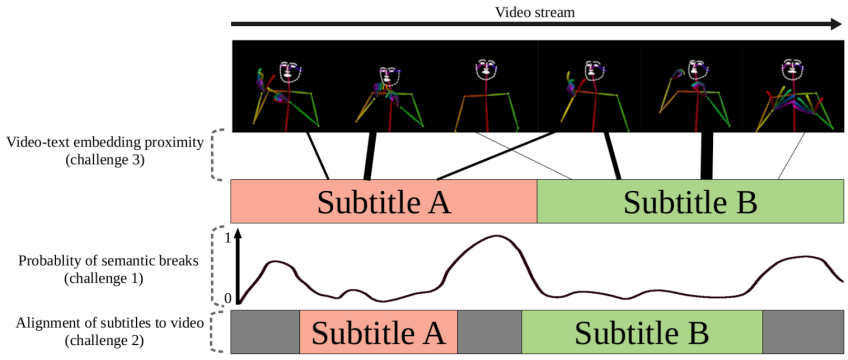


Figure 2.2: Illustration of challenges: semantic segmentation of sign language (challenge 1), alignment of text and video (challenge 2), and production of video-text embeddings for cross-modal retrieval (challenge 3)

In this challenge, we are interested in semantically segmenting a video into short units ('clauses') showing signs in their context. Concretely, we aim to detect the moments between the end of one subtitle and the beginning of the next. This challenge is useful for segmenting a video into bite-sized pieces, each of which could be translated separately. Breaking up a translation task from sign language to written language in this way can speed up the process of translation.

Moreover, this task can be considered as an intermediate task in achieving automatic alignment. A prior segmentation of a sign language video can be used to discretise the problem of matching text segments to continuous video.

The segmentation of sign language into sentence units is discussed in the linguistic literature, and automatic detection of the semantic breaks chosen by *Média'Pi!*'s subtitlers can contribute to this discussion. Different ways of defining sentences in sign language are discussed in [49], and both signers and non-signers are capable of recognizing visual cues of the start and end of sentences [71]. These visual cues could be automatically detected by using skeleton-based neural networks, such as the graph convolutional network proposed in [206]. Our challenge can help to quantitatively measure the visual cues of semantic segmentation.

2.5.2 . Alignment

We are interested in automatically aligning subtitles, or short segments of text, with the corresponding segments of sign language video. Given an ordered list of subtitle texts, can we automatically subtitle the video?

The authors of [50] conduct a similar task, aligning words to lexical sign glosses using recurrent neural networks. In [13], the authors automatically

align textual descriptions of sub-tasks to instructional videos. The order of sub-tasks described in the text annotation follows the order of actions observed in the video, and it is this feature which is exploited in their weakly-supervised learning method. This is also the case for MEDIAPI-SKEL; the order of the subtitles follows the order of the corresponding video segments.

There are numerous applications of automatic alignment of segments of text with segments of video. For example, this task can be used to create an automatic tool for subtitling videos. The process of subtitling a video manually, translating from LSF to written French, takes *Média'Pi!* almost 1 hour for each minute of video. This painstakingly long task can be simplified by automatically aligning text with sign language videos. Such tools exist for written languages, for example the software *aeneas*[4], which automatically aligns text with segments of video.

Furthermore, this task can be used to create a bilingual concordancer similar to DeepL's *Linguee*[5]. A bilingual concordancer aligns phrases in one language with phrases in another language. Such a concordancer is a translation aide tool, displaying words and phrases in their context. Whilst *Linguee* aligns text phrases with text phrases, we aim to construct an alignment between segments of text and segments of sign language video. With a concordancer, a translator can quickly search for previously translated segments of text (or even search for signs).

Finally, we can enhance existing corpora such as the BBC TV Corpus [151] by aligning the subtitles to the video stream.

2.5.3 . Video-Text Embeddings

Our third challenge is to find joint vector representations of segments of sign language video and segments of text for video-text retrieval. The goal is to find video embeddings and text embeddings in the same high-dimensional vector space, and then compute the distance between them in that space. This distance represents the semantic distance between the LSF content and the French content.

In [133], the authors present a method for video-text cross-modal retrieval, which they apply to two datasets containing short videos with textual annotations: the Microsoft Research Video to Text dataset [205] and the Microsoft Video Description dataset [39]. The method is evaluated using rank-based performance in finding the video segment that matches with a text segment, or vice versa.

One possible application of this challenge is a search engine for sign language that finds segments of video given textual input. Another application is to use the distance between video and text embeddings as a measure of loss

[4] https://www.readbeyond.it/aeneas/
[5] https://www.linguee.fr/

for the task in Section 2.5.2, which aims to find the closest match between text segments and video segments.

2.6 . Conclusion

In this chapter, we present MEDIAPI-SKEL, a new 2D-skeleton database of sign language content accurately aligned with subtitles. This corpus can be freely downloaded for public research on Ortolang[6], a language data repository.

MEDIAPI-SKEL is appropriate for training a number of sign language processing tasks beyond the classical task of sign spotting. Additionally, the corpus can be used for linguistic purposes. For example, one can quantitatively measure visual cues for semantic or grammatical structures, such as questions or lists of items. It could also be used in avatar animation from body keypoints.

In Chapter 3, we develop a method for the first of the three challenges presented in this chapter: semantic segmentation of sign language. We return to the second of the three challenges in Chapter 5, but using a different dataset. In the final Chapter 7, we present an improved version of MEDIAPI-SKEL, including RGB videos in addition to skeleton keypoints and more hours of video content.

[6] https://www.ortolang.fr/market/corpora/mediapi-skel

3 - Semantic Segmentation of Sign Language into Sentence-Like Units

In this chapter, we present baseline results for the new task of automatic segmentation of Sign Language video into sentence-like units, presented in the previous chapter. We use the MEDIAPI-SKEL corpus, presented in Chapter 2, as it contains natural Sign Language video with accurately aligned subtitles. We train a spatio-temporal graph convolutional network with a BiLSTM on 2D skeleton data to automatically detect the temporal boundaries of subtitles. In doing so, we segment Sign Language video into subtitle-units that can be translated into phrases in a written language. We achieve a ROC-AUC statistic of 0.87 at the frame level and 92% label accuracy within a time margin of 0.6s of the true labels.

The work in this chapter lead to a publication [25] at the ECCV workshop SLRTP, where it won the best paper award. I contributed to all of the experiments and the writing. The ideas come from many discussions with Annelies Braffort and Michèle Gouiffès.

3.1 . Introduction

The treatment of language as a sequence of words from a lexicon is unsuitable for SLs [72]. The notion of a 'word' in SL is ill-defined, as the beginning or end of a sign in fluent discourse is unclear. Moreover, signs can occur simultaneously, further blurring the notion of a 'word' and rendering impossible the modelisation of SL as a linear sequence of words. The iconicity of SLs means that signs are created and strongly modified according to context and meaning, rather than being drawn largely unmodified from a lexicon.

Classic natural language processing tasks including speech-to-text, word embeddings and parts-of-speech tagging currently do not have direct counterparts in SL processing. Tasks such as automatic translation between SL and written language are in a preliminary stage, with translation only possible for short and rudimentary phrases with limited vocabulary [20].

We wish to define a sentence-like unit that can be used to segment SL into short and coherent sequences that can be translated individually. This task of segmentation of SL video is useful for numerous tasks, including software for subtitling assistance, reducing sequence length for continuous SL recognition, or phrase-level alignment between SLs and spoken or written languages. Manual segmentation of SL video into sentence-like units is a fastidious and extremely time consuming task, and so we aim to automatise this problem.

Fenlon et. al. [71] demonstrate that both native signers and non-signers

can reliably segment sentence boundaries in SL using visual cues such as head rotations, nodding, blinks, eye-brow movements, pauses, lowering the hands and clasping the hands together. We aim to automatically identify such visual cues for sentence-like segmentation.

We define a *subtitle-unit* (SU) as a segment of SL video corresponding to the temporal boundaries of a written subtitle in accurately subtitled SL video. The SU is of linguistic relevance, as the person subtitling the SL video purposefully aligns phrases of text with what they consider to be equivalent phrases in SL. Implicitly, the subtitiler labels segments of SL video that can be translated into a phrase in written language.

Our key contribution is to present baseline results of the new task of automatically segmenting SL video at a sentence-like level. Our method is an adaptation of a state-of-the-art graph-based convolutional network for sequences of 2D skeleton data of natural SL. We also study the influence of different sets of articulators (body, face and hands) in this task.

After a short overview on the related work in Sec. 3.2, Sec. 3.3 introduces the corpus and Sec. 3.4 details the proposed methodology. The results are provided in Sec. 3.5.

3.2 . Related Work

To our knowledge, this chapter presents the first attempt of the task of automatic segmentation of SL into sentence-like units. This task has been suggested by Dreuw and Ney [57] as a tool for integration into a SL annotation program.

Despite a large amount of existing work for speech and text segmentation, there is debate surrounding the precise linguistic definition of a sentence in languages such as French or English [52]. Nevertheless, division by punctuation from written language is a good working solution for almost all cases. Automatic punctuation of speech can be achieved either using prosodic cues from audio or directly from a text transcription. On reference datasets, the former method tends to perform worse than the latter, but a combination of prosodic cues and a written transcription can have superior performance than either individually, as shown by Kolář and Lamel [109].

In SLs, purely oral languages, even a working notion of a sentence is unclear. Crasborn [49] proposes the pragmatic solution of identifying sentences in SL by firstly translating them into a written language and then calling a sentence the closest equivalent portion of SL to a sentence in the written language. This solution is somewhat unsatisfactory, as it requires translation to a written language.

Our definition of a SU requires translation to a written language, but our goal is to learn to segment SL into sentence-like units purely from visual cues

without translation into a written language. We note that SUs are not necessarily the same as what are sometimes called clauses, sentences or syntactic units in the linguistic literature on SL. Börstell et. al. [15] compare SUs with 'syntactic boundaries' annotated by a Deaf SL researcher. They find that many of the boundaries of the SUs overlap with the syntactic boundaries, but that there are more syntactic boundaries than there are SUs.

We consider SU boundary detection as a continuous SL recognition problem, as we learn visual cues in long sequences of video data. One main approach for continuous SL recognition consists of using RGB SL video as input, and then combining a 3D Convolutional Neural Network (CNN) with a Recursive Neural Network (RNN) to predict a sequence of words in the written language. Koller et. al. [113] use a CNN with a bi-directional LSTM (BiLSTM) and Huang et. al. [94] use a Hierachical Attention Network (HAN). Both of these articles use corpora in controlled environments with a single signer facing the camera.

Another main approach is to use sequences of skeleton data as input, which is arguably less dependent on the conditions of SL video production. Belissen et. al. [9] and Ko et. al. [107] use sequences of skeleton keypoints for continuous SL recognition, but concatenate the 2D skeleton keypoints into two vectors rather than exploiting the graph structure of the skeleton keypoints.

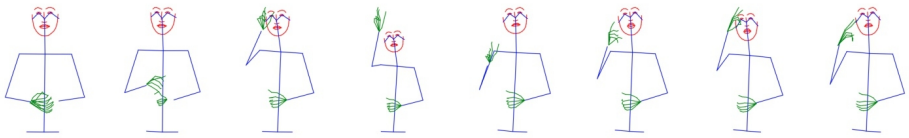
Yan et. al. [206] propose a Spatio-Temporal Graph Convolution Network (ST-GCN) for action recognition using sequences of skeleton keypoints that achieves state-of-the-art results. This model takes into account the spatio-temporal relationships between body keypoints. Our model is an adaptation of the ST-GCN, as this type of model is appropriate for our 2D skeleton video data. We combine the ST-GCN model with a BiLSTM, as we are predicting sequences not classes. This combination of a convolutional network and a BiLSTM is commonly used in language modelling [177].

3.3 . Corpus

As described in Chapter 2, the MEDIAPI-SKEL corpus [24] contains 27h of subtitled French Sign Language (LSF) video in the form of sequences of 2D skeletons (see Fig. 3.1). This corpus has the rare quality of being both natural SL (produced outside laboratory conditions) and having accurately aligned subtitles.

The subtitles in this corpus are aligned to the SL video such that the video segment corresponds to the subtitle. The original language of almost all the videos is SL, which is then translated into written language for the subtitles.[1]

[1]There are rare video segments where a hearing person is interviewed and this interview is translated into SL.



Politique : Jean-Marie Le Pen face à la justice.

Figure 3.1: MEDIAPI-SKEL corpus [24] with skeleton keypoints of LSF and aligned subtitles in written French. The graph structure connecting body keypoints (blue), face keypoints (red) and hand keypoints (green) is shown

The subtitles have been written by different people and aligned by hand, and so we expect some variation in the length and placement of the SUs. The 2D skeleton data contains 25 body keypoints, 2×21 hand keypoints and 70 facial keypoints for every person at every frame in the 27h hours of video content. Each 2-dimensional coordinate is also associated to a confidence value between 0 and 1.

This corpus contains 2.5 million frames associated to 20k subtitles, where each subtitle has an average length of 4.2 seconds and 10.9 words. The training data contains 278 videos, the validation data 40 videos, and the test data 50 videos. The average length of a video is 4.5 minutes. Videos may contain signers at different angles (not necessarily facing the camera) and around one-fifth of the videos contain multiple signers.

Since the corpus contains dialogues between multiple people in various environments, it is necessary to clean the data automatically by detecting and tracking the current signer and by removing irrelevant keypoints.

The code for our skeleton data cleaning procedure is available online.[2] The main steps consist in:

- Converting all videos to 25 frames-per-second
- Omitting the legs and feet keypoints, as they are not relevant for SL, leaving us with a total of 125 keypoints
- Tracking each person in each video using a constraint on the distance between body keypoints between consecutive frames
- Omitting people unlikely to be signers, specifically those with hands outside of the video frame, those with hands that hardly move, those that are too small (in the background of the video) or those that appear only for very short time periods (under 10 frames)

[2] https://github.com/hannahbull/clean_op_data_sl

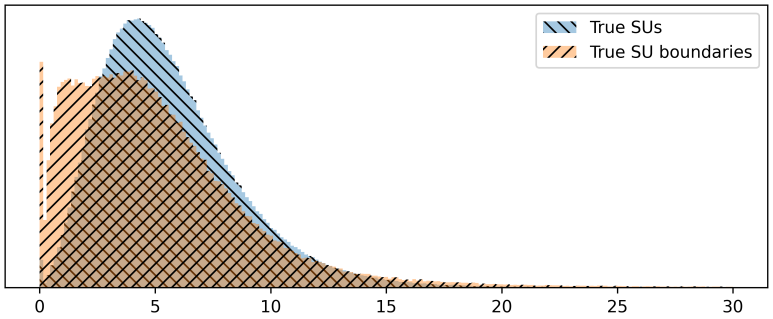


Figure 3.2: Density histogram of the average velocity of the 15 upper body keypoints of likely signers in the training set. Units are pixel distance moved per frame with 1080p resolution

- In the case of multiple potential signers, choosing the most likely signer in each second of video based on a criterion involving hand size times variation of wrist movement of the dominant hand
- Imputation of missing skeleton keypoints using past or future frames
- Temporal smoothing with a Savitzky-Golay filter

Our final input data consist of temporal sequences of variable lengths of 2D skeleton keypoints corresponding to individuals in SL video.

We label a frame of a sequence with 0 if there is no subtitle associated to that frame or if the frame is within a distance of 2 frames from a frame with no associated subtitle. We label all other frames as 1. The padding of the 0-labelled frames partially controls for the fact that the SUs are not precise at the frame-level. Frames labelled 1 are SUs, and frames labelled 0 are SU boundaries.

Fig. 3.2 shows the distribution of the average velocity of the body keypoints of likely signers in the training set by label. Sequences where there is unlikely to be a signer due to lack of hand visibility or hand movement are omitted using our data cleaning procedure. True SU boundaries tend to have lower average body keypoint velocity compared to true SUs, but velocity is an insufficient indicator to predict SU boundaries in SL discourse.

3.4 . Methodology

3.4.1 . Model

Our model is a spatio-temporal graph convolutional network (ST-GCN) following Yan et. al. [206], which we adjoin to a BiLSTM network to capture the sequential nature of the output (Fig. 5.2). The spatial graph structure

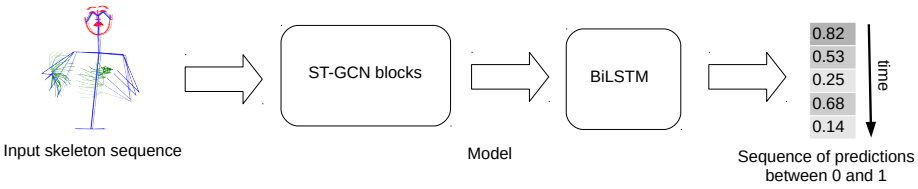


Figure 3.3: ST-GCN+BiLSTM model on skeleton sequence for SU detection

of the body keypoints, face keypoints and hand keypoints follows the human joint structure. The temporal graph structure connects body keypoints across time. The edge importance in the graph is learned during training. The convolution operation is across the spatial and temporal edges of the graph.

The ST-GCN architecture is identical to that used by Yan et. al. [206], but without temporal pooling. The model is composed of 9 layers of ST-GCN units, where the first 3 layers have 64 output units, the second 3 layers have 128 output units and the final 3 layers have 256 output units. The embedding dimension of the BiLSTM is thus 256 and we also set the hidden dimension of the BiLSTM to be 256.

Each input sequence of skeleton keypoints has a length of 125 frames, but we take every second frame of the video, so this corresponds to a sequence length of 10s. This means that we expect around two or three SUs per sequence, as the average subtitle length is 4.2s.

Each skeleton sequence is normalised such that the mean and variance of the x-coordinates and y-coordinates of the skeleton over time are equal to 0 and 1. During training, we add random flips to the horizontal dimension of the skeleton keypoints in order to take into account for left-handed and right-handed signers. We also shuffle the order of skeleton sequences at each epoch.

We use SGD optimisation with a learning rate of 0.01, a weight decay of 0.0001, Nesterov momentum of 0.9 and binary cross-entropy loss. The model is trained for 30 epochs. Due to memory constraints, the batch-size is 4.

3.4.2 . Experiments

We train our model on 278 videos and test our model on 50 videos. Our full model uses 15 body keypoints, 70 face keypoints and 2×21 hand keypoints shown respectively in blue, red and green in Fig. 3.1. In order to understand the contributions of the body, face and hand keypoints, we train the model using only the body keypoints, only the face keypoints and only the hand keypoints, as well as the body + face, the body + hand and the body + face + hand keypoints. We keep the architecture of the model constant.

Moreover, we compare the performance of our model between videos

with one signer and videos with multiple signers. The videos with multiple signers often contain dialogues between people not necessarily facing directly at the camera. This is to test the robustness of our model to more diverse scenarios.

3.4.3 . Evaluation Criteria

Our evaluation metrics should take into account that SUs are not annotated by the subtitler at a frame-level accuracy. We propose both frame-wise and unit-wise metrics, allowing for shifts in SUs.

As a flexible frame-wise metric, we propose dynamic time warping (DTW) with a window constraint as an evaluation criteria. This computes the distance between the true sequence and the predicted sequence of SUs, allowing for frames to be shifted within a certain window length w. We compute this DTW accuracy for different values of the window length w. When $w = 0$, this is the frame-wise difference between the predicted SUs and the true SUs. We also compute the DTW distance for $w \in \{5, 10, 15\}$, which corresponds to the minimum frame-wise difference between the predicted SUs and the true SUs allowing for frames to be shifted up to 5, 10 or 15 frames.

Additionally, we compute the ROC-AUC statistic, the frame-wise precision, recall and $F1$-score. The precision is given by the number of frames correctly identified with the label 0 divided by the total number of frames identified with the label 0. The recall is given by the number of frames correctly identified with the label 0 divided by the total number of true frames with the label 0. The $F1$ score is the harmonic mean of precision and recall.

Furthermore, we consider unit-wise evaluation metrics, allowing for 15 frame (0.6s) shifts in SU boundaries. We match each predicted SU boundary to the closest true SU boundary, where the closest true SU boundary is defined as the true SU boundary with the greatest intersection with the predicted SU boundary, or, in the case of no intersection, the closest true SU boundary within 15 frames. Calculating the number of matches divided by the total number of predicted SU boundaries gives us a unit-wise precision metric. In the same way, we can match each true SU boundary to the closest predicted SU boundary. The number of matches divided by the total number of true SU boundaries gives us a unit-wise recall metric. From this precision and recall metric, we can compute a unit-wise $F1$ score.

3.5 . Results and Discussion

Table 3.1 shows frame-wise evaluation metrics on the test set. Our results are encouraging and we obtain a ROC-AUC statistic of 0.87 for our predictions, with the highest score obtained using the body, face and hand keypoints. Instead of relying on the frame-wise error rate, it is important to account for slight shifts in SUs as those who subtitle the videos do not aim for accuracy at

	DTW0	DTW5	DTW10	DTW15	AUC	Prec.	Recall	F1
full	0.1660	0.1255	0.1045	0.0927	**0.8723**	0.5023	**0.7510**	0.6019
face+body	0.1560	0.1172	0.0973	0.0868	0.8708	0.5241	0.7259	0.6087
body+hands	0.1661	0.1269	0.1064	0.0952	0.8659	0.5023	0.7380	0.5977
face	0.1858	0.1483	0.1248	0.1100	0.8325	0.4624	0.6830	0.5514
body	**0.1410**	**0.1055**	**0.0882**	**0.0790**	0.8704	**0.5616**	0.7122	**0.6280**
hands	0.1821	0.1417	0.1186	0.1053	0.8554	0.4713	0.7360	0.5747
Pre-processing	*0.1406*	*0.1365*	*0.1333*	*0.1309*	*0.6039*	*0.7828*	*0.2201*	*0.3436*
Constant pred.	*0.1672*	*0.1672*	*0.1672*	*0.1672*	*0.5000*	*0.1671*	*1.0000*	*0.2865*

Table 3.1: Frame-wise evaluation metrics on the test set. The full model uses face, body and hand keypoints. The pre-processing version shows an evaluation after annotation of segments without an identified signer as not belonging to SUs. The final line shows the results for a constant prediction. DTW0 is the frame-wise prediction error. DTW5, DTW10 and DTW15 are the DTW errors respectively allowing for a 5, 10 and 15 frame discrepancy in predictions

the level of the frame. Allowing for shifts of up to 0.6s (15 frames), we obtain a frame-wise error rate of 8% when using only the body keypoints. Table 3.2 presents unit-wise evaluation results and shows that 76% of true SU boundaries can be associated to a predicted SU boundary within 15 frames.

When asking native signers to annotate sentence boundaries in SL, Fenlon et. al. [71] found inter-participant agreement of sentence boundary annotation within 1 second to be around 63%. Whilst this is not exactly the same task as subtitling SL video, we can expect that there is quite a high degree of variation in the choice of subtitle boundaries. In light of this finding, our error rate seems reasonable.

Part of the accuracy of our model is accounted for by pre-processing the data to label obvious SU boundaries, such as moments where there are no signers in the video. Such frames are correctly identified as having no associated subtitle 78% of the time, as noted in the second last line of Table 3.1. Errors here seem to be mostly due to subtitles extending beyond scenes containing signers, rather than failure to detect a signer in a scene, however further annotation of signers would be needed to verify this. Our ST-GCN+BiLSTM model makes significant improvements on top of this pre-processing.

From Table 3.1, we see that the full model has the highest ROC-AUC statistic and the highest recall, suggesting that including the facial and hand keypoints detects the most SU boundaries. However, the body model makes fewer incorrect predictions of SU boundaries and has a higher precision. Our unit-wise metrics in Table 3.2 reinforce this observation. The full model correctly identifies 76% of the true SU boundaries within 15 frames, but the body model has the highest precision with 71% of the predicted SU boundaries

	Prec.	**Recall**	**F1**
full	0.6609	**0.7631**	0.7083
face+body	0.6840	0.7408	**0.7113**
body+hands	0.6250	0.7492	0.6815
face	0.6403	0.6909	0.6646
body	**0.7090**	0.6866	0.6976
hands	0.6147	0.7619	0.6804
Pre-processing	*0.9341*	*0.0803*	*0.1478*

Table 3.2: Unit-wise evaluation metrics on the test set allowing for 15 frame (0.6s) shifts in SU boundaries. The full model uses face, body and hand keypoints. The pre-processing version shows an evaluation after annotation of segments without an identified signer as not belonging to SUs

within 15 frames of a true SU boundary. Börstell et. al. [15] find that there are more 'syntactic boundaries' than SUs. Perhaps our full model is good at learning visual cues of such 'syntactic boundaries', which do not always correspond to actual SU boundaries.

Fig. 3.4 shows an example of the predictions and true labels on a video from the test set using the full model. Most of the true SU boundaries are correctly detected, however there is an over-detection of SU boundaries. Fig. 3.7 shows that the predicted lengths of SUs using the full model is shorter than the true lengths of SUs. This difference in length is less pronounced when using the body model. Moreover, predicted SU boundaries tend to be slightly longer than the true SU boundaries. The median difference between predicted SU boundaries and the associated true SU boundaries within 15 frames is around 5-7 frames in all our models. The median absolute difference between predicted SU boundaries and the associated true SU boundaries is 7-9 frames. The problem of over-detection or under-detection of SU boundaries and differences in lengths could be alleviated by assigning length and regularity priors to the SUs. This is similar to applying shape priors in image segmentation [37], [189].

Fig. 3.5 and Fig. 3.6 show examples of correct and incorrect predictions from Fig. 3.4. The left of Fig. 3.5 shows an example of an obvious SU boundary where the signer pauses with their hands folded. This is correctly predicted by our model, albeit our predicted SU boundary is a little longer than the true boundary. The right of Fig. 3.5 shows a SU boundary with more subtle visual cues, including the head turning towards the camera and a slight deceleration of movement. This is also correctly detected by our model, but with a slight shift of about half of a second.

The left of Fig. 3.6 shows an SU boundary detected by our model but

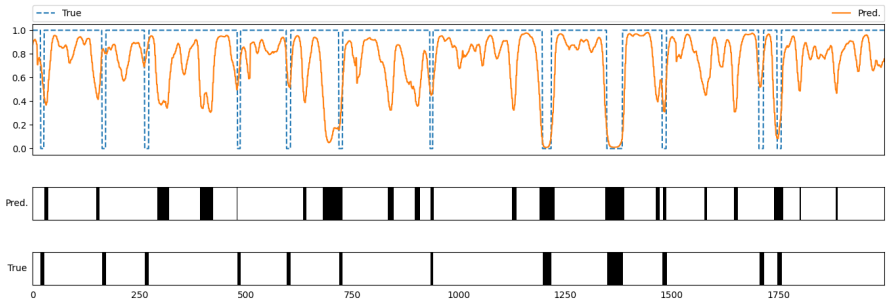


Figure 3.4: True and predicted labels for a video sequence using the full (face+body+hands) model

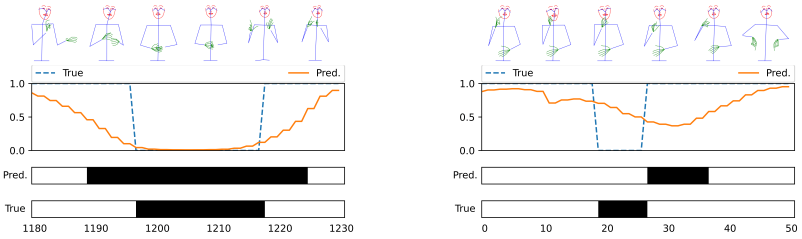


Figure 3.5: Correctly detected SU boundaries from Fig. 3.4

which is not a true SU boundary. However, this particular example could have been an SU boundary, had the subtitles for this video been aligned differently. Some of our incorrectly detected SU boundaries are thus likely to correspond to sentence-like boundaries but which are simply not annotated as such by the subtitler. The right of Fig. 3.6 shows a SU boundary not detected by our model. This particular SU boundary does not have clear visual cues, and its detection may perhaps require an understanding of the SL sequence.

Facial visual cues for semantic boundaries in SL can include blinks, eyebrow movements, head nodding or turning the head to stare directly at the camera. Manual cues include specific hand movements and the signer folding their hands together at the waist level. We thus assess whether or not including facial and hand keypoints improves SU detection. We cannot conclude that adding the face and the hand keypoints to the body model makes a significant improvement to SU detection. Nevertheless, the face keypoints or the hand keypoints alone make surprisingly accurate predictions. The face model has a ROC-AUC statistic of 0.83. Subtle facial cues are likely to be picked up by our model. Similarly, the hands alone make relatively accurate predictions.

As shown in Table 3.3, accuracy is reduced amongst test videos with more than one signer, but the ROC-AUC statistic is still relatively high at 0.84. The

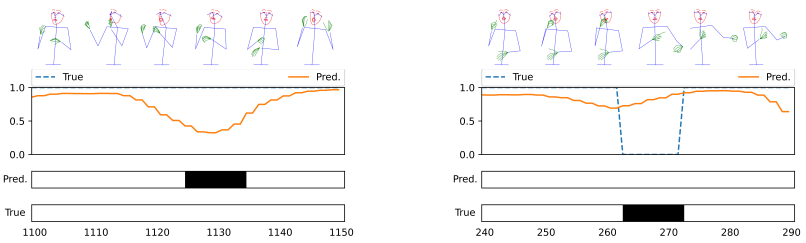


Figure 3.6: Incorrectly detected SU boundaries from Fig. 3.4

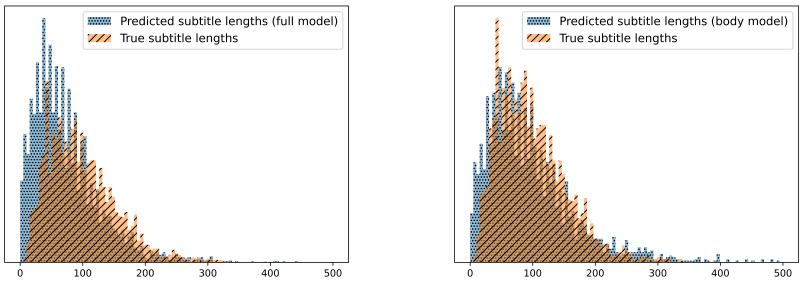


Figure 3.7: Length of true SUs compared to predicted SUs

	DTW0	DTW5	DTW10	DTW15	AUC	Prec.	Recall	F1
full 1 signer	0.1366	0.0959	0.0776	0.0686	**0.8876**	0.5144	**0.7456**	0.6088
body 1 s.	**0.1204**	**0.0838**	**0.0676**	**0.0601**	0.8854	**0.5611**	0.7140	**0.6284**
full >1 s.	0.2227	0.1824	0.1562	0.1392	**0.8388**	0.4878	**0.7579**	0.5936
body >1 s.	**0.1809**	**0.1474**	**0.1278**	**0.1156**	0.8365	**0.5622**	0.7100	**0.6275**

Table 3.3: Evaluation metrics for videos with one signer and videos with multiple signers. Models and evaluation metrics are as in Table 3.1

DTW error rate with a window length of 15 is 12%. On videos in the test set with one signer, the ROC-AUC statistic is 0.89 and the DTW error rate with a window length of 15 frames is only 6%. This suggests that our model is robust to natural SL video, including examples of dialogue between multiple signers.

3.6 . Conclusion

We provide baseline results for automatic segmentation of SL video into sentence-like units. We use natural SL video and allow multiple signers and camera angles. Our results are encouraging, given the variability of identification of semantic boundaries in a SL discourse across different annotators and given the fact that the SU annotations are not accurate at the frame level.

Our full model using face, body and hand keypoints has a high recall statistic but finds more SUs than necessary. We are interested to find out whether or not these additional SU boundaries correspond to semantic boundaries in the SL discourse that are not annotated by the subtitler. Further annotation of our test data would be required in order to see whether or not this is the case.

Some ways to improve our model include better controlling the final distribution of the SUs. For example, we would like to be able to set priors on the duration of the SUs in order to control the length of segments and the regularity of the segmentation. Identifying certain signs would also improve detection of SUs.

Segmenting sign language into subtitle-units is a first step towards aligning text sentences to sign language video. Given a text segmentation (e.g. subtitles) and an approximate alignment of text to video, we can use our SU detection model to segment sign language video and then associate the text segments to the nearest video segments.

Due to the relatively small number of hours in MEDIAPI-SKEL (27 hours), we explore the problem of subtitle alignment in the context of a new dataset, BOBSL [4], containing 1400 hours of British Sign Language video. The subtitle alignment problem is particularly applicable to this dataset, because the videos contain subtitles which are aligned to the audio track and not to the signing, in contrast to MEDIAPI-SKEL. We present the BOBSL dataset in Chapter 4, then describe a new method for subtitle alignment on BOBSL in Chapter 5, comparing to the SU-detection baseline presented in this chapter.

Part II

Using Interpreted TV Programmes for Subtitle Alignment and Dense Annotation of Sign Language Video

4 - BOBSL: A Large Dataset of Sign Language Interpreted TV Shows

In this chapter, we introduce the BBC-Oxford British Sign Language (BOBSL) dataset, a large-scale video collection of British Sign Language (BSL). BOBSL is an extended and publicly released dataset based on the BSL-1K dataset [2]. We describe the motivation for the dataset, together with statistics and available annotations. Finally, we describe several strengths and limitations of the data from the perspectives of machine learning and linguistics, note sources of bias present in the dataset, and discuss potential applications of BOBSL in the context of sign language technology. The dataset is available at https://www.robots.ox.ac.uk/~vgg/data/bobsl/.

This work was conducted as part of a collaboration with researchers from VGG, Univeristy of Oxford and from LIGM, Ecole des Ponts, Université Gustav Eiffel. There are a number of authors responsible for various aspects of acquiring, preparing, annotating and presenting the BOBSL dataset. I contributed to only some parts of the resulting publication [4]. My contributions are in part of the pre-processing of the BOBSL dataset, where I found and corrected an alignment error in the subtitles of around one quarter of the videos, and in the parts of the article on aligning subtitles to the signing.

4.1 . Introduction

To date, a central challenge in conducting sign language technology research has been a lack of large-scale public datasets for training and evaluating computational models [20]. The goal of the BBC-Oxford British Sign Language (BOBSL) dataset is to provide a collection of BSL videos to support research on tasks such as sign recognition, sign language alignment and sign language translation.

The rest of the chapter is structured as follows: in Sec. 4.2 we provide an overview of the BOBSL dataset; in Sec. 4.3, we describe the collection and annotation (both automatic and manual) of the dataset, and also the evaluation partitions. In Sec. 4.4 we discuss the opportunities and limitations of the data from the perspectives of sign linguistics and downstream applications and note several sources of bias present in the data before concluding in Sec. 4.5.

4.2 . BOBSL Dataset Overview

In this section, we first give an overview of BOBSL content and statistics (Sec. 4.2.1). Next, we compare BOBSL to existing sign language datasets (Sec. 4.2.2), outline data usage terms (Sec. 4.2.3) and describe its relationship to the BSL-1K dataset (Sec. 4.2.4).

4.2.1 . Dataset Content and Statistics

The data consists of BSL-interpreted BBC broadcast footage, along with English subtitles corresponding to the audio content, as shown in Fig 4.1. The data contains 1,962 *episodes*, which span a total of 426 differently named TV *shows*. The term *episode* refers to a single video of contiguous broadcast content, whereas a show (such as *"Countryfile"*) refers to a collection of episodes grouped thematically by the broadcaster, whose episodes typically share significant overlap in subject matter, presenters, actors or storylines. The shows can be partitioned into five genres using BBC metadata as shown in Fig. 4.2; with the majority of shows being *factual*, i.e. documentaries. These can be further divided into 22 topics, as shown in Fig. 4.3. Including horror, period and medical dramas, history, nature and science documentaries, sitcoms, children's shows, and programs covering cooking, beauty, business and travel, the BOBSL data covers a wide range of topics.

Statistics of the BOBSL data are presented in Tab. 4.1. The 1,962 episodes have a duration of approximately 1,467 hours (i.e. 45 minutes per episode on average, with the majority of episodes lasting approximately 30 or 60 minutes, as shown in Fig. 4.5). The videos have a resolution of 444×444 pixels and a frame rate of 25 fps. There are approximately 1.2M sentences extracted from English subtitles covering a total vocabulary size of 78K English words. BOBSL contains a total of 39 signers (interpreters). The data is divided into train, validation and test splits based on signers, to enable signer-independent evaluation, i.e. there is no signer overlap between the three splits. The distribution of programs associated to each signer, together with the split information is illustrated in Fig. 4.4. We note that a few signers appear very frequently.

4.2.2 . Comparison to Existing Datasets

In Tab. 4.2, we present a number of existing datasets used for sign language research – mainly for the tasks of sign recognition, sign spotting, continuous sign language recognition, sign language translation and sign language production. We refer to [114] for an extended list of corpora of European sign languages, including those used for linguistic analyses. Benchmarks have been proposed for American [7, 102, 117, 201, 58, 59], German [195, 111], Swiss-German [63, 32], Flemish [32], Chinese [36, 93, 94, 214], Finnish [191], Indian [172, 104], Greek [1], Turkish [146, 170], Korean [107] and British [135, 163, 2] sign languages. These datasets can be grouped into *isolated* signing (where the signer performs a single sign, usually at a slow speed for clarity, starting from and ending in a neutral pose) and

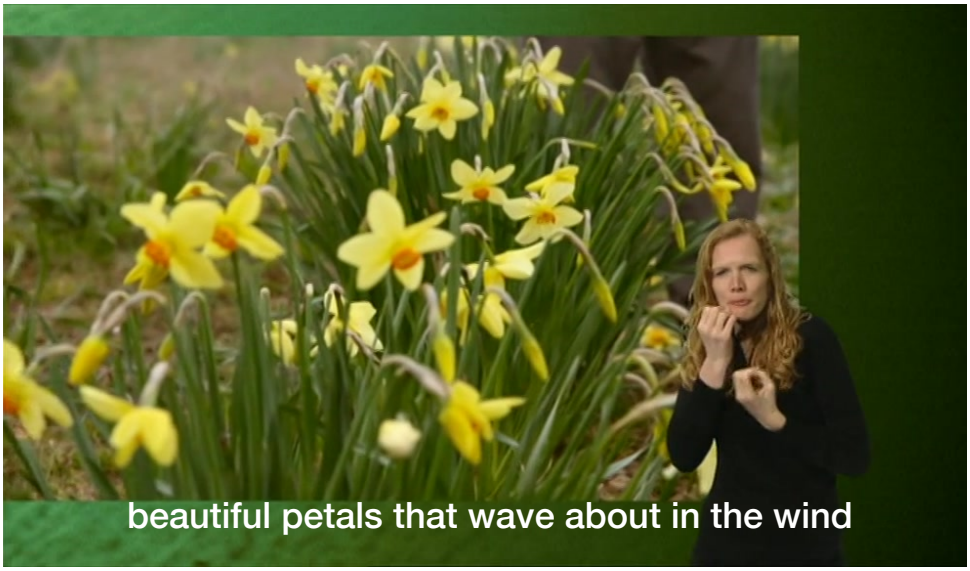


Figure 4.1: **BOBSL source data.** The source data consists of British Sign Language interpreted footage of BBC broadcasts (in this example from a *Gardeners' World* program), along with English subtitles corresponding to the audio content.

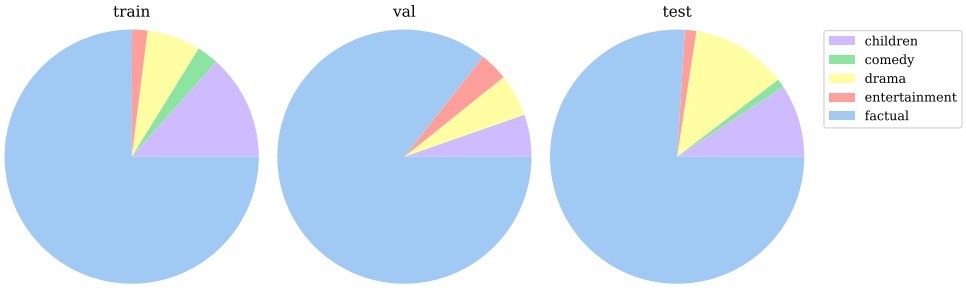


Figure 4.2: **BOBSL division into genres.** The duration of each BOBSL dataset split can be divided into 5 genres, with *factual* representing the largest proportion for train, validation and test splits.

co-articulated signing. Co-articulated signing, or "signs in context", describes signing that exhibits variation in sign form caused by immediately preceding or following signs, or signs articulated at the same time. If we are to build robust models which can understand sign language *"in the wild"*, we need to recognise co-articulated signs.

Most datasets in Tab. 4.2 fall into one or more of the following categories: (i) They have a limited number of signers – for example, Devisign [36], ASLLVD [7], ISL [104], GSL [1] have 8 or fewer signers. (ii) They have a limited vocabulary of signs – for example, Purdue RVL-SLLL [201], BOSTON104 [58], INCLUDE [172], AUTSL [170], SMILE [63] only have a few hundred signs. (iii)

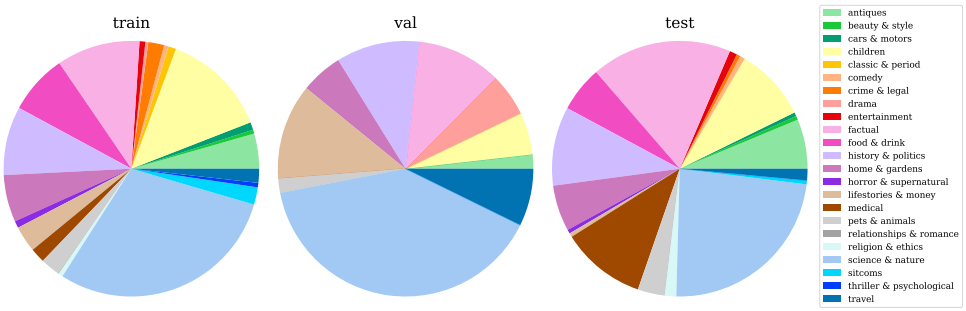


Figure 4.3: **BOBSL division into topics.** Each BOBSL dataset split can be divided into 22 topics, with *science & nature* representing the largest proportion for train, validation and test splits. The figure is best seen on computer screen and in colour.

They have a large vocabulary of signs but only of isolated signs – for example MSASL [102] and WLASL [117] have vocabularies of 1K and 2K signs, respectively. (iv) They are recorded in lab settings. (v) They are limited in total duration – for example the popular PHOENIX14T [111] dataset contains only 11 hours of content. (vi) They represent natural co-articulated signs but cover a limited domain of discourse – for example, the videos in PHOENIX14T [111] and SWISSTXT-WEATHER [32] are only from weather broadcasts.

BOBSL is most similar in content to PHOENIX14T [111], SWISSTXT-WEATHER [32], SWISSTXT-NEWS [32], VRT-NEWS [32] and BSL-1K [2]. These datasets are all built from sign language interpreted TV broadcasts. PHOENIX14T [111], SWISSTXT-WEATHER [32], SWISSTXT-NEWS [32] and VRT-NEWS [32] all provide valuable aligned subtitle annotations, but are comparatively small in scale (the latter three datasets also provide larger "RAW" unaligned variants akin to BOBSL that are approximately an order of magnitude smaller than BOBSL in duration). They are also restricted to a single domain of discourse: weather broadcasts for PHOENIX14T [111] and SWISSTXT-WEATHER [32]; news broadcasts for SWISSTXT-NEWS [32] and VRT-NEWS [32]. In contrast, BOBSL covers a variety of genres (see Fig. 4.2) and topics (see Fig. 4.3). The relationship of BOBSL to the BSL-1K dataset is discussed in Sec. 4.2.4.

In summary, the BOBSL dataset presents several advantages: it consists of co-articulated signs as opposed to isolated signs, representing more natural signing (note that BOBSL nevertheless remains distinct from conversational signing, due to its use of interpreted content). BOBSL provides the largest source of continuous signing (1,467 hours); it covers a large domain of discourse; it is automatically annotated for a large vocabulary of more than 2,000 signs. We note that since the annotations provided on the training and validation sets are obtained through automatic methods, they may contain some noise.

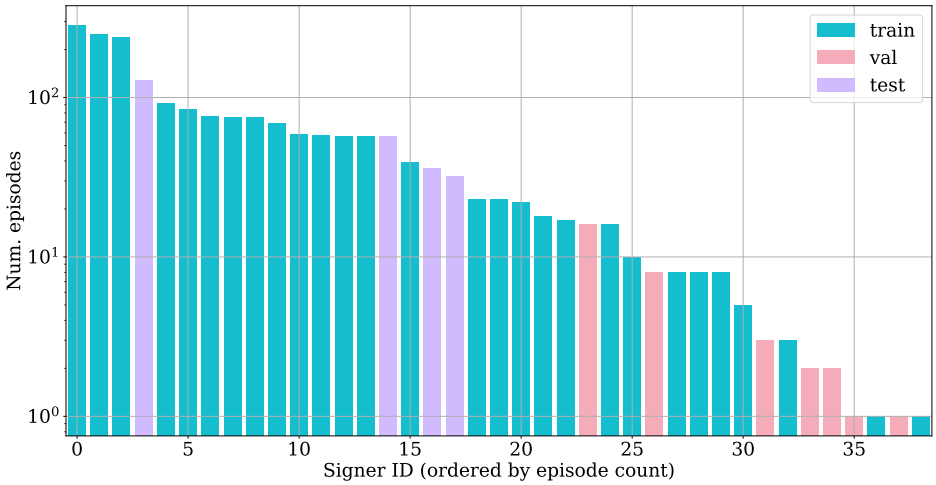


Figure 4.4: **Distribution over signers.** The number of episodes associated with each BSL interpreter in the BOBSL dataset follows a power law distribution (note the log-scale on the y-axis).

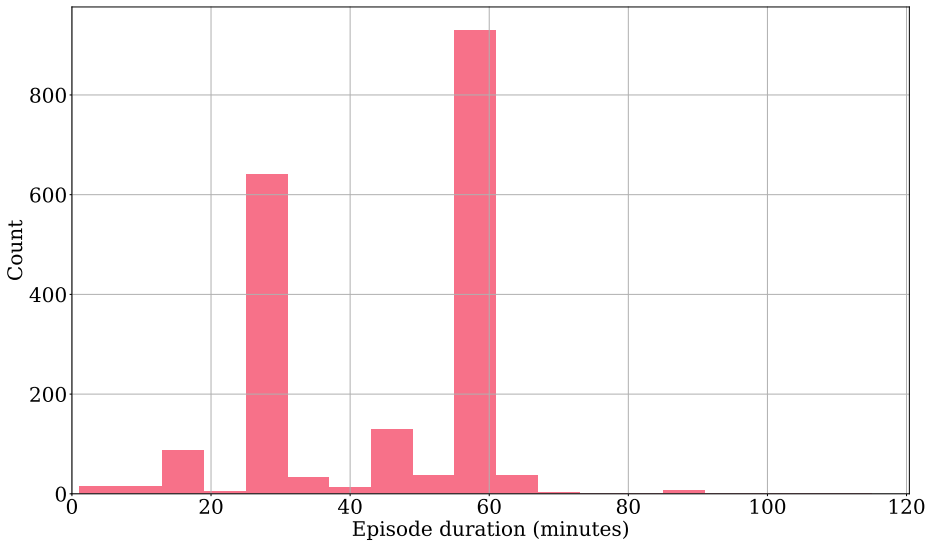


Figure 4.5: **Distribution over episode durations.** The duration of episode videos in the BOBSL dataset. The majority of episodes are either 30 minutes or 60 minutes in duration, with the longest episode lasting 120 minutes.

Split	Episodes	Num. Signers	Num. Raw Subs.	Num. Sent.	Sent. Word Count	Text Vocab.	Out-of-vocab (O-O-V)	Singletons	Avg. Dur. (mins)	Total Dur. (hours)
train	1,675	28	1,108K	1,004K	9,557K	72K	-	22.0K	44.3	1,236
val	33	7	22K	20K	205K	14K	0.8K	6.1K	50.2	28
test	254	4	192K	168K	1,593K	35K	4.8K	11.9K	48.2	204
total	1,962	39	1,322K	1,193K	11,356K	78K	-	23.6K	45.1	1,467

Table 4.1: **Statistics summarising the data distributed across the splits of BOBSL.** *Num. Signers* indicates the number of signer identities within a partition, *Num. Raw Subtitles* denotes the number of subtitles (which do not necessarily form complete sentences) associated with the original broadcasts, while *Num. Sentences* indicates the number of English sentences that were parsed from these subtitles using the process described in Sec. 4.3.4. *Text Vocabulary* indicates the vocabulary across the sentences after removing punctuation, special characters, digits etc. *Out-of-vocab* denotes the number of words that are not present in the training split, while *Singletons* denotes the number of words appearing only once in the given partition. *Duration* indicates the duration of the episodes.

4.2.3 . Research Use and Potential Changes

BSL translation services are currently supplied to the BBC by Red Bee Media Ltd. They have indicated that they and their staff are happy for their footage to be used for research purposes. However, if the position changes the dataset will need to be revised accordingly. Researchers should be mindful of this, and should be aware that the 'Permission to Use' form they will need to sign obligates them to delete portions (or, indeed, the whole) of the dataset in the future, if so instructed.

4.2.4 . Relationship to the BSL-1K Dataset

Previous work [2] introduces the BSL-1K dataset, a collection of BSL videos that were automatically annotated with sign instances via a keyword spotting method. This collection of automatic sign instances was further expanded through other methods for sign localisation [135, 187]. A short test sequence was manually annotated for temporal sign segmentation evaluation in [154, 155]. However, BSL-1K remained as an internal dataset. BOBSL represents a public, extended dataset based on BSL-1K using videos drawn from the same source distribution with no overlap between episodes to BSL-1K, but significant overlap between signers and shows, and preserving the same signer independent train, validation and test split identities for signers that appear in both datasets. The BOBSL dataset is larger than BSL-1K (1,467 hours vs 1,060 hours). BOBSL has been automatically annotated with sign instance timings using the same techniques as for BSL-1K. Through a data-sharing agreement with the BBC, BOBSL is available for non-commercial research usage.

Dataset	lang	co-articulated	sign vocab	#sign annots (avg. per sign)	text vocab	#words	#sequences	#signers	source	#hours
Devisign [36]	CSL	✗	2,000	24K (12)	-	-	-	8	lab	13-33
CSL500 [93]	CSL	✗	500	125K (250)	-	-	-	50	lab	69-139
ASLLVD [7]	ASL	✗	2,742	9K (3)	-	-	-	6	lab	4
ASL-LEX 2.0 [165]	ASL	✗	2,723	2723 (1)	-	-	-	-	lexicons, lab, web	-
MSASL [102]	ASL	✗	1,000	25K (25)	-	-	-	222	lexicons, web	25
WLASL [117]	ASL	✗	2,000	21K (11)	-	-	-	119	lexicons, web	14
BSLDict [135]	BSL	✗	9,283	14K (1)	-	-	-	148	lexicons	9
BosphorusSign22k [146]	TSL	✗	744	23K (30)	-	-	-	6	lab	19
AUTSL [170]	TSL	✗	226	38K (170)	-	-	-	43	lab	21
INCLUDE [172]	ISL	✗	263	4K (16)	-	-	-	7	lab	3
SMILE [63]	DSGS	✗	100	9K (90)	-	-	-	30	lab	-
S-pot [191]	FinSL	✓	1,211	6K (5)	-	-	4K	5	lab	9
Purdue RVL-SLLL [201]	ASL	✓	104	2K (19)	130	213	-	14	lab	-
BOSTON104 [58]	ASL	✓	104	1K (10)	-	-	201	3	lab	1
How2Sign [59]	ASL	✓	-	-	16K	598K	35K	11	lab	79
CSL100 [94]	CSL	✓	-	-	178	175K	25K	50	lab	100
CSL-Daily [214]	CSL	✓	2,000	151K (76)	2K	312K	21K	10	lab	23
SIGNUM [195]	DGS	✓	450	137K (304)	1K	166K	33K	25	lab	55
Phoenix14T [111, 29]	DGS	✓	1,066	76K (71)	3K	114K	8K	9	TV	11
KETI [107]	KSL	✓	524	15K (28)	-	-	-	14	lab	28
ISL [104]	ISL	✓	-	-	10K	-	9K	5	web	18
GSL [1]	GSL	✓	310	41K	481	44K	10K	7	lab	10
SWISSTXT-WEATHER [32]	DSGS	✓	-	-	1K	7K	1K	-	TV	1
SWISSTXT-NEWS [32]	DSGS	✓	-	-	11K	73K	6K	-	TV	9
SWISSTXT-RAW-WEATHER [32]	DSGS	✓	-	-	-	-	-	-	TV	12
SWISSTXT-RAW-NEWS [32]	DSGS	✓	-	-	-	-	-	-	TV	76
VRT-NEWS [32]	VGT	✓	-	-	7K	80K	7K	-	TV	9
VRT-RAW [32]	VGT	✓	-	-	-	-	-	-	TV	100
MEDIAPI-SKEL [24]	LSF	✓	-	-	17K	220K	20K	>100	TV	27
Dicta-Sign-LSF-v2 [9]	LSF	✓	2,551	35K (14)	-	-	-	16	lab	11
BSL Corpus [163]	BSL	✓	5K	72K (14)	-	-	-	249	lab	125
BSL-1K [2]	BSL	✓	1,064†	273K† (257)	59K	9M	1M	40	TV	1,060
BOBSL	BSL	✓	2,281†	452K† (198)	78K	11.4M	1.2M	39	TV	1,467

Table 4.2: **Summary statistics of sign language datasets.** Language, co-articulated vs. isolated signing, sign vocabulary size, total number of sign annotations, corresponding spoken language vocabulary (if provided by dataset), total number of spoken language words, number of sequences, number of signers, source of data and duration in hours for each dataset. †Denotes the statistics of the subset of annotations used for sign language recognition experiments on these datasets, but in practice larger vocabularies are annotated for details of annotations on BOBSL).

4.3 . BOBSL Dataset Construction

In this section, we describe the construction of the BOBSL dataset. We first describe the raw source data and the pre-processing pipeline employed to prepare the data for sign language research (Sec. 4.3.1). Next, we describe how the data was divided into train, validation and test splits (Sec. 4.3.2) and the automatic methods used to annotate this data with sign instance timings (Sec. 4.3.3). We detail the manual annotation processes in (Sec. 4.3.5) together with details on subtitle sentence extraction (Sec. 4.3.4). Finally, we describe the BOBSL partitions for translation and alignment tasks (Sec. 4.3.6).

Dataset genesis. This dataset has been created in partnership with the British Broadcasting Corporation (BBC), the UK's largest public service broadcaster. The UK broadcast regulator has set a threshold for the amount

of accessible content broadcasters must supply. As a result, the BBC produces subtitles for 100% of its TV output, audio description for more than 20% of its output and BSL translations for more than 5% of its output. Due to the size of its weekly broadcast output and its long-term retention of this metadata it has a comparatively large datastore of useful data for partner universities to work with.

The sort of data release represented by BOBSL is a core part of BBC R&D's remit as mandated by the UK Parliament.[1] As a result the BBC is keen to support research into accessibility services by supplying data to partner universities and administering non-commercial testing and training data to the wider academic community.

4.3.1 . Source Data and Pre-Processing

Source data. An initial collection of TV episodes were provided by the BBC. These were broadcast between 2007 and 2020 and vary from a few minutes to 120 minutes in duration (see Fig. 4.5 for the distribution of episode durations). Each episode is accompanied by a corresponding set of written English subtitles, derived from the audio track of the show. The programs span a wide range of topics (history, drama, science etc.)—a detailed summary of the content included is provided in Sec. 4.2.1. The majority of these shows are accompanied by a BSL interpreter, overlaid on the bottom right hand corner of the screen in a fixed location. Note that sign interpreters produce a *interpretation* of the speech that appears in the subtitles, as opposed to a *transcription*.[2] This means that words in the subtitles may not correspond directly to individual signs produced by the interpreters, and vice versa. The videos have a height of 576x pixels, a display aspect ratio of 16:9 and a frame rate of 25 fps.

Filtering and pre-processing. First, TV programs that were known to not contain a BSL interpreter in a fixed region of the screen were removed from the collection. A small number of videos that exhibited significant data corruptions were also removed.

Video pre-processing. Each video is cropped to include only the bottom-right 444×444 pixel region containing the BSL interpreter (see Fig. 4.6). The automatic face detection and tracking pipeline provided by the authors of [44] was used to detect and track faces, with the goal of blurring those appearing in the content behind the intepreter. There are 224,957 blurred face tracks over 170 hours of video. Some examples are shown in Fig. 4.7. The pipeline

[1] The 2016 Agreement with the Department for Media, Culture and Sport mandates the BBC to "ensure it conducts research and development activities geared to ...maintain[ing] the BBC's leading role in research and development in broadcasting ...in co-operation with suitable partners, such as university departments ..." (Section 65 of http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/about/how_we_govern/2016/agreement.pdf).

[2] There may also be discrepancies between the audio and the subtitle text.

performs well for clearly visible background faces, but there are likely to be a small number of background faces that are not blurred.

Subtitle pre-processing. After manual inspection, we observed that approximately one quarter of the subtitle files exhibited discrepancies in time alignment between the audio track and the subtitle timestamps. To address these cases, we applied standard methods of forced alignment using an acoustic model.[3]

After pre-processing the videos and subtitles, the audio track of each video was removed. The final result of these filtering and pre-processing steps was a collection of 1,962 videos containing BSL interpreters with corresponding audio-aligned written English subtitles that form the public dataset release.

Figure 4.6: **Pre-processing.** Raw broadcast footage is pre-processed by extracting a 444×444 pixel square crop from the bottom right-hand corner region occupied by the BSL interpreter in each video (illustrated by the orange dashed box).

4.3.2 . Dataset Splits

To support the development of signer-independent systems (in which models are evaluated on signers not seen during training), the dataset is divided into train, validation, test splits according to the estimated identity of the BSL interpreters.

Signers are assigned to separate splits, to produce the dataset statistics given in Tab. 4.1. The distribution of episodes associated to each signer, together with the split information is illustrated in Fig. 4.4.

[3] https://www.readbeyond.it/aeneas/, https://subsync.online/

Figure 4.7: **Background face blurring.** Faces appearing behind the interpreter are automatically tracked and blurred for anonymisation purposes.

4.3.3 . Automatic Annotation via Sign Spotting and Localisation Methods

Due to the large scale of the BOBSL dataset, exhaustive manual annotation of individual signs would be prohibitively expensive. Automatic annotation techniques for sign instance localisation are used, making use of the information within weakly-aligned subtitles. In particular: (1) the mouthing keyword spotting approach from [2], (2) the dictionary spotting approach from [135], and (3) the attention spotting approach from [187] to annotate the data. We give a brief summary of each method here and refer the reader to the original papers for further details. Fig. 4.8 provides sample annotations from each method on a sample training video.

(1) **Keyword spotting with mouthings.** A sign may consist of not just movements of the hands, but also head movements, facial expressions and mouthings [180]. Mouthings have multiple roles: they can be used to specify the meaning of a sign in the case of polysemy and to disambiguate manual homonyms [203]. Mouthings appear frequently in BSL - accompanying over 2/3 of signs in one study [179]. From an annotation perspective, mouthings provide a cue for *sign spotting*, the task of localising a given sign in a signing sequence.

The method proposed in [2] is used to spot signs, by searching for mouthings corresponding to subtitle words. A keyword spotting model is used to find whether and when a mouthing occurs within a 10s window of the weakly-aligned subtitle. The model outputs a confidence score associated to each frame, and all localisations above 0.5 threshold are considered automatic

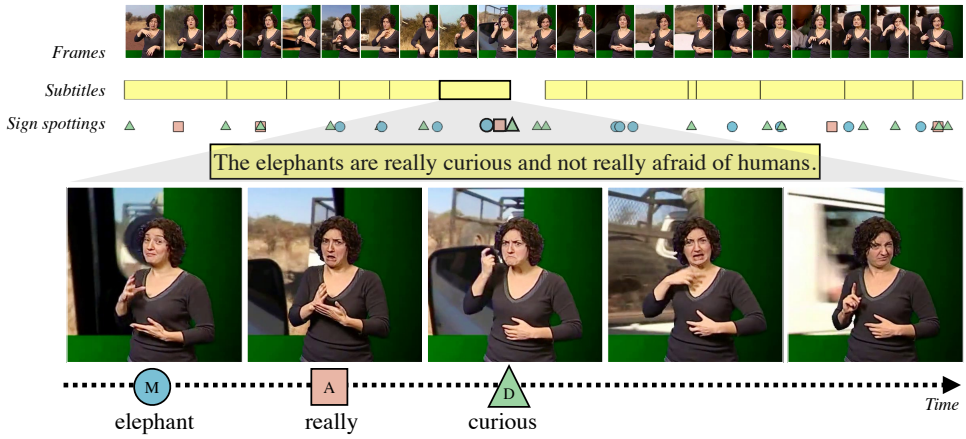


Figure 4.8: **BOBSL sample with automatic sign annotations.** A sample training video, together with the corresponding English language subtitle, and automatic annotations generated through three sign spotting techniques (*M*: mouthing, *D*: dictionary, *A*: attention, described in Sec. 4.3.3).

mouthing annotations (after a non-maximum suppression stage as in [2]). Fig. 4.14 provides statistics for the amount of annotations on the training set.

To derive the list of candidate keywords for spotting, *text normalisation* is applied to the subtitles using the method of [73]. This normalisation converts dates and numbers to their written form, e.g. 13 becomes "thirteen". The list of keywords is further filtered to words that appear in the CMU phonetic dictionary [171] with at least four phonemes. This filtering results in a final list of 43K search keywords.

The keyword spotting model used is an improved variant of the model of Stafylakis et al. [173] from [134] (described in their paper as "P2G [173] baseline"). The model is trained on "talking heads" datasets (LRW [45] and LRS2 [46]) of BBC TV broadcasts. While the model has never been trained on signers, it generalizes well to a large set of signer mouthings. As observed in [2], the peak in the posterior probability assigned to the presence of a keyword typically corresponds to (approximately) the end of the mouthing/sign. Qualitative examples of automatically retrieved signs through this method are shown in Fig. 4.9.

(2) **Sign spotting with dictionaries.** Following the method proposed in [135], given a video of an isolated sign from a dictionary, signs are located in continuous, co-articulated sign language video. A joint embedding space is used to measure similarity between isolated dictionary videos and continuous signing. This method leverages the weakly-aligned sentences by querying words in the sentence within a $\pm$4sec padded neighbourhood around the subtitle timestamps. In particular, words and phrases from the BSLDict [135] vocabulary are queried. In order to determine whether a query from the dictionary

Figure 4.9: **BOBSL automatic sign annotations through mouthings.** Examples of automatically retrieved instances of four different signs on each row (*magic*, *special*, *quality*, *wonderful*) obtained through the pipeline of keyword spotting with mouthings.

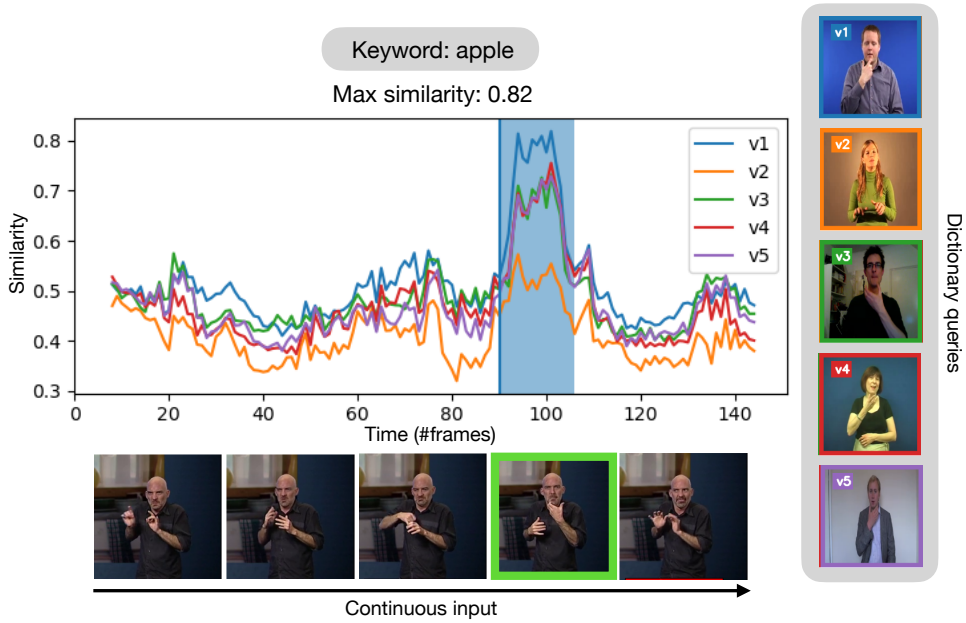


Figure 4.10: **BOBSL automatic sign annotations through dictionaries.** The localisation procedure for comparing dictionary samples for a given keyword with a continuous signing.

occurs in the sentence, the sentence is represented in its original and lemmatised forms and the query in its original and text-normalised forms. If a match is found, the dictionary videos corresponding to the word/phrase are queried.

In order to obtain the embedding space, a slightly different training procedure as [135] is used for simplicity (see [4] for details). An I3D classification model is trained jointly on continuous annotations and BSLDict samples, using the mouthing (threshold=0.8) and dictionary (threshold=0.8) spottings from BSL-1K, as well as BSLDict videos filtered to the 1K vocabulary of [2].

A single embedding for the dictionary sample is obtained by averaging features computed with multiple frame rates as in [135]. A sequence of embeddings for the BOBSL search window is obtained by applying a sliding window with a stride of 4 frames. After computing the similarity between the continuous signing search window and each of the dictionary variants for a given word/phrase, the best match is determined to be the match with the highest similarity score. All localisations with a similarity above a 0.7 threshold are considered automatic dictionary annotations. Fig. 4.14 provides statistics for the number of annotations on the training set. Fig. 4.10 contains an illustration of the similarity plots across variants.

(3) **Sign localisation with Transformer attention.** In contrast to the two previous automatic annotation methods, the approach [187] of localising signs

differs considerably in that it is *context-aware*. A Transformer model [188] is trained to predict, given an input stream of continuous signing, the sequence of corresponding written tokens. By using the trained attention mechanism of the Transformer to align written English tokens to signs, signs can be localised. More specifically, once the model is trained, new sign instances are localised for tokens that have been correctly predicted by determining the index at which the corresponding encoder-decoder attention is maximised. Even low values for the maximum attention score provide good localisations; therefore, thus no threshold is applied for attention spottings. Fig. 4.14 provides statistics for the amount of annotations on the training set.

In practice, the Transformer is trained on a subset of video-text pairs which contain at least one sign automatic annotation (from the two previously described methods) within the sentence timestamps, in order to ensure there is an approximate alignment between the source signing video and target written token sequence. The encoder input video is represented by a 1024-dimensional feature sequence, extracted from an I3D model provided by [187] which is trained on sign classification with BSLK-1K [2] for a 5K vocabulary of signs (obtained from mouthing and dictionary spottings) applied with a sliding window of stride 4. For building the target written sequences, (1) words in every sentence are lemmatised, assuming inflected versions of the same word map to the same sign, (2) the vocabulary is filtered to 18K lemmas obtained by combining the automatic annotations from mouthing (threshold=0.7) and dictionary (threshold=0.8) spottings, and (3) stop words are removed.

Recent work has also demonstrated the effectiveness of the Transformer for sign spotting with dictionaries [95]—we defer an investigation of this approach to future work.

4.3.4 . Sentence Extraction

The subtitles associated with the BOBSL episodes are approximately aligned to the audio track of the corresponding content but do not necessarily fall into well-formed sentences. To support research into tasks such as sign language translation (which often operates at the sentence-level [29, 32]) well-formed sentences are extracted from the subtitles. This is done semi-automatically by splitting subtitles on sentence boundary punctuation and employing a combination of heuristics and manual inspection to resolve ambiguous cases. To preserve an approximate time alignment between the sentences and the signing, when multiple sentences fall within a single subtitle, each sentence is assigned a duration in proportion to its written length (in characters) as a fraction of the original subtitle. Finally, sentences that correspond to descriptions of background music lyrics (these are typically unsigned) and sentences that are known to fall outside the feasible signing period (e.g. those that occur after the show credits) are

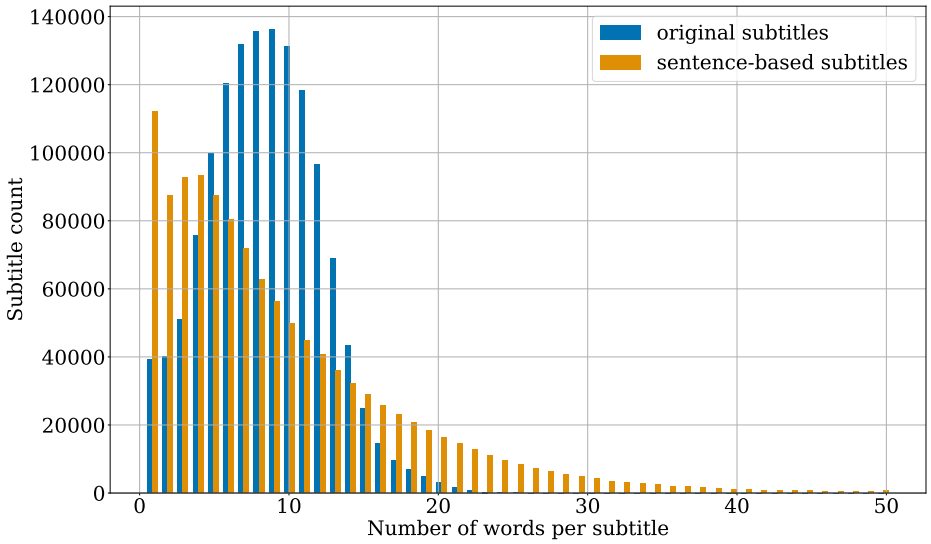


Figure 4.11: **The distribution of lengths for the original and sentence-aligned subtitles.** In contrast to the broadcast subtitles which possess a relatively small variance in length, the sentence-based subtitles exhibit a broader variance, with a greater number of very short (just a few words) and very long (more than 30 words) sequences.

removed. The result of this sentence extraction process is a collection of "sentence-based" subtitles (in which each subtitle corresponds to a single sentence), summarised in Tab. 4.1. In comparison to the original subtitles (which are relatively uniform in duration) the distribution of sentence lengths exhibits broader variance (this effect is visualised in Fig. 4.11). Note that since the sentence extraction process makes use of punctuation in the subtitles, some long subtitles may be due to missing punctuation: a manual inspection of random samples determined that this occurs relatively rarely.

4.3.5 . Manual Annotation

Sign verification. Deaf annotators proficient in BSL used a variant of the VIA tool that was adapted for whole-sign verification [61] (see Fig. 4.12), similarly to the process used by [2]. To enable efficient collection, labels were collected for temporal proposals for signs in the test split by verifying/discarding automatic spottings that were assigned high confidence scores by the automatic sign spotting techniques (above 0.9 confidence for mouthing annotations, above 0.8 for the dictionary annotations). When viewing a temporal proposal, the video could be played at different speeds (and replayed if needed). For each proposed spotting location, the annotator is able to indicate: (i) whether the sign is correct, incorrect, or that they are unsure, (ii) whether fingerspelling (using the manual alphabet to spell English words) was used, (iii) further com-

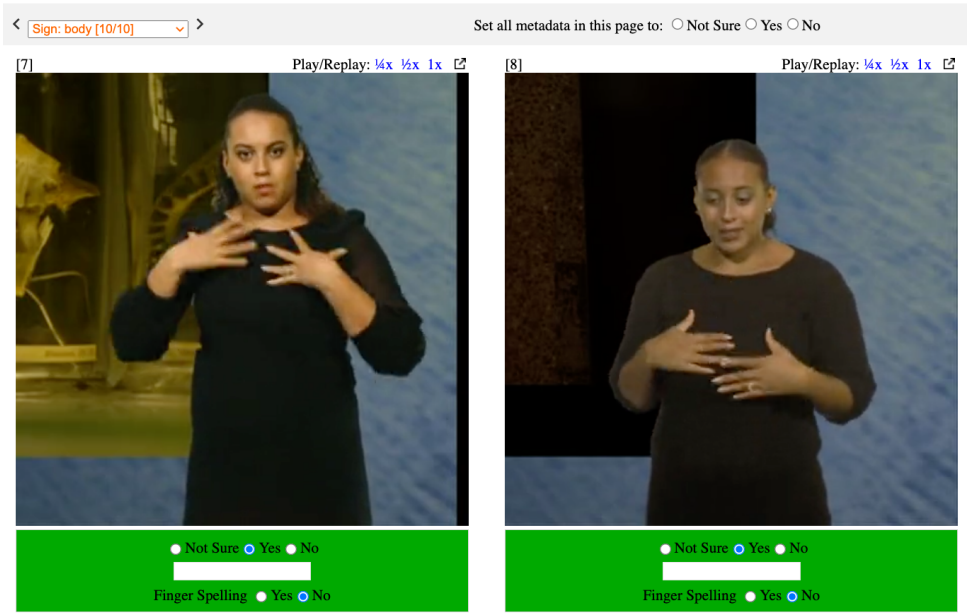


Figure 4.12: **Manual annotations.** A screenshot of the VIA Whole-Sign Verification Tool. Given proposed temporal windows from the automatic sign spotting methods described in Sec. 4.3, annotators can mark the proposals as correct, incorrect or unsure, and provide additional metadata (see Sec. 4.3.5 for details).

ments, including the meaning of the sign (if the proposed meaning was incorrect), and any other observations.

For quality control, a small random sample of the annotations were further verified by a deaf native signer of BSL. Of the mouthing spottings within a 2,281 vocabulary with a confidence of at least 0.9 that were annotated, 63.6% were marked correct, yielding 9,263 verified signs spanning 1,653 classes. The latter figure includes predictions that were corrected by annotators, as well as a small number of verified low confidence signs that were annotated during early development. Of the dictionary spottings within the 2,281 vocabulary with a confidence of at least 0.8 that were annotated, 75.8% were marked correct, yielding 15,782 verified signs (spanning 765 classes) after including corrections. These verification statistics also exclude a small number signs that were tagged by annotators as "inappropriate" in modern BSL signing.

Sentence alignment. To support research into the tasks of sign language alignment and translation, the sentences for a subset of the episodes are manually aligned with the signing content (they are initially coarsely aligned with the audio content). The audio-aligned sentences differ from the signing-aligned subtitles in both start time and duration, as shown in Fig. 4.15. To perform the alignment, an adapted version of the VIA tool is used, shown in Fig. 4.13. The annotator is presented with a list of sentences for which they

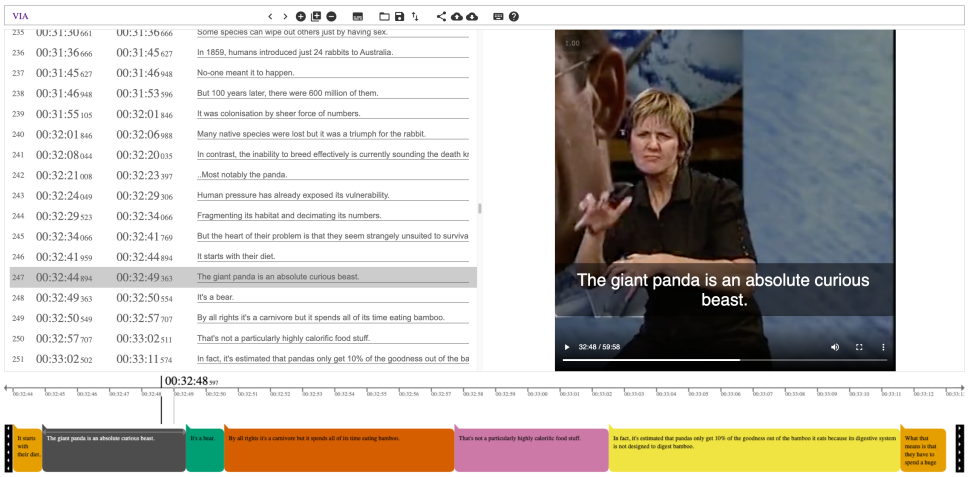


Figure 4.13: **Sentence alignment tool.** A screenshot of the VIA sentence alignment tool [61]. The annotator uses a "draggable" visualisation of the temporal extent of the sentences at the bottom of the screen to perform alignment, with the ability to pause and replay segments.

are able to adjust timings by clicking and dragging elements on a webpage (this methodology is similar to the alignment tool described in the concurrent approach of [32]). These sentence-level alignment annotations are available.

4.3.6 . BOBSL Partitions for Sentence Alignment and Translation Evaluations

In order to develop methods for sign language sentence alignment and translation, we need aligned continuous signing segments and corresponding English sentences. We propose to make use of two levels of alignment: (i) audio-aligned video-sentence alignments that have been filtered using automatic spotting annotations to select sentences that are likely to be reasonably well aligned to the signing (these are available in large numbers); (ii) manual video-sentence alignments (these are available in smaller numbers).

Spotting-filtered signing video-sentence alignments. These correspond to video segments for which an automatic sign instance annotation falls within the corresponding sentence timestamps (we restrict ourselves to annotations obtained from mouthings and dictionaries with confidence over 0.8 and use all annotations obtained through attention) and the word matching the sign occurs in the sentence. This indicates a probable approximate alignment between the signing video and corresponding sentence. For the sentence timestamps, we use the audio-aligned timestamps shifted by +2.7 seconds – this is the average shift calculated between audio-aligned and signing-aligned sentences in our manual training set (Sent-Train$_{H}$) described next. We define these splits as Sent-Train$_{SF}$, Sent-Val$_{SF}$, Sent-Test$_{SF}$. These

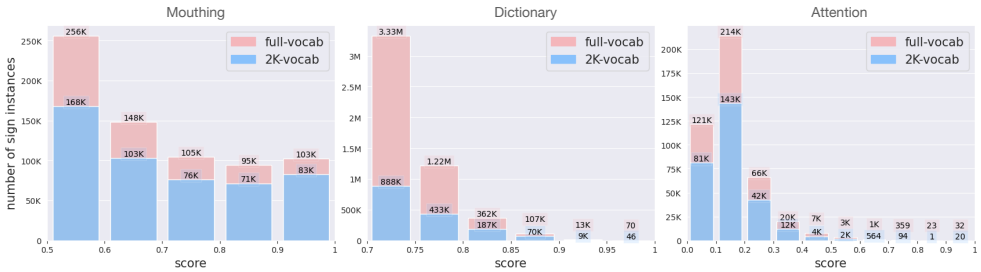


Figure 4.14: **Automatic training set of sign instances.** There are several million automatic annotations through sign spotting, with varying levels of noise. Different subsets of the training set obtained from mouthing (M), dictionary (D) and attention (A) spottings are shown according to their confidence scores. Note that the range of scores for each annotation type is different; minimal thresholds of 0.5, 0.7, 0.0 are necessary for M, D, A, respectively.

spotting-filtered alignments enable large-scale training over multiple domains of discourse.

Manual signing video-sentence alignments. These manual sentence-level alignments are obtained through the process described in Sec. 4.3.5, with statistics shown in Tab. 4.3. There is a total of 32K manually aligned sentences for a total duration of 46 hours. The training set episodes are chosen to maximise the number of signers. Given access only to the manual training set, the number of out-of-vocabulary (OOV) words is 1,127 words for the validation set and 8,030 words for the test set. The distribution of show topics for the different splits is shown in Fig. 4.16, with *science & nature* representing the largest proportion for all dataset splits (see Fig. 4.3). We define these splits as Sent-Train$_H$, Sent-Val$_H$, Sent-Test.

4.4 . Opportunities and Limitations of the Data

In this section we discuss some of the opportunities and limitations of the data from several perspectives: sign linguistics (Sec. 4.4.1), annotator observations (Sec. 4.4.2) and dataset bias (Sec. 4.4.3).

4.4.1 . A Sign Linguistics Perspective

The availability of this dataset represents a positive advance for enabling studies from a linguistics perspective. One challenge with existing technologically-focused research on sign languages is that it has made use of small databases, with few signers, limited content and limited naturalness. The present dataset is large-scale, with a broad range of content, and produced by signers of recognised high levels of proficiency. Nevertheless, there are limitations that should be recognised. First among these is that although this is a relatively large dataset, it includes only 39 signers, all

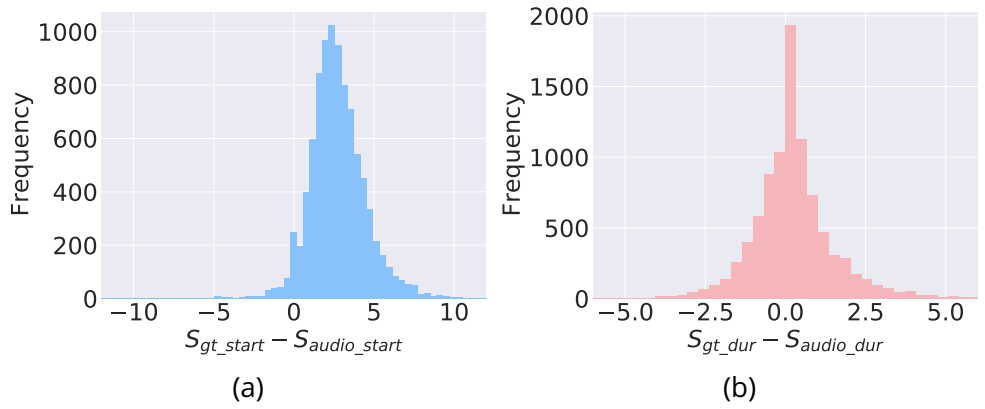


Figure 4.15: **Audio-aligned versus signing-aligned subtitles.** We plot the distribution of temporal shifts between the signing-aligned and audio-aligned subtitles for Sent-Train$_H$ by showing the differences in subtitle (a) start times and (b) duration.

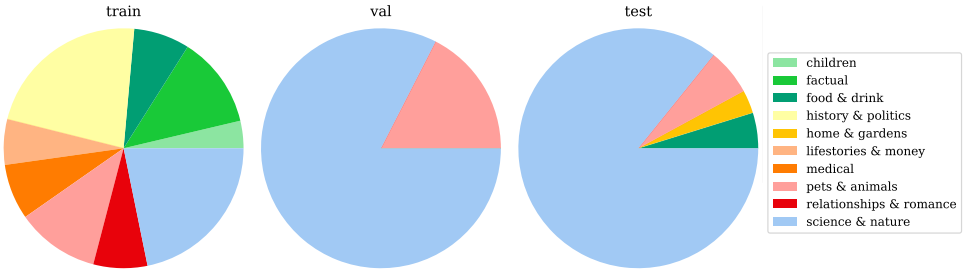


Figure 4.16: **Manual signing video-sentence alignment data divided into topics.** While *science & nature* represent the largest proportion of the validation and test set, the training set covers a broader range of themes.

using the same formal linguistic register, and—because the signing is in the context of broadcast television—little of the well-documented regional lexical variation in BSL [174] is apparent. Secondly, all of the material is translated from English. There is research evidence of systematic differences between interpreted and non-interpreted language [51]. with evidence that differences in forms of language are reduced in interpreted texts. Finally, as an additional observation, we note that there is some evidence of differences between the output of hearing and deaf interpreters [176], which may manifest in the BOBSL data.

4.4.2 . Observations from the Annotation Process

During the process of constructing the dataset, several observations arose from the annotation process that provide useful additional context for working with BOBSL. First, it was highlighted that it is frequently the case that not

Split	Episodes	Signers	Sentences	Vocab.	O-O-V	Singletons	Duration (hours)
Sent-Train$_{H}$	16	16	9,168	8,906	-	4,371	13
Sent-Val$_{H}$	4	3	1,973	3,528	1,127	1,837	3
Sent-Train$_{SF}$	1673	28	294,944	8,954	-	23	1,236
Sent-Val$_{SF}$	33	4	7,594	6,318	0	337	28
Sent-Test†	36	3	20,870	13,641	(H: 8,030, SF: 6,490)*	5,604	31

Table 4.3: **Aligned sentence-level subtitles.** Statistics summarising the BOBSL data for which manually aligned sentence-level subtitles (indicated with an H subscript) and automatically "spotting-filtered" sentence-level subtitles (indicated with an SF subscript) are available. See Sec. 4.3.5 for a description of the annotation process and Sec. 4.3.6 for details on how these splits were constructed. †Note that Sent-Test consists of human-aligned sentences. *Out-of-vocabulary (O-O-V) statistics reported w.r.t Sent-Train$_{H}$ and Sent-Train$_{SF}$.

all words present in the subtitles are captured by the signing of the BSL interpreter. Instances when this occurs are tagged and provided as part of the manually aligned sentence annotations to support further analysis. Second, it was noted that a small number of signs are used that would no longer be considered appropriate in modern BSL. These signs have been identified in the manually verified spottings of the test set, and are excluded from evaluation. However, we note that there are likely to be other occurrences of such signs in the rest of the data. We highlight this property to researchers working with the dataset, with particular relevance for research that uses the data to train sign language production models.

4.4.3 . Data Bias

While there are several promising research opportunities associated with BOBSL, it is important to also recognise the limitations of the dataset. Here we note factors that may have implications for the generalisation of models trained on this data. First, the data was gathered from TV broadcast footage: consequently, the content of the signing reflects the content of TV shows, rather than spontaneous, conversational signing. A second consequence is that the distribution of interpreters follows that of the original broadcasts, in which not all demographics are equally represented. A third consequence of using broadcast interpretations is that the interpreters may choose not to convey information from the audio stream that they consider to be redundant to the visual stream of the footage. Additional potential sources of bias stem from our use of automatic annotation: (1) First, the distribution of signs that were annotated by spotting mouthings skew towards signs that are more commonly associated with mouthing patterns, as well as towards interpret-

ers who sign with more pronounced spoken components. (2) Second, by constructing benchmark test sets for sign classification through *human verification of automatic sign proposals*, the distribution of test set signs will exhibit higher similarity to the training set distribution than would be expected if the test set was annotated without automatic proposals. There is a trade-off here: our semi-automatic *"propose and verify"* pipeline has the benefit of significantly enhanced annotator efficiency (enabling the creation of much larger and more comprehensive test sets than would otherwise be possible). However, as a consequence of the bias introduced by the *propose and verify* pipeline, researchers and practitioners should note the gap that remains between evaluation performance on the BOBSL test sets and expected performance on real world signing. Noting these important caveats, we nevertheless hope that BOBSL forms a useful, large-scale benchmark to spur progress within the research community.

4.5 . Conclusion

We introduced BOBSL, a large-scale dataset of British Sign Language. We hope that this dataset will provide a useful resource for researchers in the computer vision, natural language processing and sign linguistics communities. BOBSL contains around 1400 hours of BSL with written English subtitles. There are around 55 hours of video for which the written English subtitles have been manually aligned to the signing (13 hours in the training set). These manually aligned subtitles can be used as strong supervision for the task of automatically aligning subtitles to sign language video. The remaining subtitles are aligned to the audio, and can be used as weak supervision for this task. There are automatic annotations using mouthing, dictionary and attention spotting methods. These annotations are sparse, and there are many words present in the subtitle text and present in the videos in the form of lexical signs. In order to fully annotate the lexical signs in BOBSL, we need to significantly increase the yield of automatic annotations.

In the following two chapters, we use BOBSL to train models to align subtitles to sign language video (Chapter 5), and to densely annotate lexical signs (Chapter 6).

5 - Aligning Subtitles to Signing in Interpreted TV Data

This chapter extends upon the segmentation method proposed in Chapter 3, where we segment sign language video into subtitle units using prosodic cues. In this chapter, we train a model to jointly segment and align subtitle units to subtitle texts using both prosodic and semantic cues. We train and evaluate our model on three British Sign Language datasets, BSL-1K [2], BSL Corpus [164] and BOBSL, presented in Chapter 4.

Our goal is to temporally align asynchronous subtitles in sign language videos. In particular, we focus on sign-language interpreted TV broadcast data comprising (i) a video of continuous signing, and (ii) subtitles corresponding to the audio content. We propose a Transformer architecture tailored for this task, which we train on manually annotated alignments. We use BERT subtitle embeddings and CNN video representations learned for sign recognition to encode the two signals, which interact through a series of attention layers. Our model outputs frame-level predictions, i.e., for each video frame, whether it belongs to the queried subtitle or not. Through extensive evaluations, we show substantial improvements over existing alignment baselines that do not make use of subtitle text embeddings for learning, such as in Chapter 3. Our automatic alignment model opens up possibilities for advancing machine translation of sign languages via providing continuously synchronized video-text data.

The resulting publication of this chapter [26] is the result of a collaboration between researchers at VGG, University of Oxford and LIGM, Ecole des Ponts, Université Gustav Eiffel. In particular, Triantafyllos Afouras and myself contributed equally to setting up and running all the experiments in this chapter. The writing was shared amongst all co-authors. The ideas come from many discussions between all of the co-authors.

5.1 . Introduction

Our goal in this chapter is to temporally localise subtitles in continuous signing video. Automatic alignment of subtitle text to signing content has great potential for a wide range of applications including assistive tools for education and translation, indexing of sign language video corpora, efficient subtitling technology for signing vloggers[1], and automatic construction

[1]Unlike spoken vlogs that benefit from automatic closed captioning on sites such as YouTube, signing vlog creators who wish to provide written subtitles must both translate *and* align their subtitles manually.

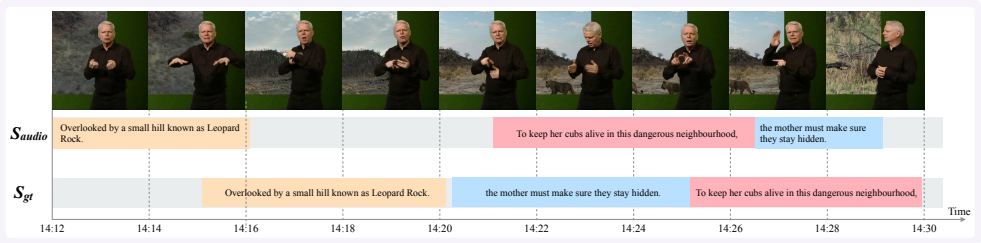


Figure 5.1: **Subtitle alignment**: We study the task of aligning subtitles to continuous signing in sign language interpreted TV broadcast data. The subtitles in such settings usually correspond to and are aligned with the audio content (top: audio subtitles, S_{audio}) but are unaligned with the accompanying signing (bottom: Ground Truth annotation of the signing corresponding to the subtitle, S_{gt}). This is a *very challenging* task as (i) the *order* of subtitles varies between spoken and sign languages, (ii) the *duration* of a subtitle differs considerably between signing and speech, and (iii) the signing corresponds to a *translation* of the speech as opposed to a transcription.

of large-scale sign language datasets that support computer vision and linguistic research.

Despite recent advances in computer vision, machine translation between continuous signing and written language remains largely unsolved [20]. Recent works [29, 31] have shown promising translation results, but to date these have been achieved only in *constrained* settings where continuous signing is *manually pre-segmented* into clips, with each clip associated to a written sentence from a *limited vocabulary*. Two key bottlenecks for scaling up translation to continuous signing depicting unconstrained vocabularies are (i) the segmentation of signing into sentence-like units, and (ii) the availability of large-scale sign language training data.

Manual alignment of subtitles to sign language video is tedious – an expert fluent in sign language takes approximately 10-15 hours to align subtitles to 1 hour of continuous sign language video. In this work, we focus on the task of aligning a particular known subtitle within a given temporal signing window. We explore this task in the context of sign language interpreted TV broadcast footage – a readily available and large-scale source of data – where the subtitles are synchronised with the audio, but the corresponding sign language translations are largely unaligned due to differences between spoken and sign languages as well as lags from the live interpretation.

Subtitle alignment to continuous signing remains a *very challenging* task. First, sign languages have grammatical structures that vary considerably from those of spoken languages [180], and as a result the *ordering* of words within a subtitle as well as the subtitles themselves is often not maintained in the signing (see Fig. 6.2). Second, the *duration* of a subtitle varies considerably between signing and speech due to differences in speed and grammar. Third,

the signing corresponds to an *interpretation* of the speech that appears in the subtitles as opposed to a transcription: there is no direct one-to-one mapping between audio/subtitle words and signs produced by interpreters, and entire parts may not be signed.[2]

Previous work exploiting such weakly-aligned data has mainly focused on finding sparse correspondences between keywords in the subtitle and individual signs [2, 135, 187], as opposed to localising the start and end times of a complete subtitle text in continuous signing. Though, as we show, localising isolated signs identified by keyword spotting nevertheless forms a useful pre-training task for full subtitle alignment. In Chapter 3, we consider the task of segmenting a continuous signing video into subtitle units purely based on body keypoints. In fact, similarly to speech which can be segmented based on prosodic cues such as pauses, sign sentence boundaries can *to an extent* be detected through visual cues such as lowering the hands, head movement, pauses, and facial expressions [71]. However, as shown in our evaluations in Sec. 5.4, such approaches based on prosody-only perform poorly in our setting, where subtitles do not necessarily correspond to complete sign sentences with clear visual boundaries.

In this chapter, we instead propose to use *the subtitle text as an additional signal* for better alignment. We make the following three contributions: (1) we show that encoding the subtitle text as input to the alignment model significantly improves the temporal localisation quality as opposed to only relying on visual cues to segment continuous sign language videos into subtitle units; (2) we design a novel formulation for the subtitle alignment task based on Transformers; and (3) we present a comprehensive study ablating our design choices and provide promising results for this new task when evaluating on unseen signers and content.

Sec. 5.2 details related work, Sec. 5.3 describes our method, Sec. 5.4 presents our experimental results and before concluding in Sec. 5.5.

5.2 . Related Work

Here, we review relevant works on temporal localisation at the levels of individual signs and sequences, in addition to more general temporal alignment methods from the literature.

Temporal localisation of individual signs. A rich body of work has considered the task of localising sparse sign instances in continuous signing, often referred to as "sign spotting". Early efforts using signing gloves [122] were followed by methods employing hand-crafted visual features to represent the hands, face and motion that were integrated with CRFs [208, 207], HMMs [161]

[2]Note that there may also be discrepancies between the audio and the written subtitle transcription.

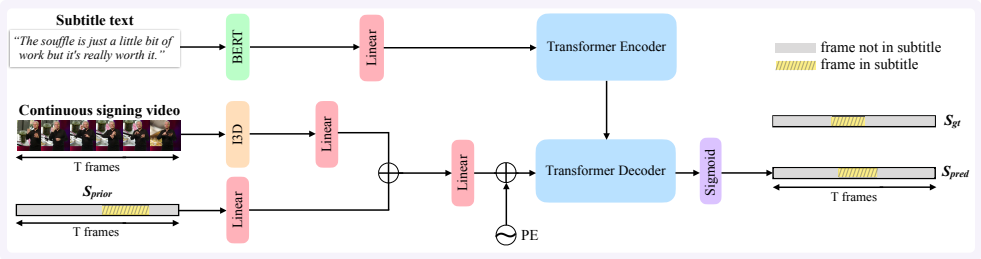


Figure 5.2: **SAT model overview:** We input to our model (i) token embeddings of the subtitle text we wish to align, (ii) a sequence of video features extracted from a continuous sign language video segment and (iii) the shifted temporal boundaries of the audio-aligned subtitle, S_{prior}. Using these inputs, the model outputs a vector of values between 0 and 1 of length T. Its first and last values above a threshold τ delimit the predicted temporal boundaries for the query subtitle. The location of the subtitle with respect to the window is represented in dashed yellow.

and HSP Trees [145]. Several studies have sought to employ subtitles as weak supervision for learning to localise and classify signs, using apriori mining [48] and multiple-instance learning [22, 23, 150]. More recent work has leveraged cues such as mouthings [2] and visual dictionaries [135] and by making use of deep neural network features with sliding window classifiers [119] and attention learned via a proxy translation task [187]. In deviation from these works, our objective is to localise complete subtitle units, rather than individual signs.

Temporal localisation of sign sequences. The alignment of subtitles to continuous signing was considered in creative early work by combining cues from multiple sparse correspondences [67], but under the assumption that ordering of words in subtitles are preserved in the signing (which does not hold in our problem setting). Other sequence-level sign language temporal localisation tasks that have received attention in the literature include category-agnostic sign segmentation [66, 155], active signer detection [42, 14, 138, 168] and diarisation [79, 78, 3]—each considers a temporal granularity that differs from subtitle units. In Chapter 3, we employ a keypoint-based model to segment continuous signing into sentence-like units without knowledge of the written subtitles during inference. Our approach relaxes this assumption and considers instead the practical scenario in which we assume access to the written subtitle to be aligned. We compare our approach with our approach in Chapter 3 in Sec. 5.4.

Continuous sign language recognition. Hybrid models coupling CNNs with HMMs [112, 113], attention mechanisms [94] and CTC losses [28, 41] have been studied for continuous sign language recognition, with recent extensions to

sequence-to-sequence models [29] and Transformers [31, 118] to tackle the task of sign language translation. These models produce either implicit or explicit alignments over a signing sequence corresponding to a sentence. However, these approaches have only been demonstrated to work on *pre-segmented* sentences of signing [29].

Aligning bodies of text to video. The Dynamic Time Warping (DTW) algorithm [142] has been applied to the problem of aligning sequences of movies to transcripts [65, 160] and plots synopses [183] using cues such as character recognition and subtitle content. It has also been successfully applied to the problem of aligning generic text descriptions against untrimmed video [13]. While effective, these methods require the preservation of sequence ordering across modalities, which does not hold in our problem setting. We nevertheless show in Sec. 5.3 how DTW can be used as a secondary stage of processing that resolves conflicting local alignments on the re-ordered subtitle prediction timings via a global objective. The fixed ordering assumption is relaxed by the work of [184], which aligns book chapters to video scenes. Their approach, however, which works through matching sparse character identifications against specific shots, is not applicable in our setting where shot boundaries do not provide a natural segmentation of the signing content.

Natural language grounding in videos. Our work is also related to the task of natural language grounding, which aims to locate a temporal segment within an untrimmed video sequence corresponding to a given natural language query. Existing methods have considered two-stage *propose and rank* approaches [88, 77, 124, 204], iterative grounding agents trained with reinforcement learning [87, 199] and single-stage regression models [211, 80, 40, 212]. Our proposed subtitle alignment task differs from natural language grounding in three ways: (i) The signing content is more *fine-grained*—the visual appearance of a signing sequence remains very similar across frames, necessitating nuanced recognition of body dynamics; (ii) Differently from language grounding, each subtitle to be aligned comes with its own reference location, providing an instance-specific prior over the start time and duration. As we show in Sec. 5.4, our effective use of this reference is important to achieving good performance, and our model is specifically designed to take advantage of this cue; (iii) Subtitles occupy mutually exclusive temporal regions, a property that we further exploit to improve alignment quality, but that does not hold in general for natural language grounding.

5.3 . Method

In this section, we describe our Transformer-based subtitle alignment model operating on a single subtitle and a short video segment (Sec. 5.3.1), our pretraining on sparse sign spottings (Sec. 5.3.2), and our final step that globally adjusts multiple subtitles in a long video using DTW (Sec. 5.3.3).

Problem formulation. As inputs to the model, we provide (i) token embeddings of the subtitle text we wish to align to signing, (ii) a sequence of video features extracted from a continuous sign language video segment, as well as (iii) prior estimates of the temporal boundaries for the given query, which we refer to as S_{prior}. The latter is provided as an approximate location and duration cue of the signing-aligned subtitle. Using these inputs, we predict a binary vector of the same length as the video features, where a consecutive sequence of 1s denotes the temporal location of the subtitle.

5.3.1 . Subtitle Aligner Transformer

The core of our model is a Transformer [188], as shown in Fig. 5.2, which we refer to as Subtitle Aligner Transformer (SAT). In contrast to the common approach of feeding video frames as input to the encoder [54, 34], we input the video frames to the *decoder* side in order for the model to learn the association between the frame-level features and the output vector of the same duration. We first describe the structure of the Transformer, and then the text and video feature extraction.

Encoder. The input to the encoder is a sequence of text embeddings corresponding to the subtitle we wish to align. Positional encodings are not used on the encoder side of the Transformer since the text embeddings (see below) already contain positional information. The encoder is a stack of Transformer layers, each containing a multi-head attention mechanism followed by a feed-forward network and embedding dimensionalities of size d_{model}.

Decoder. The decoder is a stack of Transformer layers that attend on the encoded sequence.[3] The input to the decoder consists of a sequence of video features encoding the visual signing information from the video, as well as a binary vector representing a prior estimate of the location of the signing-aligned subtitle (S_{prior}). Positional encodings are added to the decoder input in order for the model to exploit the temporal ordering of the signing. The final layer of the model is a linear layer with a sigmoid activation which outputs T predictions in the range $[0, 1]$ one for each video frame. Values of this output vector, S_{pred}, that are above a threshold τ correspond to the predicted temporal location of the queried subtitle text.

Text features. Each subtitle is encoded using a BERT [56] model, pretrained on a large text corpus with a masked language modelling task, to produce a sequence of 768-dimensional vectors, one for each token in the sentence. To match the input dimension of the encoder Transformer, these embeddings

[3] Note: There is no auto-regression.

are first linearly projected to d_{model}.

Video features. The visual features are 1024-dimensional embeddings extracted from the I3D [35] sign classification model made publicly available by the authors of [187]. The features are pre-extracted over sign language video segments. A visual feature sequence of length T is used as input to the model.

Prior position encoding. Besides the video features, the input to the decoder also includes a subtitle timing estimate as a prior position and duration cue. This prior estimate is encoded as a binary vector of length T, where 1 indicates that the associated video frame is within the temporal boundaries of the subtitle, and 0 otherwise. The video and prior inputs are fused via concatenation before being passed as input to the decoder. Before the concatenation both inputs are linearly projected to the same dimension. The fusion output is finally projected to d_{model} in order to be input to the Transformer decoder.

Training objective. The model is trained with a binary cross entropy loss between the predicted vector and the ground truth S_{gt} of the signing-aligned subtitle within the video segment:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} S_{gt}^{t} \log S_{pred}^{t} + (1 - S_{gt}^{t}) \log\big(1 - S_{pred}^{t}\big).$$

5.3.2 . Word Pretraining with Individual Sign Locations

SAT is designed for alignment of subtitles to video signing streams. However, the same architecture can be used without any alterations to align smaller text units, e.g. single words. Given that we have access to sparse sign annotations from mouthings [2] and dictionary exemplars [135], we can use these to initialise the model weights and incorporate this knowledge via a potentially easier single-sign spotting task. We obtain timings of the sparse word-level annotations and assume a fixed single-second width as the precise sign boundaries are not available. The model is then trained to spot the single sign occurrence within a video window of size T. In our experiments, we demonstrate the advantages of such a pretraining strategy.

5.3.3 . Global Alignment with DTW

Our model does not take into account global information from the length of the video (e.g. 1-hour), rather it looks for signing associated to a given subtitle within a short temporal window T (e.g. 20-seconds). Hence, there may be overlaps between predictions for different subtitles; we resolve these overlap conflicts using DTW [142]. We find an order-preserving global alignment from all elements of a sequence of video frames to all elements of sequence of subtitles, maximising the sum of sigmoid outputs of our model in our cost function for each subtitle query.

As DTW aligns all frames in a video sequence to subtitles, we select all frames of the signing video which are likely to be associated with subtitle quer-

ies. Specifically, we select all frames associated to an output score over τ_{dtw}. In the case where our model outputs only values below τ_{dtw} for a particular subtitle, we instead select all frames within the prior location S_{prior}.

We order the subtitles by the mid-point of their predicted temporal location. This allows the predicted subtitles to follow a different order to the original subtitles, because the order of phrases in the sign language interpretation does not necessarily follow the order of phrases of the written English subtitles (see Sec. 5.4.6 for further details).

We construct a cost matrix of dimension (i) the number of frames by (ii) the number of subtitles, and with entries of $1-p_{ij}$, where p_{ij} is the sigmoid output corresponding to frame i with subtitle j as the encoder input. We apply the DTW algorithm to this cost matrix of aligning video frames to subtitles. This maximises the overall sum of the sigmoid outputs of the model under the ordering and allocation constraints of DTW.

If not otherwise mentioned, our full SAT model uses DTW postprocessing.

5.4 . Experiments

In this section, we first give implementation details (Sec. 5.4.1) and describe the datasets and evaluation metrics used in this work (Sec. 5.4.2). We then compare the results of the proposed SAT model against strong baselines (Sec. 5.4.3) and present a series of ablation studies (Sec. 5.4.4). Next, we demonstrate the performance of our model on additional datasets (Sec. 5.4.5). Finally, we provide qualitative results and discuss limitations (Sec. 5.4.6).

5.4.1 . Implementation Details

Architecture. For both the encoder and the decoder we use 2 identical Transformer layers with 2 heads and size $d_{model} = 512$ each.

Backbone pretraining. The I3D model is pretrained to perform 1064-way classification across the sign spotting instances with mouthings [2] and dictionary exemplars [135] (further details can be found in [187]). The model is then frozen and used to densely pre-extract visual features with stride 1 over the clips of the datasets.

Prior input selection. As the prior estimate input S_{prior} we use the temporal location of the audio-aligned subtitle S_{audio} shifted by +3.2 seconds. This value, which we denote with S^{+}_{audio}, corresponds to the average temporal shift between the audio-aligned subtitles S_{audio} and the ground truth subtitles S_{gt} in our training data (see Fig. 5.3a).

Search windows. During training, we randomly select a search window of 20 seconds around the location of the ground truth subtitle S_{gt}, select the densely extracted video features for this window, and temporally subsample them by a factor of 4. All videos are sampled at 25 FPS, therefore this results

in $T = 125$ frames. During testing, we select a search window of the same length centered around the shifted subtitle location S^{+}_{audio}. An ablation study on the window size can be found in Tab. 5.12.

Text augmentation. During training, we augment the text query inputs randomly to reduce overfitting. For 50% of the samples, we shuffle the word order and add or delete up to two words.

Text embeddings. For the text embeddings, we use a pretrained BERT model from HuggingFace[4] with a standard architecture of 12-layers, 12-heads and 768 model size. The model is pretrained on BookCorpus[5] and English Wikipedia[6].

Positional encodings. For the input to the video encoder, we use 512-dimensional sinusoidal positional encodings as in [188]. The positional encodings are added to the video features before feeding to the Transformer.

Output thresholding. The output of our model is a temporal sequence of predictions between 0 and 1. For the single-subtitle SAT model, we consider the start of the subtitle to be the first time when the prediction is above $\tau = 0.5$ and the end of the subtitle to be the last time when the prediction is above $\tau = 0.5$ in the search window. When we apply a global alignment step with DTW to correct for overlapping subtitles, we no longer use these thresholds, but rather the temporal sequence of predictions between 0 and 1.

Parameters. We set thresholds τ to 0.5, τ_{dtw} to 0.4. We use the Adam optimiser with a batch size of 64. We train with a learning rate of 10^{-5} at the word-pretraining stage, and of 5×10^{-6} at finetuning with subtitles. At the word pretraining stage, the model is trained over 5 epochs. In one epoch of word pretraining, there are approximately 700K sign instances (including sign spotting both with mouthings and dictionaries). At this point the word alignment model obtains a frame-level accuracy of 30.38% and F1@50 of 40.75% on the 1630 sign instances of the test set episodes. During full-sentence finetuning, the model is trained over 80 epochs.

5.4.2 . Data and Evaluation Metrics

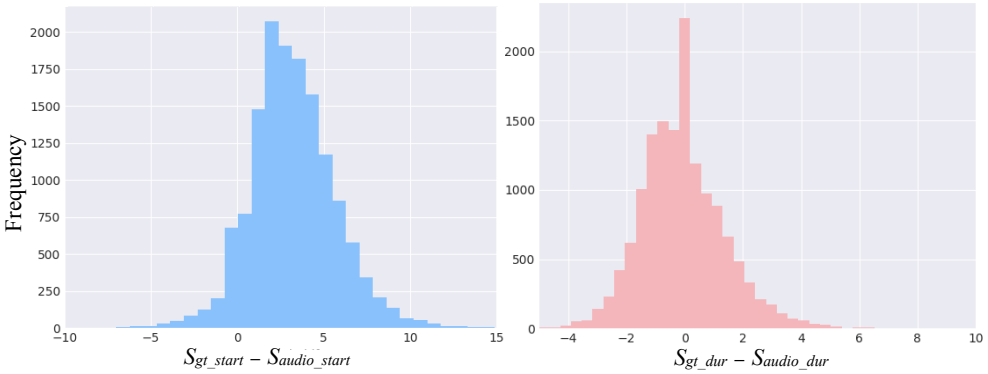
Statistics on the number of videos, hours, subtitles and vocabulary of each of our training and evaluation datasets are provided in Tab. 5.1.

BSL-1K$_{aligned}$ is a subset of BSL-1K [2], covering 24 different television programmes (food, nature, travel and lifestyle documentaries). The subtitles were originally aligned to the audio, but we have manually aligned them to the signing. The unaligned subtitles (i.e. those that are synchronised with the audio track, rather than the signing) differ from the signing-aligned subtitles in both start time and duration. In particular, Fig. 5.3, shows that there is no fixed shift or temporal scaling that can be consistently applied to transform audio-synchronised subtitles to their signing-aligned counterparts. We note

[4] https://huggingface.co/bert-base-uncased

[5] https://yknzhu.wixsite.com/mbweb

[6] https://en.wikipedia.org



(a) (b)

Figure 5.3: **S_{gt} vs. S_{audio}**: We plot the distribution of temporal shifts between ground-truth (S_{gt}) and audio-aligned (S_{audio}) subtitles on the training split of the BSL-1K$_{aligned}$ dataset by showing the differences in subtitle (a) start times and (b) duration. We observe the difficulty of the subtitle alignment task: (i) there is no fixed shift between ground-truth and audio-aligned subtitle timings, and (ii) the subtitle duration varies between spoken and signed languages.

		#vids.	#hours	#subs	#inst.	Vocab.	OOV
BSL-1K$_{aligned}$	Train	20	14.4	13.8K	128.1K	8.6K	\
	Test	4	3.3	2.0K	18.6K	2.8K	0.7K
BSL Corpus	Train	191	22.9	33.7K	261.5K	7.5K	\
	Val	15	1.5	2.6K	18.1K	1.8K	0.2K
	Test	21	2.6	3.8K	27.3K	2.4K	0.4K
BOBSL	Train	16	13	9.2K	91.5K	8.9K	\
(manually aligned)	Val	4	3	2.0K	23.6K	3.5K	1.1K
	Test	36	31	20.9K	248.9K	13.6K	8.0K

Table 5.1: **Datasets:** number of videos, hours, subtitles, word instances, vocabulary size and number of out-of-vocabulary (OOV) words.

that the differences exhibit an approximately Gaussian distribution, with the exception of an accentuated peak at 0 in Fig. 5.3b; we attribute this to the fact that if the duration of the subtitle is approximately correct, annotators tend to not further refine the boundaries.

The training set contains 7 cooking, 9 food-related travel, 1 environment-related travel and 3 lifestyle documentary shows. The test set contains 2 nature and 2 cooking shows. The 4 test episodes are chosen to evaluate the alignment model in different settings: seen/unseen signer and seen/unseen programme genre (which affects the number of out-of-vocabulary words) as shown in Tab. 5.2.

The signing-aligned subtitles were annotated by one deaf native BSL signer and a random subset was verified by another deaf native BSL signer, taking around 200 hours for the 24 episodes. The instruction was to shift the start and end times of each subtitle to correspond to the signing using the open-source VIA tool [61]. The process was refined over several iterations, incorporating annotator feedback. A handful of subtitles were excluded due to annotation uncertainty.

	#vids.	#hours	#subs	#inst.	Vocab.	OOV
Train	20	14.4	13.8K	128.1K	8.6K	\
Test (total)	4	3.3	2.0K	18.6K	2.8K	726
signer$_{seen}$, genre$_{seen}$	1	0.7	648	6.1K	1.3K	188
signer$_{seen}$, genre$_{unseen}$	1	0.9	465	4.1K	1.0K	233
signer$_{unseen}$, genre$_{seen}$	1	0.7	506	5.6K	1.1K	99
signer$_{unseen}$, genre$_{unseen}$	1	1.0	360	2.8K	882	234

Table 5.2: **BSL-1K$_{aligned}$:** The test set videos were chosen to evaluate performance on episodes with either signers or genre unseen during training.

BSL Corpus [163, 164] is a public dataset of videos of deaf signers gathered from several regions across the UK and accompanied by a variety of linguistic annotations. Unlike BSL-1K, the subtitles in this dataset are aligned to signing, and the translation direction is from sign language to English. We therefore simulate unaligned data by perturbing the subtitle locations in our experiments.

For our task, we employ the *FreeTranslation* annotation tier, which provides written English subtitles to accompany portions of the *Conversation* and *Interview* subsets of the corpus. In total, the annotations cover a total of 227 videos after cropping to include a single signer. Of these, 141 are sourced from the *Interview* subset and 86 videos are sourced from the *Conversation* subset. For consistency with prior work, we follow the train, validation and test partition employed by [154, 2]. However, since this partition does not

fully span the dataset, we add any dataset instances that were not present in the partition to the training set.

BOBSL is a dataset similar to BSL-1K$_{aligned}$ in style and content. See Chapter 4 for a detailed description of BOBSL and the manually aligned episodes. The test set contains 36 videos, almost all of which are factual documentaries related to nature, science and the environment. There are also a handful of food-related shows.

Evaluation metrics. We consider two main evaluation metrics: (i) frame-level accuracy, and (ii) $F1$-score. For the $F1$-score, hits and misses of subtitle alignment to sign language video are counted under three temporal overlap thresholds (IoU $\in \{0.1, 0.25, 0.50\}$) between predicted S$_{pred}$ and manually aligned S$_{gt}$ subtitles, denoted as F1@.10, F1@.25, F1@.50, respectively.

5.4.3 . Comparison to Baselines

Simple temporal shift baseline (S$^{+}_{audio}$). As a first baseline we use the shifted audio-aligned subtitles S$^{+}_{audio}$. Only a third of the shifted-audio subtitles S$^{+}_{audio}$ have more than 50% overlap (IoU) with the ground truth aligned subtitles.

Prosodic cues baseline (Chapter 3). We compare to the method used in Chapter 3, which is a model based on 2D body keypoints. In contrast to our framework, this method only uses visual prosodic cues and does not use semantic information from the query subtitle. It has been trained on a large-scale sign language corpus with aligned subtitles, and the pretrained model is public. The model consists of ST-GCN [206] and BiLSTM layers and segments sign language video into subtitle units. However, this is a different task than alignment, i.e. segments have no correspondence to subtitles. To obtain an association from each predicted segment to a subtitle, we align the shifted subtitles S$^{+}_{audio}$ to a subtitle-unit segmentation of Chapter 3 using DTW, where the cost of alignment is the temporal distance.

Heuristic baseline based on sparse sign spottings. Inspired by previous works that approached the alignment task through sparse correspondences [67], we implement a heuristic approach to align the subtitles using a combination of sign spotting and active signer detection. Sign spotting, performed by [2, 135], searches in the temporal vicinity of each audio-synchronised subtitle (the search window is constructed by padding the original subtitle by four seconds at each end) for individual sign instances corresponding to words that appear in the subtitle. From these sparse sign localisations, we perform subtitle alignment in four stages. First, we segment the episode into sequences that contain active signing, following [3]. Second, for any subtitle containing words that were spotted in the signing (assigned a posterior probability of 0.8 or greater by the model of [135]), we shift the subtitle such that its centre falls on the mean position of the spotted signs. Third, we transform all subtitles without spottings by affine transformations such that they fall within the "gaps" between those subtitles that contained

Method	frame-acc	F1@.10	F1@.25	F1@.50
S_{audio}	44.67	45.82	30.51	12.57
S^{+}_{audio}	60.76	71.69	60.74	36.10
Sign-spotting heuristics	61.71	69.23	59.60	36.04
Chapter 3	62.14	73.93	64.25	38.16
SAT (random subtitle)	65.52	70.30	60.36	40.04
SAT w/out DTW	65.81	74.32	64.69	41.27
SAT	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.3: **Comparison to baselines:** We show significant improvements by training a Subtitle Aligner Transformer (SAT) over several baselines. Moreover, providing a random subtitle as the text input results in poor performance, demonstrating that our model does indeed rely on token embeddings, and does not simply learn prosodic cues to align the subtitles. We obtain a further boost by correcting the overlaps of our predicted subtitles using DTW.

spotted signs, while preserving ordering (we use one such transformation per gap). Finally, we expand the duration of subtitles locally (applying a single scaling factor to each subtitle) in left to right ordering, such that they maximally fill the active signing segments predicted by the first stage. We note that only 15% of the subtitles in our test set can be confidently associated to a sign spotting, therefore relying only on sign localisation is expected to be insufficient for subtitle alignment.

A comparison of our model to the above baselines is given in Tab. 5.3. The simple temporal shift baseline and the heuristic baseline based on sparse sign spottings perform similarly, but are a significant improvement over the non-shifted subtitles S_{audio}. Using prosodic cues through the model in Chapter 3 results in a slight improvement over these two baselines. Our model significantly outperforms all baselines by exploiting the subtitle text to find the associated video segment. Indeed, when providing random subtitle text during training, our model is forced to rely on prosodic cues and fails to outperform the baseline F1 scores. Using DTW to resolve overlaps in predicted subtitles boosts our model performance.

5.4.4 . Ablation Study

We ablate the effects of inputting the prior estimate $S_{prior} = S^{+}_{audio}$ to the model, modifying the text input to the encoder, pretraining on sign localisation, using alternative model formulations, changing the number of attention heads, evaluating on seen and unseen signers and genres, training on different amounts of data, using different text encodings, changing the duration

Additional input	frame-acc	F1@.10	F1@.25	F1@.50
w/out S_{audio}	61.37	59.03	49.35	30.66
w/ S_{audio}	67.81	74.69	66.53	45.10
w/ S^{+}_{audio} 3.2-sec shift	**68.72**	**77.80**	**69.29**	**48.15**
w/ S_{audio} centre position	61.40	58.07	51.13	35.01
w/ S^{+}_{audio} rand. duration	68.61	75.10	66.84	46.72

Table 5.4: **Inputting S_{prior} variants:** Without information on the approximate position and duration of the subtitle, our model fails to improve upon our baseline methods. In particular, when setting the input S_{prior} to be systematically in the centre of the search window and with the duration of S_{audio}, model performance is poor. When using S^{+}_{audio} in its correct location in the search window, but varying the duration randomly of up to 2s, performance is relatively high. This suggests that position is a stronger cue than duration.

of the search window, shifting the search window and the prior estimates, as well as sampling the prior estimate from a Gaussian distribution during training.

Knowledge of S_{prior}. We experiment with several versions of inputs as additional information to the alignment task. Tab. 5.4 summarises the results. We first observe a significant drop in performance when S_{prior} is not provided (48.15 vs 30.66 F1@.50), suggesting that the position and duration of the corresponding audio content allows an approximate localisation cue, enabling the model to refine this via a series of attention layers. Inputting the 3.2 seconds shifted subtitle timings ($S_{prior} = S^{+}_{audio}$) performs better than inputting the audio-aligned subtitle timings ($S_{prior} = S_{audio}$). Nevertheless, our model still performs well when the average subtitle lag is unknown and the audio-aligned subtitle timings are used. Moreover, we carry out two additional experiments to investigate whether this cue is more important for providing a position prior or a duration prior. First, we always input the subtitle timing centred with respect to the search window. The poor performance of this model suggests the importance of the position. Second, we preserve the shifted location, but randomly change the input subtitle duration at training time by up to 2s. This slightly reduces the performance, therefore we infer that duration cues are less essential for the model than location cues.

Effect of text input to the encoder. We perform a series of ablations regarding the text encoding, including: no text augmentations, adding extra positional encodings to the BERT text features, and using the sentence embedding only (the output embedding corresponding to the BERT “CLS" token)

Method	frame-acc	F1@.10	F1@.25	F1@.50
w/o augmentations	67.35	75.72	66.85	45.31
w/ augmentations	**68.72**	**77.80**	**69.29**	**48.15**
w/ aug. + positional enc.	68.21	74.89	67.14	46.36
w/ aug. sentence emb.	66.18	72.99	63.71	41.71

Table 5.5: **Text ablations:** Our best model uses word embeddings without positional encodings as well as text augmentations during training (shuffling words in 50% of the subtitles, adding and deleting up to 2 words). Adding positional encodings to the BERT text features does not improve our model. Using sentence embeddings instead of token embeddings for the subtitle query degrades performance.

Pretraining	frame-acc	F1@.10	F1@.25	F1@.50
w/o word pretraining	67.26	76.18	66.19	42.47
w/ word pretraining	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.6: **Pretraining for sign localisation:** By pretraining our model to locate individual words within a given temporal window, we boost performance of subtitle alignment.

instead of the sequence of individual token embeddings. Tab. 5.5 presents the results on BSL-1K$_{aligned}$ with these text ablations. Augmenting the subtitle text improves performance, while adding extra positional encodings or using the sentence embedding degrades performance.

Effect of sign localisation pretraining. As explained in Sec. 5.3.2, we initially pretrain our model for temporal localisation of individual signs on a large set of word-video training pairs. In Tab. 5.6, we measure the effect of this pretraining and conclude that it provides a good initialisation for finetuning on long subtitles.

Model formulation. We consider an alternative version of the Transformer model, inspired by the DETR model in [34] for object detection in images. This model inputs image features into the Transformer encoder and text query into the Transformer decoder. Similarly, we input the sign language video features into the Transformer encoder. On the decoder side, we input the subtitle text features as well as either (i) the start and end times or (ii) the shift and scale of the shifted subtitles S^{+}_{audio} relative to the temporal window. We then consider the problem of subtitle alignment as a regression problem, and aim to predict (i) the start and end times or (ii) the shift and scale of the subtitle relative to the temporal window. As a further ablation, we also consider the same model architecture (with subtitle features and the start and end times as decoder input), but outputting a fixed binary vector of length T, which we

Prior input	Loss	frame-acc	F1@.10	F1@.25	F1@.50
shift/scale	shift/scale regress.	59.23	70.55	59.00	33.71
start/end	start/end regress.	60.04	72.20	60.41	34.33
start/end	binary classif.	60.48	74.05	62.75	35.07
binary	binary classif. (SAT)	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.7: **Model formulation:** We present an ablation where we experiment with a DETR-style Transformer model [34]. Video features are inputs to the Transformer encoder, and the subtitle query is fed to the Transformer decoder. Moreover, on the decoder side, we input either the start and end times or the shift and scale of the shifted subtitles S^+_{audio} relative to the temporal window, and use a regression model to predict the true values. This model fails to produce satisfactory results. Changing the regression model to a classification one by instead predicting a binary vector of length T (as in the SAT model) results in a small improvement; however SAT outperforms all the alternative models with a large margin.

No. heads	frame-acc	F1@.10	F1@.25	F1@.50
1	66.00	75.35	66.13	44.08
2	**68.72**	**77.80**	**69.29**	**48.15**
4	67.99	75.50	67.60	46.97

Table 5.8: **Number of attention heads:** We choose 2-head attention for our final model.

train with a binary classification objective (as in SAT).

The results in Tab. 5.7 suggest that our proposed approach with video features as input to the Transformer decoder enables significantly better learning, perhaps by providing a one-to-one mapping between video inputs and the frame-wise outputs. Another possible explanation for our proposed model's superiority is that it outputs alignment scores between subtitles and individual frames which allows for better conflict resolution strategies for overlapping subtitle predictions.

Number of attention heads. In Tab. 5.8, we ablate 1, 2 and 4 attention heads. We conclude that the model with 2-head attention performs best.

Performance on unseen signers/genres. Tab. 5.9 shows the SAT model results by test set episode. Our model tends to result in larger improvements over the S^+_{audio} baseline for signers seen in the training episodes, but still outperforms the S^+_{audio} baseline for unseen signers in unseen genres. More training data would be needed to better generalise to unseen signers.

Test episode signer	genre	Method	frame-acc	F1@.10	F1@.25	F1@.50
seen	*seen*	S^+_{audio}	45.48	66.92	55.02	31.84
		SAT	**60.23**	**77.74**	**68.47**	**49.00**
seen	*unseen*	S^+_{audio}	64.31	74.84	64.73	34.19
		SAT	**72.56**	**81.29**	**74.19**	**52.47**
unseen	*seen*	S^+_{audio}	56.30	**80.79**	69.70	44.95
		SAT	**63.68**	80.32	**72.40**	**52.82**
unseen	*unseen*	S^+_{audio}	71.84	63.29	53.16	33.76
		SAT	**74.93**	**69.76**	**59.92**	**34.32**

Table 5.9: **Performance breakdown by test episode:** Our model improves upon the S^+_{audio} baseline for all the combinations of seen/unseen for signer and genre. The improvements however are greater in the test episodes where the signer has been seen during training.

Amount of training data. By increasing the amount of training data, we improve performance of our model on the test set. Tab. 5.10 shows our results when training on random subsets of 25%, 50% and 75% of the videos in our training data. For subset selection, we randomly sample 4 times, and report the average performance across 4 trainings, as well as the standard deviation.

#training videos	frame-acc	F1@.10	F1@.25	F1@.50
5	$66.62^{\pm 0.16}$	$75.55^{\pm 0.86}$	$66.04^{\pm 1.09}$	$43.24^{\pm 0.81}$
10	$67.40^{\pm 0.28}$	$75.74^{\pm 0.25}$	$66.60^{\pm 0.25}$	$45.41^{\pm 0.88}$
15	$67.71^{\pm 0.23}$	$75.24^{\pm 0.43}$	$66.29^{\pm 0.84}$	$46.16^{\pm 0.66}$
20	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.10: **Amount of training data:** We train with a subset of our videos, using 5, 10, or 15 episodes instead of the total 20 episodes available. We observe increased performance as we increase the training size.

Text encoding choice. We experiment with word2vec [131] encodings for subtitle words instead of BERT. We use the pretrained word2vec model from [130], forming sentence embeddings by max pooling the encodings of all words over the channel dimension. In Tab. 5.11, we see that this results in lower performance compared to using the BERT encodings. We hypothesize that this is due to word2vec using a limited vocabulary, ignoring word order, and lacking the large-scale pretraining of the BERT model.

Method	frame-acc	F1@.10	F1@.25	F1@.50
word2vec	67.16	74.59	64.96	42.06
BERT	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.11: **Text encoding:** We experiment with word2vec encodings instead of BERT to embed words in the subtitle.

Window size	frame-acc	F1@.10	F1@.25	F1@.50
8 sec	66.98	73.12	64.66	44.13
12 sec	68.63	75.52	67.56	47.29
16 sec	68.51	76.18	68.63	48.10
20 sec	**68.72**	**77.80**	**69.29**	**48.15**

Table 5.12: **Search window size T:** We vary T between 50 and 125 frames (corresponding to 8- and 20-second inputs, respectively). Larger windows tend to perform better, possibly due to increased contextual information and the fact that the difference between S_{audio} and the aligned subtitle S_{gt} can be in the order of 10s.

Size of the search window T. In Tab. 5.12, we report the performance against different choices for input duration T We conclude that larger search windows generally improve performance, at the cost of computational complexity. This might be due to increased supervision, since with larger windows the training sees more negative examples, as well as due to better coverage at test time. A too short window size inhibits recovery of the correct location, if the correct location falls outside of the window boundaries.

Sensitivity analysis. During inference, we predict the location of a subtitle within a 20 second search window surrounding the location of S^{+}_{audio}. In order to analyse the sensitivity of the choice of search window, we shift the window by 1s, 3s and 5s at inference time. Tab. 5.13 shows that the choice of window within a margin of a few seconds does not have a large impact on the results.

However, if we keep the position of the search window constant and change the position of the prior estimate S^{+}_{audio}, then this has a significant effect on results. Tab. 5.14 shows the results of an experiment where we shift the prior estimate S^{+}_{audio} by 1s, 3s and 5s at inference time. The performance degrades when the model is given a worse prior as input, i.e., shifting S^{+}_{audio}.

Sampling the prior estimate. We consider an alternative choice of prior where we randomly sample S_{audio} during training from a Gaussian distribution with sample mean (3.2s) and standard deviation (3.6s) of the difference between the start of S_{gt} and S_{audio}. This choice seems equally valid in comparison to our original prior, which shifts S_{audio} by the estimated mean of 3.2s.

Shift window	frame-acc	F1@.10	F1@.25	F1@.50
0s	**68.72**	**77.80**	**69.29**	48.15
1s	68.53	76.99	69.23	47.69
3s	68.53	76.99	68.32	47.90
5s	68.32	76.58	68.42	**48.50**

Table 5.13: **Shifting search window:** We shift the search window at inference time by 1s, 3s and 5s. This does not have a major impact on results.

Shift prior	frame-acc	F1@.10	F1@.25	F1@.50
0s	**68.72**	**77.80**	**69.29**	**48.15**
1s	68.26	75.77	67.36	45.67
3s	58.69	58.08	47.80	28.18
5s	46.11	35.49	26.21	12.52

Table 5.14: **Shifting prior estimate S^+_{audio}:** By shifting the location of the prior S^+_{audio} at inference time by respectively 1s, 3s and 5s, the performance degrades.

We obtain similar results, i.e. a slightly higher frame accuracy (69.15 vs 68.72), but slightly lower F1 scores ({F1@.10, F1@.25, F1@.50}={75.42 vs 77.80, 67.61 vs 69.29, 47.59 vs 48.15}).

5.4.5 . Performance on Different Datasets

We demonstrate our model's performance on two more datasets: the BSL Corpus [163, 164] and the BOBSL dataset, which we introduced in Chapter 4.

BSL Corpus. The subtitles in this dataset are aligned to the sign language, and so we randomly shift and scale the subtitles in order to create artificial training data. We then train our SAT model to learn the correct alignment of subtitles to video in the BSL Corpus. We train the model (i) without any pretraining, (ii) with only word pretraining (on BSL-1K) and (iii) with SAT pretraining on BSL-1K$_{aligned}$. We report results in Tab. 5.15.

At each subtitle, we apply a random shift following a normal distribution with standard deviation σ_{pos} and a random change of duration of the subtitle also following a normal distribution with standard deviation σ_{dur}. Tab. 5.15 shows that our model is able to partially recover the correct original alignment. Larger shifts make it more difficult for our model to recover the correct original alignment, but random changes in subtitle duration seems to have less effect. This is consistent with the results in Tab. 5.4, where changing the duration of S^+_{audio} does not greatly impact results. Word pretraining on BSL-

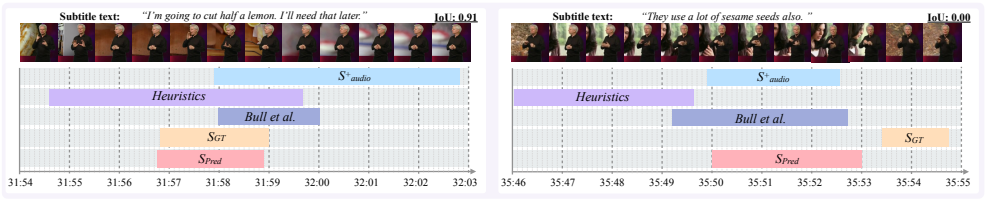


Figure 5.4: **Qualitative results:** This figure shows short time windows of 9s with shifted audio-aligned subtitles (S^{+}_{audio}), heuristic and Chapter 3 baselines ([26]), ground truth signing-aligned subtitles (S_{gt}) and our predicted signing-aligned subtitles (S_{pred}). Note that in practice, we input 20 seconds of video during training and testing as our search window. We depict shorter, "zoomed in" 9 second windows here for clearer visualisation. The right shows a failure case.

Rand. perturb. (σ_{pos}, σ_{dur})	Method	frame-acc	F1@.10	F1@.25	F1@.50
(3.5s, 1.5s)	Rand. shift & scale	63.24	37.13	26.54	12.47
	SAT w/out pretrain.	73.73	51.51	43.33	27.98
	SAT pretrain.	75.77	55.55	47.45	32.57
	SAT w/ word pretrain.	**76.29**	**57.65**	**50.35**	**34.54**
(4.5s, 1.5s)	Rand. shift & scale	60.18	29.52	20.61	10.00
	SAT pretrain.	73.69	48.41	41.34	28.06
	SAT w/ word pretrain.	**74.29**	**51.33**	**44.37**	**30.13**
(3.5s, 2s)	Rand. shift & scale	62.62	37.47	26.82	11.87
	SAT pretrain.	75.79	55.31	47.24	32.89
	SAT w/ word pretrain.	**76.00**	**57.86**	**50.43**	**33.79**

Table 5.15: **BSL Corpus:** We randomly shift and scale the correctly aligned subtitles in BSL Corpus to simulate unaligned data and then use our SAT model to recover the original correct alignments.

1K helps the model, but SAT pretraining on BSL-1K$_{aligned}$ does not. Word pretraining may help the SAT model recognise certain signs in BSL, but domain difference between BSL Corpus and BSL-1K$_{aligned}$ subtitles may explain why SAT pretraining on BSL-1K$_{aligned}$ does not lead to any significant gains on BSL Corpus.

BOBSL. The BOBSL test set allows us to evaluate our model on a larger and more diverse set of videos than the BSL-1K$_{aligned}$ test set. We report results in Tab. 5.16 and show further qualitative analysis in Fig. 5.9.

We follow the same procedure as for BSL-1K$_{aligned}$, but pre-train the model using the 1675 episodes in Sent-Train$_{SF}$ in addition to Sign-Train^{M,D}. We firstly pretrain SAT on word-level boundaries from Sign-Train^{M,D} with confidence scores above 0.8, where we predict a 1-second interval centred at the automatic mouthing or dictionary sign instance annotation in a randomly

chosen 20-second search window around the annotation. We do not input a prior alignment to the decoder. Secondly, we finetune this model using the sentence-level boundaries from Sent-Train$_{SF}$, where we use random shifts of these sentence-level boundaries of up to 3 seconds as a prior alignment. Thirdly, we further finetune the model on sentence-level boundaries from Sent-Train$_{H}$. We use 2.7-second shifted audio-aligned subtitles S^{+}_{audio} as a prior alignment, where 2.7 is the average lag between the audio-aligned and annotated signing-aligned sentences in the BOBSL training set (see Fig. 4.15). We apply additional random shifts of up to 2 seconds during training for data augmentation. When training on sentence-boundaries, we randomly select a search window of 20 seconds around the location of the prior alignment and filter to sentences longer than 0.5 seconds. We also randomly shuffle the words in 50% of the sentences and drop 15% of words during training as a data augmentation step.

We use the Adam optimiser with a batch size of 64. We train with a learning rate of 10^{-5} at the word-pretraining stage, 0.5×10^{-5} at finetuning with sentence-level boundaries from Sent-Train$_{SF}$ and 10^{-6} at finetuning with sentence-level boundaries from Sent-Train$_{H}$. At the word pretraining stage, the model is trained over 22 epochs. During the full-sentence finetuning on Sent-Train$_{SF}$ and Sent-Train$_{H}$, the model is trained over 44 and 143 epochs respectively.

We report the performance of baseline sign language alignment methods in Tab. 5.16 on Sent-Test: (i) the original audio-aligned subtitles (S_{audio}), (ii) the shifted (by +2.7 seconds) audio-aligned subtitles (S^{+}_{audio}) and (iii) SAT model. We observe that SAT performs best.

Method	frame-acc	F1@.10	F1@.25	F1@.50
S_{audio}	40.27	46.80	33.88	14.33
S^{+}_{audio}	62.33	73.01	64.28	44.75
SAT	**70.37**	**73.33**	**66.32**	**53.18**

Table 5.16: **Sign language alignment on Sent-Test.** We report baselines for sign language alignment on the 36 manually aligned episodes. We observe a significant improvement for SAT over the simpler baseline methods.

5.4.6 . Qualitative Analysis

Results on BSL-1K$_{aligned}$. Fig. 5.4 illustrates several test examples on BSL-1K$_{aligned}$. The timeline shows the ground truth alignment (S_{gt}), our prediction (S_{pred}), as well as the S^{+}_{audio} baseline, alongside a sample of video frames and the query subtitle text. While the shifted baseline S^{+}_{audio} provides an approx-

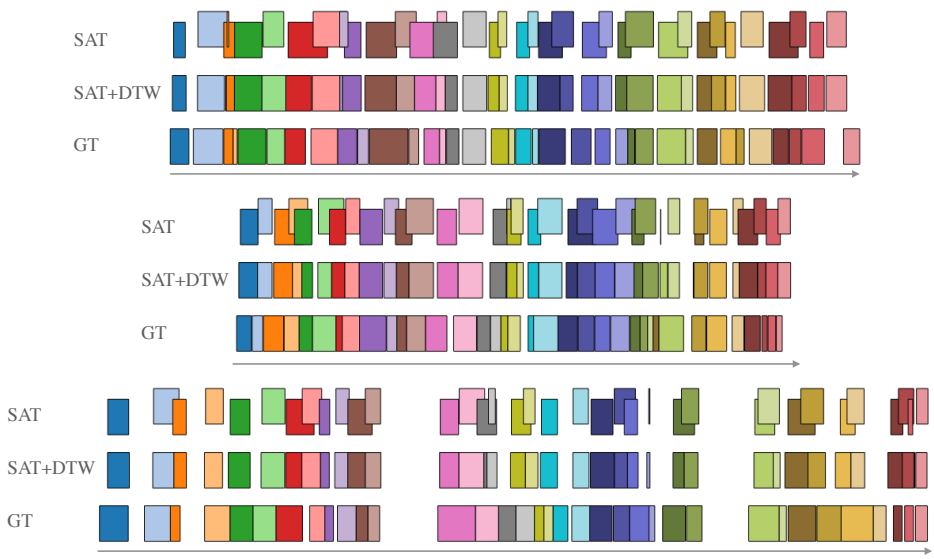


Figure 5.5: **DTW:** Our SAT model predicts the locations of subtitles independently of each other, and thus there can be overlaps in subtitle localisations. Using a global alignment step with DTW, we resolve these overlaps and improve performance.

imate position, it is largely unaligned. Our model effectively learns to attend to both visual and textual cues. A typical failure mode happens when the prior position encoding is significantly far from the ground truth (see Fig. 5.4 right).

Effect of global alignment with DTW. In Fig. 5.5, we present results before and after the global alignment with DTW on a long timeline. We observe that the single-subtitle Transformer model produces overlapping regions between consecutive subtitles which are resolved after the global DTW stage. Consequently, we see that the overall duration of subtitles decreases after DTW (see Fig. 5.6). During the DTW stage, we order subtitles by their predicted order, not by the original order of S_{audio}. Indeed, in BSL-1K$_{aligned}$, 1.6% of subtitles in S_{gt} do not respect the original order of S_{audio}. On the test set, 1.6% of subtitles in S_{pred} switch position with respect to S_{audio}.

Results on BSL-1K$_{aligned}$. Fig. 5.7 demonstrates qualitative results on BSL-1K$_{aligned}$.

Results on BSL Corpus. Fig. 5.8 demonstrates qualitative results on BSL Corpus.

Results on BOBSL. Fig. 5.9 demonstrates qualitative results on BOBSL.

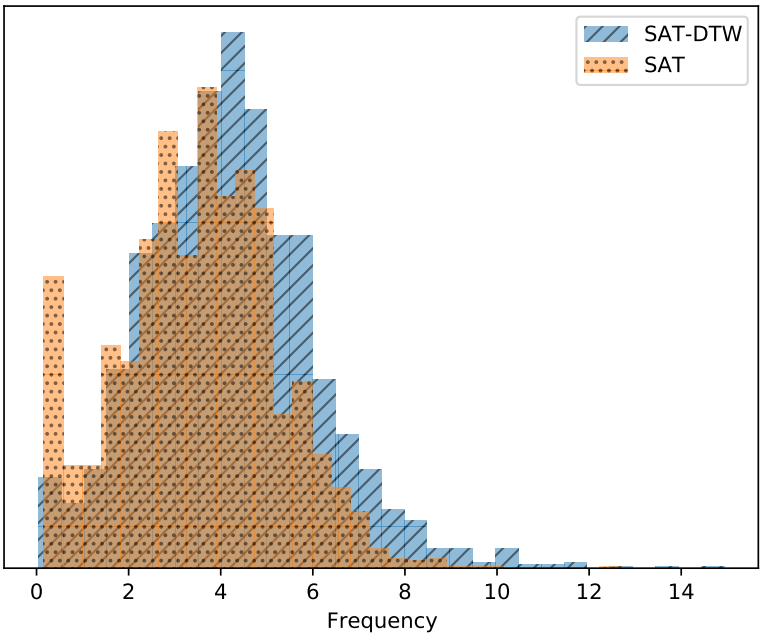


Figure 5.6: **Duration before and after DTW**: The median duration of S_{gt} is 3.3s. Before DTW, the median duration of our predicted subtitles is 4.1s, but after DTW the median duration is reduced back down to 3.5s by resolving conflicts in overlapping subtitles.

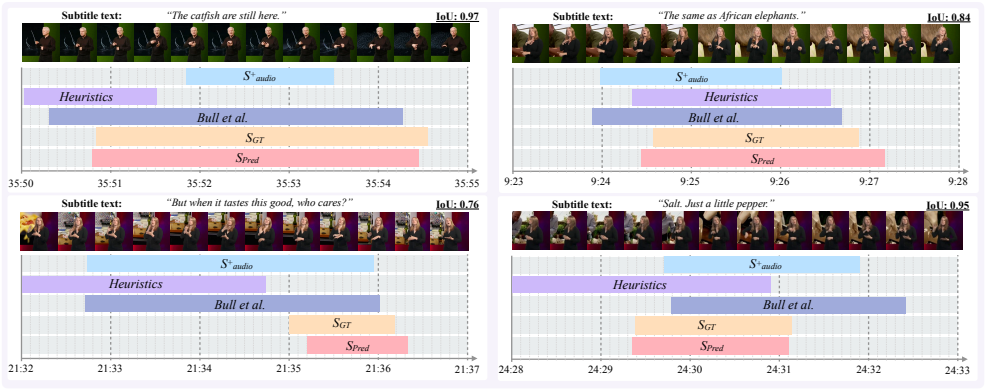


Figure 5.7: **Qualitative results on BSL-1K$_{aligned}$:** This figure shows short time windows of 5s with shifted audio-aligned subtitles (S^{+}_{audio}), ground truth signing-aligned subtitles (S_{gt}) and our predicted signing-aligned subtitles (S_{pred}). In practice, we input 20 seconds of video during training and testing as our search window.

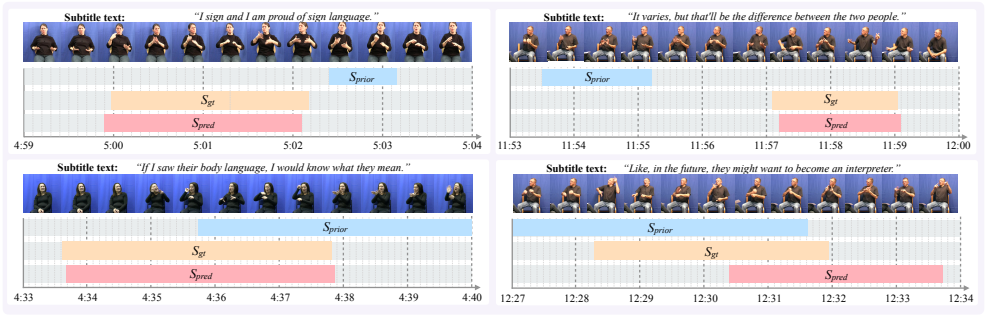


Figure 5.8: **Qualitative results on BSL Corpus:** This figure shows short time windows of 5s and 7s with shifted and rescaled subtitles (S_{prior}), ground truth aligned subtitles (S_{gt}) and our predicted subtitles (S_{pred}). In practice, we input 20 seconds of video during training and testing for our search window. The shifted and rescaled subtitles (S_{prior}) are created using a random shift with standard deviation of 3.5s and a random change in length of standard deviation 1.5s.

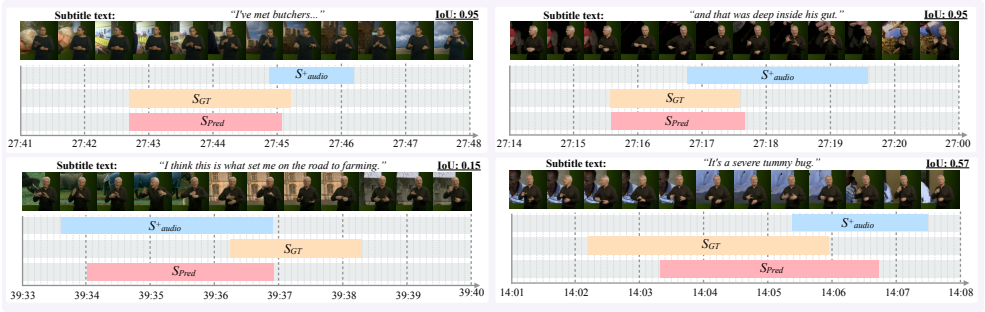


Figure 5.9: **Qualitative results on BOBSL:** This figure shows short time windows of 7s with shifted audio-aligned subtitles (S^{+}_{audio}), ground truth signing-aligned subtitles (S_{gt}) and our predicted signing-aligned subtitles (S_{pred}).

5.5 . Conclusion

We presented a Transformer-based approach to synchronise subtitles with sign language video content in interpreted data. We showed that knowledge of subtitle content is essential to effectively align subtitles to signing. We hope that our work will be a stepping stone to obtain video-subtitle pairs that allow training of unconstrained machine translation systems for sign languages.

In Chapter 6, we use our Subtitle Aligner Transformer to improve automatic annotation methods of lexical signs. By automatically aligning subtitles to sign language video, we narrow the search window in order to localise signs corresponding to words from the subtitle text.

6 - Dense Annotation of Sign Language Video

BOBSL (Chapter 4) is a large dataset consisting of sign language interpreted TV broadcasts, comprising (i) a video of continuous signing and (ii) subtitles corresponding to the audio content. However, one key challenge in the usability of such data is the lack of sign annotations. In Chapter 5, we provide an method for automatically aligning sentences to the sign content. In this chapter, we aim to localise individual words in the sign video.

We propose a simple, scalable framework to *vastly* increase the *density* of automatic annotations. Our contributions are the following: (1) we significantly improve previous annotation methods by making use of synonyms and subtitle-signing alignment from Chapter 5; (2) we show the value of pseudo-labelling from a sign recognition model as a way of sign spotting; (3) we propose a novel approach for increasing our annotations of *known* and *unknown* classes based on *in-domain exemplars*; (4) on the BOBSL BSL sign language corpus, we increase the number of confident automatic annotations from 670K to 5M. We make these annotations publicly available to support the sign language research community.

This chapter was published in ECCV 2022 [136]. I contributed to many different aspects of this chapter, including improving yield of annotations by using synonyms and subtitle-signing alignment, developing a new method of automatic annotation using exemplars and running experiments. All co-authors contributed to the writing and to the ideas developed in this work.

6.1 . Introduction

An important factor impeding progress in automatic sign language recognition – in contrast to automatic speech recognition – has been the lack of large-scale training data. To address this issue, researchers have recently made use of sign language interpreted TV broadcasts, comprising (i) a video of continuous signing, and (ii) subtitles corresponding to the audio content, to build datasets such as Content4All [27] (190 hours) and BOBSL, described in Chapter 4, (1460 hours).

Although such datasets are orders of magnitude larger than the long-standing benchmark RWTH-PHOENIX [29] (9 hours) and cover a much wider domain of discourse (not restricted to only weather news), the supervision they provide on the signed content is limited in that it is *weak* and *noisy*. It is weak because the subtitles are temporally aligned with the audio content and not necessarily with the signing. The supervision is also noisy because the presence of a word in the subtitle does not necessarily imply that the word is signed; and subtitles can be signed in different ways. Recent works

have shown that training automatic sign language translation models on such *weak* and *noisy* supervision leads to low performance [27, 187, 4].

In an attempt to increase the value of such interpreted datasets, multiple works [2, 135, 187] have leveraged the subtitles to perform lexical *sign spotting* in an approximately aligned continuous signing segment – where the aim is to determine *whether* and *when* a subtitle word is signed. Methods include using visual keyword spotting to identify signer mouthings [2], learning a joint embedding with sign language dictionary video clips [135], and exploiting the attention mechanism of a transformer translation model trained on weak, noisy subtitle-signing pairs [187]. These works leverage the approximate subtitle timings and subtitle content to significantly reduce the correspondence search space between temporal windows of signs and spoken language words. Although such methods are effective at automatically annotating signs, they only find *sparse* correspondences between keywords in the subtitle and individual signs.

Our goal in this chapter is to produce *dense* sign annotations, as shown in Fig. 6.2. We define densification in two ways: (i) reducing gaps in the timeline so that we have a densely spotted signing sequence; and also (ii) increasing the number of words we recall in the corresponding subtitle. This process can be seen as automatic annotation of lexical signs. Automatic dense annotation of large-vocabulary sign language videos has a large range of applications including: (i) *substantially* improving recall for retrieval or intelligent fast forwards of online sign language videos; (ii) enabling *large-scale* linguistic analysis between spoken and signed languages; (iii) providing *supervision* and *improved alignment* for continuous sign language recognition and translation systems.

In this chapter, we ask the following questions: (1) Can we improve current methods to improve the yield of automatic sign annotations whilst maintaining precision? (2) Can we increase the vocabulary of annotated signs over previous methods? (3) Can we 'fill in the gaps' that current spotting methods miss? The answer is yes, to all three questions, and we demonstrate this on the recently released BOBSL dataset of British Sign Language (BSL) signer interpreted video.

We make the following four contributions: (1) we significantly improve previous methods by making use of synonyms and subtitle-signing alignment; (2) we show the value of pseudo-labelling from a sign recognition model as a way of sign spotting; (3) we propose a novel approach for increasing our annotations of *known* and *unknown* sign classes based on in-domain exemplars; (4) we will make all 5 million automatic annotations publicly available to support the sign language research community. Our increased yield and vocabulary size is shown in Fig. 6.1. Our final vocabulary of 24.8K represents the vocabulary of English words (including named entities) from the subtitles which have been automatically associated to a sign instance; different words

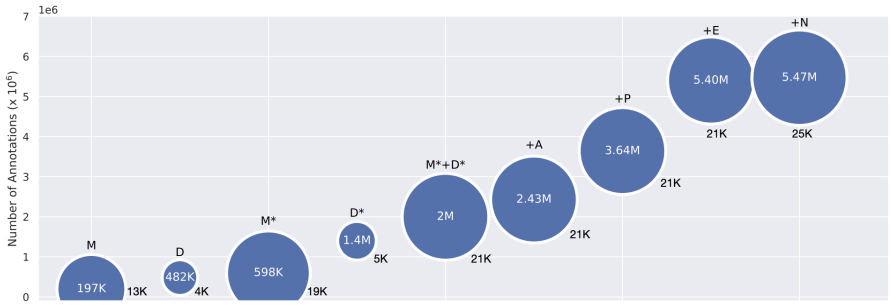


Figure 6.1: **Yield of automatic annotations and vocabulary size:** We highlight the increase in the number of automatic annotations and vocabulary size at each stage in our proposed framework. M, D, A refer to annotations from previous methods. M*, D*, P, E, N refer to new and improved annotations described later in this chapter. The number of annotations is shown within each circle. The vocabulary size is reported below each circle and also represented by the circle diameter.

may have the same sign.

We note that this chapter is focused on *interpreted* data, which can differ from *conversational* signing in terms of style, vocabulary and speed [20]. Although our long-term aim is to move to conversational signing, learning good representations of signs from interpreted data can be a 'stepping stone' in this direction. Moreover, non-lexical signs, such as pointing signs, depicting signs and spatially located signs, are essential elements of sign language, but our method is limited to the annotation of lexical signs associated to words in the text.

In this chapter, we firstly provide a summary of the relevant literature in Sec. 6.2, describe our densification methods in Sec. 6.3, present our experimental resuls in Sec. 6.4 and conclude in Sec. 6.6.

6.2 . Related Work

Our work relates to several themes which we give a brief overview of below.

Sign Spotting. One line of research has focused on the task of *sign spotting*, which seeks to detect signs from a given vocabulary in a target video. Early efforts for sign spotting employed lower-level features (colour histograms and geometric cues) in combination with Conditional Random Fields [207], Hidden Markov Models (HMMs) [191] and Sequential Interval Patterns [145] for temporal modelling. A related body of work has sought to localise signs while leveraging weak supervision from audio-aligned subtitles. These include the use of external dictionaries [119, 135, 95] and other localisation cues

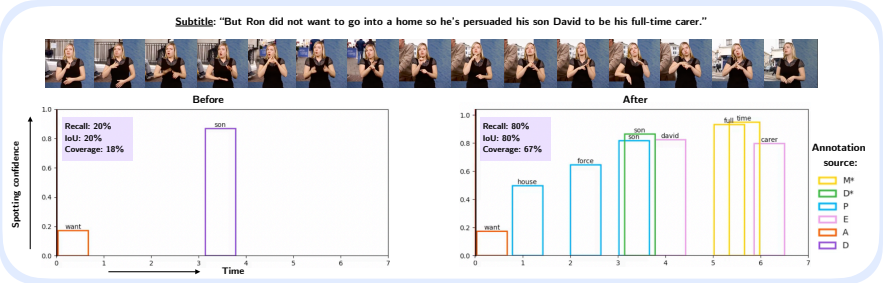


Figure 6.2: **Densification**: For continuous sign language, we show automatic sign annotation timelines, along with their confidence and annotation source, *before* and *after* our framework is applied. M, D, A refer to automatic annotations from previous methods from mouthings [2], dictionaries [135] and the Transformer attention [187]. M*, D*, P, E, N refer to new and improved automatic annotations collected in this work.

such as mouthing [2] and Transformer attention [187]. The performance of these approaches depends on the quality of the visual features, keywords, and the search window. In this work, we show improved yield of existing sign spotting techniques by employing automatic subtitle alignment techniques to adjust the time window and incorporating synonyms when forming the keywords. Going further beyond the spotting task explored in prior work, we use the automatic spottings to initiate additional algorithms for sign discovery based on *in-domain* exemplar matching (Sec. 6.3.1). This is similar to dictionary-based sign spotting techniques [135, 95] except we do not source the exemplars from external dictionaries, avoiding the domain gap issue. Besides in-domain *sign* exemplars as in [95], we explore the weak *subtitle* exemplars with unknown sign locations.

A recent progress in mouthing-based keyword spotting was presented by *Transpotter* [152]. This architecture comprises a transformer joint encoder of visual features and phoneme features that is trained to regress both the presence and location of the target keyword in a sequence from mouthing patterns. Preliminary small-scale experimental results reported by Prajwal et al. [152] demonstrated that Transpotter can perform visual keyword spotting in signing footage. Here, we showcase its suitability for the large-scale annotation regime, and further train it on sign language data to obtain a greater density of sign annotations.

In this work, we demonstrate the additional value of *pseudo-labelling* [209, 115] with a sign classifier as an effective mechanism for sign spotting. While pseudo-labelling has been explored previously for category-agnostic sign segmentation [155] and temporal alignment of glosses [113, 41] to the best of our knowledge, this is the first use of pseudo-labelling for sign spotting by directly leveraging the predictions of a sign classifier in combination with a pseudo-

label filter constructed from the subtitles themselves.

Sign Language Recognition. Efforts to develop visual systems for sign recognition stretch back to work in 1988 from Tamaura and Kawasaki [181], who sought to classify signs from hand location and motion features. There were later efforts to design hand-crafted features for sign recognition [38, 175, 193, 194, 144]. Deep convolutional neural networks then came to dominate sign representation [112], particularly via 3D convolutional architectures [102, 117, 2, 119] with extensions to focus model capacity around human skeletons [93] and non-manual features [90].

In the domain of continuous sign language recognition, in which the objective is to infer a sequence of sign glosses, prior work has explored HMMs [8, 111] in combination with Dynamic Time Warping (DTW) [213], RNNs [50] and architectures capable of learning effectively from CTC losses [215, 41]. Recently, sign representation learning methods inspired by BERT [56] have shown the potential to learn effective representations for both isolated [89] and continuous [216] recognition. Koller [110] provides an extensive survey of the sign recognition literature, highlighting the extremely limited supply of datasets with large-scale vocabularies suitable for continuous sign language recognition. In our work, we aim to take a step towards addressing this gap by developing "densification" techniques for constructing such datasets automatically.

Sign Language Translation. The task of translating sign language video to spoken language sentences was first tackled with neural machine translation by Camgöz et al. [29], who also introduced the PHOENIX-Weather-2014T dataset to facilitate research on this topic. Several frameworks have been proposed to employ transformers for this task [31, 210], with extensions to improve temporal modelling [118], multi-channel cues [30] and signer independence [97]. Related work has also sought to contribute to progress on this task by exploiting monolingual data [214] and gloss sequence synthesis [139, 120]. To date, various works have shown promise on the PHOENIX-Weather 2014T [29] and CSL Daily [214] benchmarks. However, sign language translation has not yet been demonstrated for a large vocabulary across multiple domains of discourse. Differently from the works above, this chapter focuses on developing methods that are applicable to large/open vocabulary regimes.

Weakly-supervised Object Discovery and Localisation. Our approach is also related to the rich body of literature on object cosegmentation [157, 100, 106, 158], weakly supervised object localisation [143, 55, 167, 197, 81], object colocalisation [182, 101] and unsupervised object discovery and localisation [43, 192]. Here, we propose an algorithm for discovering and localising novel signs (i.e. for which we have no labelled examples), but instead have weak supervision in the form of subtitles containing keywords of interest. Moving beyond initial work that sought to learn from subtitles in an aligned setting [67], classical approaches for sign discovery using subtitles have included Multiple Instance Learning where

the subtitles are considered as positive and negative bags for a particular keyword [22, 105, 150] and a priori mining [48]. Differently from these works, we first bootstrap our sign discovery process with sign spotting to both obtain initial candidates and learn robust sign representations, then propagate these examples across video data by leveraging the similarities between the resulting representations together with noisy constraints imposed by the subtitle content.

6.3 . Densification

Our goal is to leverage several ways of sign spotting to achieve dense annotation on continuous signing data. To this end, we introduce both new sources of automatic annotations, and also improve the existing sign spotting techniques. We start by presenting two new spotting methods using in-domain exemplars: to mine more sign instances with individual *exemplar signs* (Sec. 6.3.1) and to discover novel signs with weak *exemplar subtitles* (Sec. 6.3.2). We also show the value of pseudo-labelling from a sign recognition model for sign spotting (Sec. 6.3.3). We then describe key improvements to previous work which substantially increase the yield of automatic annotations (Sec. 6.3.4). Finally, we present our evaluation framework to measure the quality of our sign spottings in a large-vocabulary setting (Sec. 6.3.5). The contributions of each source of annotation are assessed in the experimental results.

6.3.1 . Mining more Spottings through In-domain Exemplars (E)

The key idea is: given a continuous signing video clip and a set of exemplar clips of a particular sign, we can use the exemplars to search for that sign within the video clip. In our case, the exemplars are obtained from other *automatic* spotting methods (M*, D*, A, P), described in Sec. 6.3.3 and Sec. 6.3.4, and come from the same *domain* of sign language interpreted data, i.e. the same training set. We hypothesise that signs from the same domain are more likely to be signed in a similar way and in turn help recognition; in contrast, for example, to signs from a different domain such as dictionaries.

Formally, suppose we have a reference video V_0 in which we wish to localise a particular sign w, whose corresponding word occurs in the subtitle. We also have N video exemplars $V_1, \ldots, V_N$ of the sign w. For each video, V_i, let $\mathcal{C}_i$ denote the set of possible temporal locations of the sign w and let $c = (f, p) \in \mathcal{C}_i$ denote a candidate with features f at temporal location p. We compute a score map between our reference video V_0 and each exemplar $V_1, \ldots, V_N$ by computing the cosine similarity between each feature at each position in $c_0 \in \mathcal{C}_0$ and $(c_1, c_2, \ldots c_n) \in \mathcal{C}_1 \times \cdots \times \mathcal{C}_N$. This results in N score maps of dimension $|\mathcal{C}_0| \times |\mathcal{C}_i|$ for $i = 1 \ldots N$. We then apply a max operation over the temporal dimension of the exemplars, giving us N vectors of length

$|\mathcal{C}_0|$, which we call $M_1, \ldots, M_N$.

We subsequently apply a voting scheme to find the location of the common sign w in V_0. Specifically, we let $L = \frac{1}{N}\sum_{i=1}^{N} 1_{(M_i > h)}$ for a threshold h, where the vector $1_{(M_i>h)}$ takes the value 1 for entries of M_i which are greater than h and 0 otherwise. The candidate location of w in V_0 is then $c = (f, p) \in \mathcal{C}_0$ where p corresponds to the position of the maximum non-zero entry in the vector L (see Fig. 6.3 for a visual illustration). If there are multiple maxima, we assign p to be the midpoint of the largest connected component. If all entries in L are zero, we conclude w is not present. We perform two variants of this approach using mean and max pooling of the score maps (instead of voting). We note that for a given signing sequence, we only focus on finding signs for words in the subtitle that have *not* been annotated by other methods.

We choose N video exemplars of spottings of signs that we wish to find in the reference video. For an exemplar sign, we choose 8 consecutive stride-4 features surrounding each spotting ($|\mathcal{C}_i| = 8$ for $i = 1 \ldots N$), where the features come from the last layer of the M* + D* + A [187] + P MLP model of Tab. 6.6 and are 256 dimensional. The values of N are shown in the fourth column of Tab. 6.5. For the reference video, we choose a subtitle with 2s padding on either side, and use stride-4 features as candidate locations of signs.

The methods 'avg' and 'max' noted in the fifth column of Tab. 6.5 are computed slightly differently to the method 'vote' described in Sec. 6.3.1. As before, we compute the cosine similarity between each feature at each position of the reference video $c_0 \in \mathcal{C}_0$ and each position of the spottings exemplars $(c_1, \ldots, c_n) \in \mathcal{C}_1 \times \cdots \times \mathcal{C}_N$. The cosine similarity is rescaled to the interval $[0, 1]$. This results in N score maps of dimension $|\mathcal{C}_0| \times |\mathcal{C}_i|$ for $i = 1 \ldots N$, which for us can be represented as a matrix $\mathcal{M}$ of dimension $|\mathcal{C}_0| \times 8 \times N$. We take either the average or the maximum value of $\mathcal{M}$ over the N exemplars to obtain a matrix $\mathcal{M}'$ of dimension $|\mathcal{C}_0| \times 8$. We then take the maximum of $|\mathcal{C}_0| \times 8$ across the exemplar temporal dimension to obtain a vector L of dimension $|\mathcal{C}_0|$. We consider the first element of L above a threshold h to be the corresponding sign in the reference video. For the version where we take the average value of $\mathcal{M}$ over the N exemplars, we let $h = 0.7$; for the version where we take the maximum value of $\mathcal{M}$ over the N exemplars, we let $h = 0.8$.

6.3.2 . Discovering Novel Sign Classes (N)

One limitation of our proposed method in Sec. 6.3.1 is that we are only able to collect more sign instances from a *closed* vocabulary, determined by sign exemplars obtained from other methods (described in Sec. 6.3.3 and Sec. 6.3.4). Here, we extend our approach to localise *novel* signs, for which we have no exemplar signs but whose corresponding word appears in the subtitle text. We follow our approach described in Sec. 6.3.1, computing score

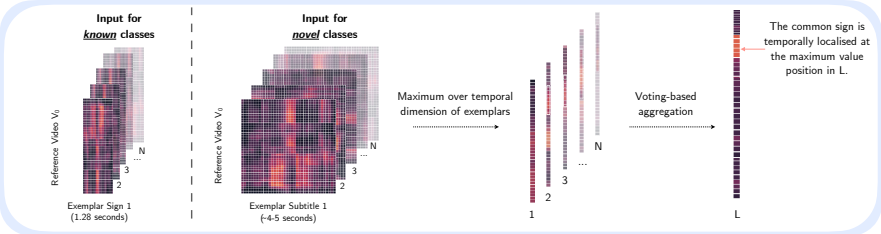


Figure 6.3: **Sign spotting through exemplars to find instances of known classes (E) and novel classes (N)**: By comparing a reference video V_0 to a set of exemplars (either sign exemplars for known sign class instances or weak subtitle exemplars for novel sign class instances), we can find the common lexical sign in the collection. We (1) form a set of score maps by calculating the cosine similarities between reference and exemplar representations; (2) we perform a maximum operation over the temporal dimension of exemplars; (3) we apply a voting-based aggregation to find the temporal location of the commmon sign in V_0. The duration of exemplar signs is fixed.

maps between our reference video and exemplar subtitles (instead of exemplar signs, see Fig. 6.3). We note that by 'exemplar subtitle', we are referring to the video frames corresponding to the subtitle timestamps. Non-lexical signs, such as pointing signs or pause gestures, are very common in sign language. To avoid annotating such non-lexical signs as the common sign across V_0 and $V_1, \ldots, V_N$, we also choose N^- negative subtitle exemplars $U_1 \ldots U_{N^-}$ presumed to not contain w (due to the absence of w in the subtitle). We compute L^+ and L^- using the score maps from positive exemplars $V_1, \ldots, V_N$ and negative exemplars $U_1, \ldots, U_{N^-}$ respectively. We then let $L = L^+ - L^-$.

In order to find a sign corresponding to a word w in a reference video, we take $N = 9$ positive exemplars corresponding to subtitles containing w, and $N' = 27$ negative exemplars corresponding to subtitles not containing w. We do not use padding around either the reference video nor the exemplars. The confidences for these spottings correspond to the proportion of the N exemplars with a cosine similarity match above a threshold h, i.e. the maximum value of L^+. We consider all novel sign classes with a confidence threshold above 0, that is, with at least one match amongst the positive exemplars.

6.3.3 . Pseudo-labelling as a Form of Sign Spotting (P)

We propose to re-purpose a pretrained large-vocabulary sign classification model (see vocabulary expansion in Sec. 6.3.5) for the task of sign spotting. Specifically, we predict a sign class from a fixed vocabulary for each time step in a continuous signing video clip. We subsequently filter the predicted signs to words which occur in the corresponding English subtitle. Similarly to [187], here the task is to recognise the sign from scratch, without a query keyword. The subtitle is only used as a post-processing step to filter out signs

which are less likely performed (due to absence in the subtitle).

6.3.4 . Improving the Old (M*, D*)

Here, we briefly describe our improvements over the existing sign spotting techniques.

Better Mouthings with an Upgraded KWS from Transpotter [152]. In previous work [4], an improved BiLSTM-based visual-only keyword spotting model of Stafylakis et al. [173] from [134] (named "P2G [173] baseline") is used to automatically annotate signs via mouthings. In this work, we make use of the recently proposed transformer-based *Transpotter* architecture [152], provided by the authors, that achieves state-of-the-art results in visual keyword spotting on lipreading datasets. We follow the procedure described in [2, 4] to query words in the subtitle in continuous signing video clips.

Finetuning KWS on Sign Language Data through Bootstrapping. The visual keyword spotting Transpotter architecture in [152] is trained on silent speech segments, which differ considerably from signer mouthings. In fact, signers do not mouth continuously and sometimes only partly mouth words [16]. In order to reduce this severe domain gap, we propose a dual-stage finetuning strategy. First, we extract high-confidence mouthing annotations using the pre-trained Transpotter from [152] on the BOBSL training data. We query for the words in the subtitle and obtain the temporal localization of the word in the video. We finetune on this pseudo-labeled data using the same training pipeline of [152], where the spotted mouthings (word-video pairs) act as positive samples. For the negative samples, we pair a given word with a randomly sampled video segment from the dataset. As we observe the Transpotter to predict a large number of false positives, we remedy this by sampling a larger number of negative pairs in each batch. We also do a second round of fine-tuning by training on the pseudo-labels from the finetuned model of the first stage. We did not achieve significant improvements with further iterations.

Better Search Window with Subtitle Alignment with SAT [26]. One challenge in using sign language interpreted TV broadcasts is that the original subtitles are not aligned to the signing, but to the audio track. In [4], a signing query window is defined as the audio-aligned subtitle timings together with padding on both sides to account for the misalignment. We automatically align spoken language text subtitles to the signing video by using the SAT model introduced in [26], trained on manually aligned and pseudo-labelled subtitles as described in [4]. By using subtitles which are better aligned to the signing, we reduce the probability of missing spottings.

Better Keywords with Synonyms and Similar Words. To determine whether a keyword belongs to a subtitle, previous works [4] check whether the raw form, the lemmatised form, or the text normalised form (e.g. *two* instead of *2*) appears in the subtitle text. We notice that this is sub-optimal

as multiple words may correspond to the same sign, often due to (i) English synonyms, (ii) identical signs for similar words, or (iii) ambiguities in spoken language. For example, *dad* and *father* or *today* and *now* can be the same signs in BSL. In this work, we investigate whether the automatic annotation yield could be improved by querying words beyond the subtitle, by querying synonyms and similar words to the words in the subtitle. We collect the additional words to query through (i) English synonyms from WordNet [68], (ii) the metadata present in online sign language dictionaries such as SignBSL[1] [135] and BSL Sign-Bank[2] which provide a set of 'related words' for each sign video entry; (iii) words with GloVe [149] cosine similarity above 0.9 to account for ambiguities in spoken language.

6.3.5 . Evaluation Framework

Our framework consists of three stages: (a) a costly end-to-end classification training to learn sign category aware video features given an initial set of sign-clip annotation pairs; (b) a lightweight classification training given pre-extracted video features for a large number of annotations; (c) a sliding window evaluation of the trained lightweight model by comparing dense sign predictions against the subtitles (see Sec. 6.4.1). These stages are illustrated in Fig. 6.4. Note that the *annotations* we refer to are always *automatically* localised sign spottings from continuous videos using subtitle information. The motivation for the video backbone and lightweight classifier is purely related to computational costs. Unlike traditional video recognition datasets, we work with untrimmed video data of 1400 hours, where the set of sign-clip pairs is not fixed. Instead, our goal is to increase the number of sign-clip pairs within the continuous stream, and assess the quality of the expanded annotation yield on the proxy task of continuous sign language recognition. Next, we describe the training stages for the video backbone and the lightweight classifier.

Improving the I3D Feature Extractor through Vocabulary Expansion. Following previous works [102, 117, 2, 4], we use the I3D spatio-temporal convolutional architecture to train an end-to-end sign recognition model. We input 16 consecutive RGB frames and output class probabilities. As explained above, this model forms the basis of sign video representation which corresponds to the spatio-temporally pooled latent embedding before the classification layer. The prior work of [4] trains this classifier on the BOBSL dataset (see Sec. 6.4.1) with 2K categories obtained through the vocabulary of mouthing spottings. As a first step, we perform a vocabulary expansion and construct a significantly increased vocabulary of 8K categories. This is achieved by including each sign that has at least 5 training spottings above 0.7 confidence from both mouthing (M) and dictionary (D) annotations. The confidence for the mouthing annotation corresponds to the probability that a text keyword

[1] www.signbsl.com

[2] bslsignbank.ucl.ac.uk

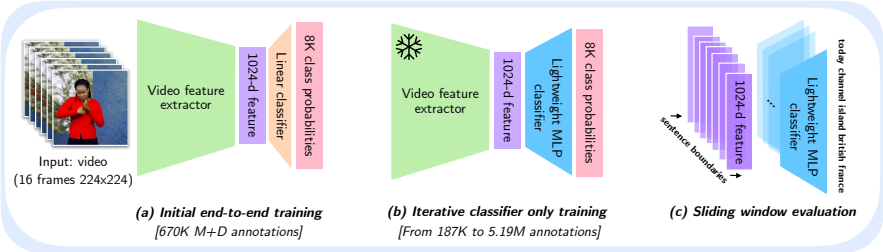


Figure 6.4: **Evaluation framework**: (a) Video features are obtained by training an I3D architecture end-to-end given M + D annotations from [4]. The I3D ingests 16-frames of video and has a linear classifier for 8K sign categories. The end-to-end training is a costly procedure which is not affordable to repeat for each set of our new sign spottings that are on the order of several million training samples. (b) As new sets of spottings are generated, a light weight MLP classifier is trained on the pre-extracted I3D features. This relatively inexpensive training procedure means that we benefit from new annotations without the expense of end-to-end training. (c) The MLP is applied in a sliding window fashion to the signing sequence to generate sign predictions.

(corresponding to the sign) is mouthed at a certain time frame, as computed in [2]. The confidence for the dictionary annotation corresponds to the cosine similarity (normalised between 0-1) between the representations of a dictionary clip of the sign and the continuous signing at each time frame, as in [135]. The resulting M+D training set comprises 670K annotations, with a long-tailed distribution. Furthermore, we note that the categories are noisy where multiple categories may correspond to the same sign, and vice versa. Despite this noise, we empirically show that this model provides better performance than its 2K-vocabulary counterpart. We use our improved I3D model for two purposes: as the frozen feature extractor and as the source of pseudo-labelling for sign spotting (see Sec. 6.3.3).

Lightweight Sign Recognition Model. Following [187], we opt for a 4-layer MLP module (with one residual connection) to assess the quality of different sets of annotations. Given pre-extracted features, this model is trained for sign recognition into 8K categories. We note that we do not train on a larger vocabulary to avoid the presence of many singletons in the training set. The efficiency of the MLP allows faster experimentation to analyse the value of each of our sign spotting sets. The input is one randomly sampled feature around the sign spotting location (the receptive field of one feature 16 frames). The MLP weights are randomly initialised. Additional training and implementation details are given in Sec. 6.3.8.

6.3.6 . Obtaining Synonyms for Querying Keywords and for Eval-

uation

We use synonyms both when querying keywords for spottings and when evaluating the performance of our MLP model. For these two purposes, we construct two different lists of synonyms. The first list is used for querying keywords for spottings and is large and flexible. The second list is a subset of the first; it is used to deem a prediction correct when evaluating our MLP model and is therefore more restrictive.

The first list is an extensive list of synonyms combined from multiple sources: the online dictionaries SignBSL[3], and BSL SignBank[4] 'related words' propositions for each sign video entry; words from the English synonym list from WordNet [68] as well as words with GloVe [149] cosine similarity above 0.9. In order to reduce noise, we remove synonyms with GloVe cosine similarity of less than 0.5. The second list of synonyms is a subset of the first, but we do not add all words with GloVe [149] cosine similarity above 0.9. Instead, amongst words with GloVe similarity above 0.9, we keep only those predicted to be sign synonyms by a simple sign synonym detection model. The sign synonym model is a 4 layer MLP model predicting whether or not two video features correspond to the same or different signs. The model is trained on pairs of M+D+A spottings from [4], and evaluated using the validation split with 33 videos, rather than the 36 aligned test set episodes used in the rest of the chapter. At evaluation, we search for sign synonyms from our first list only amongst words with GloVe similarity of 0.9 and above. For each potential pair or synonyms with more than 5 spottings in the evaluation set, we consider the pair to be sign synonyms if it is predicted to be identical for at least 50% of the evaluation set examples. Tab. 6.1 shows examples of synonyms.

6.3.7 . Different Automatic Annotation Approaches

We provide a summary of the different approaches mentioned in this chapter for annotating signs automatically in sign language interpreted TV shows, which consist of continuous signing and weakly-aligned English subtitles. We highlight specifically the limitations of different approaches and their dependencies.

- M refers to automatic sign annotations obtained in previous work [2] from mouthings, as signers often mouth a word and sign it simultaneously. Specifically, the sign annotations are obtained by querying subtitle words in a signing window with a mouthing-based keyword spotting model and saving the most confident model predictions. Mouthing is a strong signal, but it cannot be used to annotate all data (since

[3] www.signbsl.com

[4] bslsignbank.ucl.ac.uk

Word	Synonyms
change	evolution, diversity, conversion, switch, variety, convert, other, acquire, transform, amend, transformation, deepen, selection, evolve, adaptation, alteration, amendment, various, adapt, transfer, become, exchange, alter, modify, variation, modification, vary, among, shift
bus	coach, heap, metro, subway, tube, underground, vehicle, bus stop
rare	uncommon, few
content	message, capacity, substance, subject, context, insert, relief
architect	designer
airplane	aeroplane
skyscraper	city
king	royal, prince, princess, mogul, queen, power, tycoon, baron

Table 6.1: **Examples of synonyms**: Our list of synonyms contains English words with similar meaning or words that can be signed using the same sign.

signers do not mouth continuously). Furthermore, these automatic annotations are skewed towards words with 'easy' mouthings.

- D refers to automatic sign annotations obtained in previous work [135] by leveraging online sign language dictionary clips. In more detail, a joint embedding space is learned between the *isolated* dictionary video clips and the *continuous* signing video sequences. At inference time, the cosine similarity between the continuous signing sequence and dictionary clips corresponding to subtitle words is calculated. The sign annotations correspond to the dictionary clips with highest similarity. Although these automatic annotations are not limited to signs accompanied by mouthings, they are limited to the vocabulary of the online sign dictionary. Furthermore, they are biased to an extent towards mouthings since the joint embedding space is learned using M annotations.

- A refers to automatic sign annotations obtained in previous work [187] by using the localisation ability from the attention mechanism of a video-to-text Transformer model. The encoder takes as input pre-computed video features (from a sign recognition model trained with M and D annotations) and outputs a sequence of word stems. The sign annotations correspond to words which are correctly predicted and the sign timestamps are obtained by looking at the temporal position where the encoder-decoder attention is maximised. Compared to mouthing (M) and dictionary (D) annotations, the attention (A) annotations are obtained by taking context into account.

- M* refers to new and improved mouthing annotations obtained in this work. In fact, we upgrade to a state-of-the-art keyword spotting model (Transpotter [152]) and finetune this model on signer mouthings. We also use subtitles which are better aligned to the signing for centering

our querying windows. This enables the number of detected mouthings and therefore automatic sign annotations to be greatly expanded.

- D* refers to new and improved dictionary annotations obtained in this work by (i) using subtitles which are better aligned to the signing for centering our querying windows, and (ii) expanding the query set to dictionary clips corresponding to similar words and synonyms to words in the subtitles.

- P refers to new sign annotations obtained in this work through pseudo-labelling. In fact, we train a large-vocabulary (8K) sign classification model with automatic annotations from mouthings (M), dictionaries (D) and attention (A) and use it to pseudo-label. We firstly predict a sign class at each time step in a continuous signing video clip. We then filter the predicted signs to words in the corresponding subtitle.

- E and N are automatic sign annotations obtained in this work by relying on in-domain occurrences of signs. We localise a sign w in a reference video V_0 given (i) the word corresponding to w occurs in the subtitle associated with V_0, and (ii) other exemplar videos $V_1 \dots V_N$ with w in the associated subtitles. When mining instances of *known* classes E, the exemplar videos $V_1 \dots V_N$ are short video segments of the sign w from previous annotation methods. When mining instances of *novel* classes N, the exemplar videos $V_1 \dots V_N$ are longer, subtitle-length videos that have w in their corresponding subtitle. E and N are collected by calculating a matrix of cosine similarities between video features of the reference and exemplar videos. These video features are extracted from the last layer of a sign recognition MLP model trained with M*, D*, A [187], and P (see Tab. 6.6). These in-domain methods are necessary as not all signs have mouthing cues, and signs in continuous signing may differ from their isolated realisations in dictionaries.

6.3.8 . Implementation Details

Transpotter Finetuning

In Section 6.3.4, we discuss the domain gap between the lip movements in videos with the audio track removed (for example, from TV programmes) and the mouthings in sign language videos. As the Transpotter [152] is trained on the former, we finetune it on the pseudo-annotated sign language mouthings to reduce the domain gap. In this section, we describe the process of extracting pseudo annotations and the subsequent finetuning.

Extracting Pseudo-annotations: We start with a pre-defined list of keywords that are at least 3 phonemes in length according to the CMU dictionary [171] and find all occurrences of these keywords in the subtitles.

We take the video segment corresponding to the subtitle as our search window. We add 10 second padding (as also done in [4]) on either side of this video segment to account for the temporal misalignment between the continuous signing and audio-aligned subtitles. We query for the keywords present in the subtitle in order to obtain the temporal localization of each keyword in the video segment. As the video segment is much longer than the segments seen by the model during training, we perform a windowed inference with short 3 second windows. We have a 1.2 second overlap between successive windows. We run two windowed passes through the video, where the start time of the second pass is delayed by one second. This is to ensure that in at least one of these passes, the desired sign (often < 1 second in length) occurs completely within the short window. The Transpotter outputs a per-frame probability indicating whether a word is uttered at that frame. We save the frame number with the maximum probability as a possible annotation for the word and later filter these annotations based on confidence values.

Finetuning: We perform two rounds of finetuning. We first extract pseudo-labels using the Transpotter model from [152], pretrained on silent speech videos. We filter the mouthings with a confidence ≥ 0.7 as positive samples. In each batch, we oversample negative word-video pairs, in order to reduce false positives. We finetune the pre-trained Transpotter at a low learning rate of $1e^{-6}$ using the AdamW optimizer [126]. After convergence, we extract annotations with this more accurate finetuned model. We finetune the model a second time using the same hyper-parameters as above but resuming from the model weights from the first stage of finetuning. Further rounds of finetuning bring negligible improvements. Our final mouthing annotations M^* are extracted using this model.

How Does Finetuning Help? We observe that the Transpotter pre-trained on silent speech segments produces a large number of false detections on signing video segments as shown in Fig 6.5.

After finetuning, the model is less likely to erroneously predict a query word. The decreased number of false positives is reflected by a reduction in overall size of the automatically annotated dataset, noted in Tab. 6.4. The finetuned model only spots $412K$ mouthings compared to the pre-trained model's $661K$. Despite a $1.5\times$ reduction in dataset size, the MLP model achieves better performance when trained on the $412K$ mouthings. Thus, finetuning the Transpotter improves downstream task performance, while also enabling faster and more efficient training of our MLP classifier due to fewer training samples.

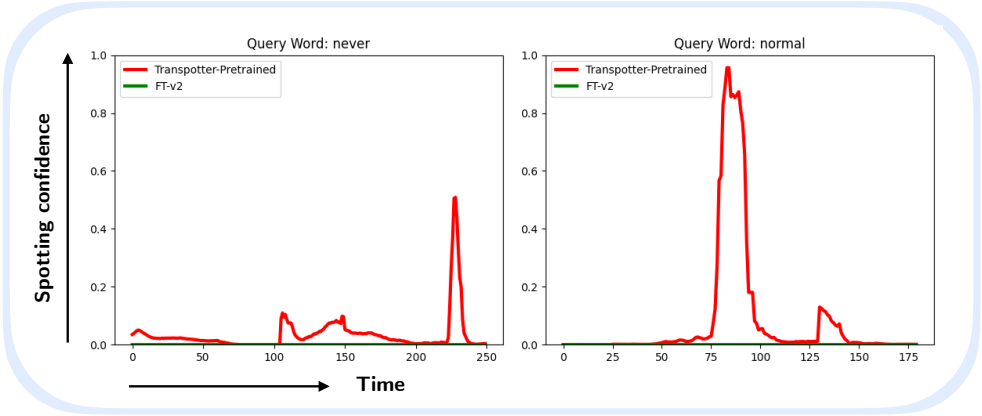


Figure 6.5: **Finetuning the Transpotter on pseudo-annotations leads to fewer false positive detections**: We show two qualitative examples to illustrate the impact of finetuning the mouthing model. For a given query word and a short video segment, we plot the per-frame confidence scores of the two models, i.e. before and after finetuning. We can see that the pre-trained Transpotter spots mouthings even though they are not present, whereas the finetuned model correctly predicts near-zero confidence, indicating that the word is indeed not mouthed in the given video segments.

Video Backbone (I3D)

Here, we describe the training of our I3D video backbone, which is used as the frozen feature extractor and as the source of pseudo-labelling for sign spotting.

As shown in Tab. 6.3.8, we start with the M+D baseline from [4] which is initialised with Kinetics [35] pretraining. This model is trained with sign annotations with confidence above 0.8, resulting in 426K training samples from a 2,281 sized vocabulary. The model takes as input 16 consecutive video frames at 25 fps and a cropped 224×224 spatial region (from an initial 256×256 region). The input to the model is therefore $3 \times 16 \times 224 \times 224$, since our frames are RGB. For each sign annotation from mouthing (M), a sequence of 16 contiguous frames is randomly sampled from a window covering 15 frames before the time associated with the annotation and 4 frames after the annotation, i.e., $[-15, 4]$ around the mouthing peak. For dictionary annotations, the window around the similarity peak is $[-3, 22]$. I3D is trained for 25 epochs using SGD with momentum (with a momentum value of 0.9), with a batch-size of size 4. An initial learning rate of 0.01 is decayed by a factor of 10 after 20 epochs. Augmentations are applied during training including spatial cropping and color augmentations as well as scale and horizontal flip augmentations. The model produces a 1024-dimensional embedding (following average pooling) which is passed to our last linear layer, which outputs scores with the dimensionality of the number of classes. When evaluating the I3D predic-

tions, I3D is run in a sliding window manner over the continuous signing with a stride of 4.

We explore how changing our pretraining effects performance: instead of only pretraining on Kinetics, we use a publicly released model (available on the webpage for [187]) which is first pretrained on Kinetics then finetuned on BSL1K [2] on a 5K vocabulary size. As shown in Tab. 6.3.8, this marginally improves performance on our downstream task of continuous sign recognition.

We explore how expanding the vocabulary from 2K to 8K varies performance: this increases the number of training instances with confidence over 0.8 from 426K to 670K. In this case, our model is only trained for 17 epochs (due to computational costs) with an initial learning rate of 3e-2, reduced by a factor of 10 at epoch 12. As shown in Tab. 6.3.8, this increases our recall from 25.5 to 26.3 and coverage from 15.5 to 16.3. This final model is chosen as our frozen feature extractor and as our source of pseudo-labelling for sign spotting: both features and class predictions are obtained by running I3D in a sliding window fashion with a stride of 4.

Lightweight Classifier (MLP)

As new sets of spottings are generated, a light weight MLP classifier is trained on the pre-extracted I3D features. Our 4-layer MLP module has layers of dimension (1024,512,256,8K) where the last layer corresponds to the number of sign classes and contains LeakyRelu activations in between. The first linear layer also has a residual connection on the 1024-dimensional I3D input features. The MLP is trained with a batch size of 128 for 15 epochs, with the learning rate initially set to 1e-2 and decayed by a factor of 10 at epochs 5 and 10. When evaluating the MLP predictions, the MLP is run in a sliding window fashion, outputting one feature for each I3D input feature (where the I3D features are extracted with a stride of 4).

6.4 . Experiments

We start by describing our dataset and evaluation metrics (Sec. 6.4.1). We then present experimental results on the contribution of each source of annotation and show qualitative examples (Sec. 6.4.2).

6.4.1 . Data and Evaluation Protocol

BOBSL [4] is a public dataset consisting of British Sign Language interpreted BBC broadcast footage, along with English subtitles corresponding to the audio content. The data contains 1,962 episodes, which have a total duration of 1,467 hours spanning 426 different TV shows. BOBSL has a total 1,193K subtitles covering a total vocabulary of 78K words. We note that in this work

we use the word *subtitle* to refer to the processed BOBSL sentences from [4] as opposed to the raw subtitles. There are a total of 39 signers in the dataset. Further dataset statistics can be found in [4]. For a subset of 36 episodes in BOBSL, referred to as SENT-TEST in [4], the English subtitles have been manually aligned *temporally* to the continuous signing video. We make use of this test set to evaluate the quality of our predicted automatic annotations. SENT-TEST covers a total duration of 31 hours and contains 20,870 English subtitles. The total vocabulary of English words is 13,641, of which 5,604 are singletons. The 3 signers in SENT-TEST are different to the signers in the training set, this enables signer-independent BSL recognition to be evaluated.

Evaluation protocol. Given an English subtitle and the *temporally* aligned continuous signing video clip, we evaluate our predicted signs for the clip using (i) *intersection over union* (IoU); (ii) *recall* between signs and the English word sequence; and (iii) *temporal coverage*: this is defined as the proportion of frames in the clip assigned to signs that occur in the word sequence, where a sign is given a fixed duration of 16 frames (for 25Hz video). Note that none of these metrics depend on the word order of the English subtitle, only the words it contains. All metrics are rescaled from the range 0-1 to 0-100 percentage for readability.

For this evaluation, stop words are filtered out since often they are not signed. This reduces the number of test subtitles from 20,870 to 20,547: subtitles such as “is it?”, “Oh!”, “but no” are removed. The sign and word sequences are also lemmatised. We also remove repetitions from the predicted sign sequence and allow the prediction of synonyms of words in the English subtitle. This processing is highlighted in Fig. 6.6, where the IoU and recall are computed for a pair of predicted signs and English text. While this evaluation is suboptimal due to the simplified word-sign correspondence assumption, it tests the capacity of the sign recognition model in a large-vocabulary scenario, necessary for open-vocabulary sign language technologies.

Note, the predicted signs for a clip can be produced in two ways. In the first way, the signs are obtained from the automatic annotations using knowledge of the content of the English subtitle – we refer to these as *Spottings*. In the second, signs are predicted directly from the clip using the MLP sign predictions, without access to the corresponding English subtitle. These are referred to as *MLP predictions*. Spottings are evaluated using all the words; this metric is important to monitor how dense we can automatically annotate the data. The MLP evaluation is limited to the fixed classification vocabulary (of size 8K in our experiments). We note that when different annotations are combined, the sign spotting methods are applied independently.

6.4.2 . Results

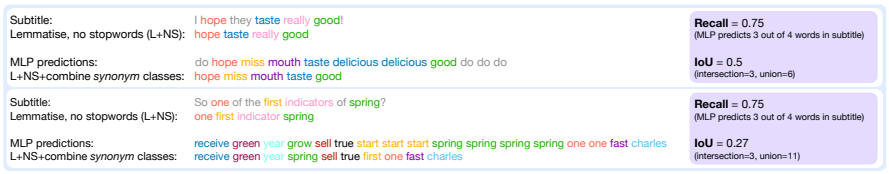


Figure 6.6: **Evaluation illustration on sample prediction:** We illustrate the processing applied to the predicted sign sequence from the MLP predictions and corresponding English subtitle for calculating our metrics. As the MLP model predicts one sign per time-step, some predictions are repeated and irrelevant words appear at transition periods between signs, decreasing the IoU. Some signs are not predicted as they are not signed, showing the limitations of using the subtitle to measure performance.

			I3D predictions (subtitle independent)		
Annot. source	Num. I3D train annot.	Vocab. size	Recall	IoU	Coverage
M [2]+D [135]	426K	2K	25.5	6.4	15.5
M [2]+D [135]	670K	8K	**26.3**	**7.9**	**16.3**

Table 6.3: **Comparison of I3D video features**: We highlight the improved performance of I3D on the test set (SENT-TEST) when trained on a larger vocabulary (8K instead of 2K) with more samples (670K instead of 426K).

Comparison of Video Features. By finetuning our Kinetics pretrained I3D model on BOBSL M+D annotations from [4] using an 8K vocabulary instead of a 2K vocabulary, we improve predictions on the test set, as shown in 6.3. We increase the recall from 25.5 to 26.3 and the coverage from 15.5 to 16.3. We therefore use the 8K M+D model for the rest our experiments as the frozen feature extractor. We note that we restrict the M+D annotations to the high-confidence ones (over 0.8 threshold) used for the I3D baseline in [4], as these present an appropriate signal-to-noise ratio. We use the same threshold for subsequent automatic annotations unless stated otherwise.

Oracle. As the MLPs are trained on a restricted 8K vocabulary, it is not possible to predict the full vocabulary of 13,641 words present in the test set subtitles. Furthermore, not all words in the subtitle are signed and vice versa. This means a recall, IoU and coverage of 100% is not achievable between predicted signs and English subtitle words. However, we propose an oracle in Tab. 6.4 whereby we measure the recall and IoU assuming each word in the subtitle, which either falls within the 8K vocabulary or corresponds to a synonym of a word in the 8K vocabulary, is signed and correctly predicted. The oracle achieves a recall of 86.7 and IoU of 86.3. For the coverage metric, we assume each correctly predicted sign has a duration of 16 frames and no signs overlap. The resulting oracle coverage is 55.2. This low coverage is partly due to the signer pausing within subtitles and also due to the presence of non-

Annotation source	Subtitle alignment	Synonyms	Training set full vocab	Training set #ann. [full]	Training set #ann. [8K]	Spottings [full] (subtitle dependent) Recall	Spottings [full] (subtitle dependent) IoU	Spottings [full] (subtitle dependent) Coverage	MLAP predictions [8K] (subtitle independent) Recall	MLAP predictions [8K] (subtitle independent) IoU	MLAP predictions [8K] (subtitle independent) Coverage
Oracle			-	-	-	-	-	-	86.7	86.3	55.2
Translation baseline [4]			-	-	-	-	-	-	11.7	8.3	7.6
M [2]			13.6K	197K	187K	2.5	2.2	1.3	15.1	3.2	8.7
M [152] (no finetuning)			21.5K	725K	661K	9.4	8.3	4.9	20.4	4.8	11.9
M [152]			18.6K	445K	412K	7.1	6.5	3.9	23.6	4.8	13.8
M [152] (M*)	✓		19.6K	598K	552K	8.9	8.2	4.9	27.4	6.3	16.7
M [152]	✓	✓	19.6K	1.38M	1.25M	11.8	10.4	6.1	25.3	6.2	16.3
D [135]			4.4K	482K	482K	6.5	6.3	3.7	24.0	7.2	15.1
D [135]	✓		4.5K	535K	535K	7.0	6.9	4.0	24.2	7.3	15.3
D [135] (D*)	✓	✓	5.0K	1.40M	1.39M	12.5	11.6	7.0	26.0	7.3	16.9
M*+ D*	✓	✓(D-only)	20.9K	2.00M	1.94M	19.0	17.6	10.5	29.0	7.9	18.4
M*+ D*+ A [187]	✓	✓(D-only)	20.9K	2.43M	2.37M	**21.9**	**20.1**	**11.8**	**29.6**	**9.1**	**19.0**

Table 6.4: **Improved mouthing and dictionary spottings:** We evaluate different sets of spottings and their respective MLP predictions. M [152] shows our finetuned version for all the rows in the last block. We quantify the effects of subtitle alignment and querying synonyms. We also show the oracle performance and a translation baseline.

lexical signs. In fact, the percentage of fully lexical signs in three other sign language corpora (Auslan [99], ASL [99] and LSF [10]) is estimated to be only 70-85% of total signing.

Translation Baseline. Although the goal in this work is not translation, but achieving dense annotations, we can nevertheless compare our MLP predictions to the translation baseline in [4]. Using the test set translation predictions from this model, we perform the same processing as highlighted in Fig. 6.6 to calculate our metrics. As shown in Tab. 6.4, all our simple MLP models clearly outperform the transformer-based translation model used in [4], demonstrating that we are able to recognise more signs in the English subtitle.

Improving Mouthing and Dictionary Spottings. As shown in Tab. 6.4, by using the Transpotter [152] for spotting mouthings M, our yield of total annotations triples from 197K to 725K. The quality of these new annotations is reflected in the increased performance of the MLP: the recall increases from 15.1 to 20.4 and the coverage from 8.7 to 11.9. Finetuning the keyword spotter on sign language data through pseudo-labelling also helps considerably despite the drop in the number of training annotations since there are less false positives; recall increases from 20.4 to 23.6 and coverage from 11.9 to 13.8. Subtitle alignment improves the yield of both mouthing and dictionary annotations, as shown in Tab. 6.4. This translates to a significant boost for mouthings on the MLP performance; the recall increases from 23.6 to 27.4 and the coverage from 13.8 to 16.7. For dictionary annotations, the improvement by using aligned subtitles is less striking. By querying synonyms when searching for mouthings, the yield more than doubles. However, these additional annotations seem to be quite noisy as they decrease the performance of our MLP. Due to the nature of sign language interpretation, it is possible

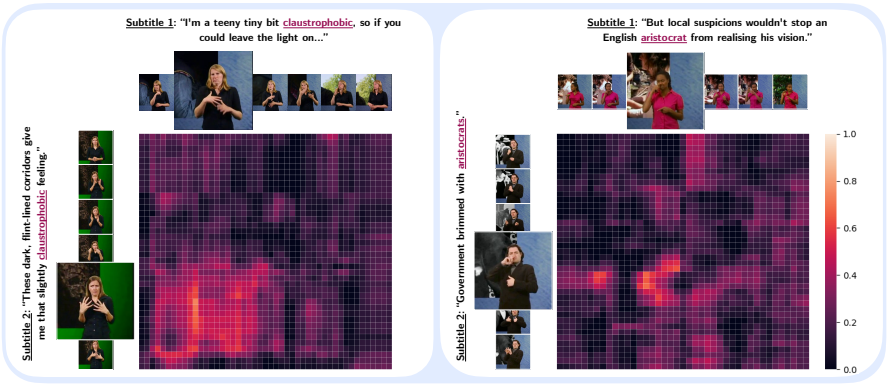


Figure 6.7: **Discovering novel sign classes (N)**: For two pairs of continuous signing sentences, we plot the score maps (as described in Sec. 6.3.1) between their feature sequences. We highlight the ability of our approach to spot novel sign classes.

that signers are far more likely to mouth a word which is actually in the written subtitle than a synonym of that word. We therefore do not query synonyms for mouthing spottings. For dictionary spottings, we observe the opposite effect. By incorporating synonyms, the yield of dictionary spottings more than doubles and the recall of the MLP predictions also increases from 24.2 to 26.0. We denote our best performing mouthing and dictionary spottings with M* and D*, respectively. Adding attention spottings from [4] (with a threshold of 0) adds around 400K additional annotations and boosts the MLP performance; increasing recall from 29.0 to 29.6 and coverage from 18.4 to 19.0, compared to the oracle recall of 86.7 and coverage of 55.2.

Sign Recognition as a Form of Pseudo-labelling. Pseudo-labels P are a source of over 1M new annotations (when using a threshold of 0.5) on top of our best M*, D*, A spottings. As shown in Tab. 6.6, they greatly increase the spottings recall from 21.9 to 25.4 and coverage from 11.8 to 13.9, while only marginally increasing the recall and coverage for MLP predictions. As the pseudo-labels come from our 8K I3D model in Tab. 6.3 whose frozen features are also used for training the MLP, P may not be providing additional information for our downstream evaluation. Nevertheless, they provide a great source of additional spottings (not found by previous methods) for our goal of dense annotation.

Mining more Examples of Known and Novel Sign Classes with In-domain Exemplars. By explicitly querying words in the subtitle text which are not present in our annotations, we can obtain significantly more annotations. Tab. 6.5 shows multiple methods to use exemplar signs to find additional annotations for these signs. The best performing method takes spotting exemplars from across the whole training set, irrespective of signer or episode, and uses the voting scheme described in Sec. 6.3.1 to localise signs. By us-

Ann. src.	ex. data	ex. thres	ex. #	ex. pooling	Training set full vocab	Training set #ann. [full]	Training set #ann. [8K]	Spottings [full] (subtitle dependent) Recall	Spottings [full] (subtitle dependent) IoU	Spottings [full] (subtitle dependent) Coverage	MLP predictions [8K] (subtitle independent) Recall	MLP predictions [8K] (subtitle independent) IoU	MLP predictions [8K] (subtitle independent) Coverage
E	same ep.	0	var	avg	11.6K	869K	833K	10.4	9.6	5.8	25.1	6.9	15.3
E	same signer	0	20	avg	15.9K	505K	421K	7.8	7.5	4.4	23.1	5.6	14.2
E	all	0	20	avg	16.7K	351K	252K	5.7	5.7	3.3	21.5	5.1	13.4
E	all	0.5	20	avg	16.6K	370K	261K	5.9	5.8	3.4	21.9	5.2	13.5
E	all	0.8	20	avg	16.6K	458K	358K	7.4	7.3	4.3	25.2	6.2	15.7
E	all	0.8	20	max	15.4K	1.48M	1.38M	20.2	18.6	10.8	27.6	8.4	17.7
E	all	0.8	10	max	15.4K	1.07M	982K	15.2	14.0	8.3	27.9	8.0	17.7
E	all	0.8	5	max	15.3K	740K	664K	10.7	10.0	6.0	27.6	7.6	17.4
E	all	0.8	20	vote	15.9K	1.76M	1.63M	**25.8**	**23.3**	**13.5**	**28.4**	**8.5**	**18.1**
E	all	0.8	10	vote	15.8K	1.32M	1.21M	20.0	18.1	10.7	28.4	8.3	18.1

Table 6.5: **Ablation on mining exemplar-based spottings for known signs E:** We perform different ablations for mining known signs which have been unannotated by previous methods (M*, D*, A, P). We experiment with the source of exemplar data (same episode, same signer, all data), the confidence of exemplar signs (0,0.5,0.8), the number of samples of exemplar data (5,10,20) and the pooling mechanism (average, max, vote). We evaluate on the test set (SENT-TEST).

ing 20 spotting exemplars, we acquire 1.63M additional annotations. An MLP model trained *only* on these additional annotations achieves a recall of 28.4 and coverage of 18.1. Tab. 6.6 illustrates the impact of combining these additional annotations from spotting exemplars to M*, D*, A and P annotations. With the additional exemplar-based annotations E, recall increases from 29.8 to 30.7 and coverage increases from 19.2 to 19.8, where the oracle recall and coverage are 86.7 and 55.2. Furthermore, by mining instances of novel sign classes N (see Fig. 6.7), we increase our total vocabulary to 24.8K and total number of annotations to 5.47M.

6.5 . Qualitative Examples

6.5.1 . Densification Visualisations

In Fig. 6.8, we show visualisations of our densified sign sequences after our framework is applied.

6.5.2 . Known Classes Spottings Visualisations

In Fig. 6.9, we show visualisations of our score maps for annotating instances of known classes through our in-domain exemplar signs.

6.5.3 . Novel Classes Spottings Visualisations

We show visualisations of score maps for annotating instances of novel classes through our in-domain weak exemplar subtitles. Fig. 6.10 illustrates the necessity of using negative samples to avoid incorrectly identifying signs

Annotation source	Training set full vocab	#ann. [full]	#ann. [8K]	Spottings [full] (subtitle dependent) Recall	IoU	Coverage	MLP predictions [8K] (subtitle independent) Recall	IoU	Coverage
M* + D* + A [187] + P	20.9K	3.64M	3.56M	25.4	23.5	13.9	29.8	8.9	19.2
M* + D* + A [187] + P + E	20.9K	5.40M	5.19M	45.3	40.7	23.3	**30.7**	**9.5**	**19.8**
M* + D* + A [187] + P + E + N	24.8K	5.47M	-	**45.6**	**40.9**	**23.4**	-	-	-

Table 6.6: **Pseudo-label spottings P & Exemplar-based sign spottings for known E and novel classes N:** We highlight the boost in annotations by adding our pseudo-label annotations (P) as well as exemplar-based spottings of known (E) and novel (N) classes. We evaluate Spottings and MLP predictions on the test set (SENT-TEST). For the novel classes, we only show the evaluation of spottings since these are beyond the 8K training vocabulary of the MLP.

common to many subtitles such as pointing signs, pause gestures or other common gestures as the common lexical sign across exemplars. Fig. 6.11 shows a failure case, where we cannot identify the sign for 'mandible' due to two different realisations of the sign depending on context.

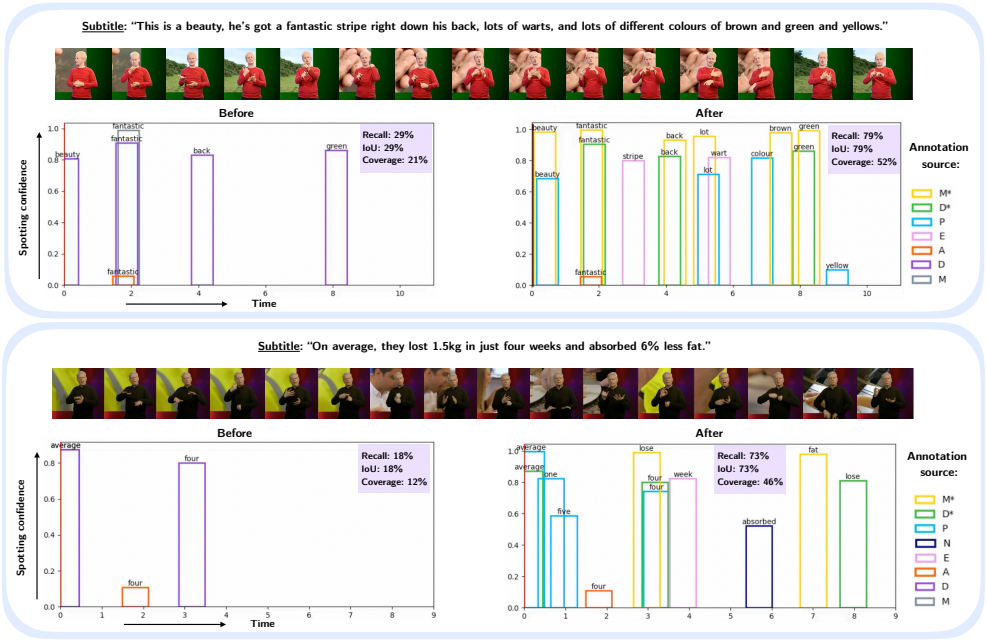


Figure 6.8: **Densification**: For two continuous signing sequences, we show plots of automatic sign annotation timelines, along with their confidence and annotation source, *before* and *after* our framework is applied. We observe that our method enables *densification* by two measures: removing gaps in the timeline so that we have a dense signing sequence spotted; and also increasing the number of words in the corresponding spoken language subtitle we recall. M, D, A refer to spottings obtained from previous methods from mouthings [2], dictionaries [135] and attentions [187] respectively. M^*, D^*, P, E, N refer to new and improved spottings from mouthings [152], dictionaries [135], I3D sign recognition pseudo-labels, in-domain exemplar spottings of known sign classes as well as in-domain exemplar spottings of novel classes respectively.

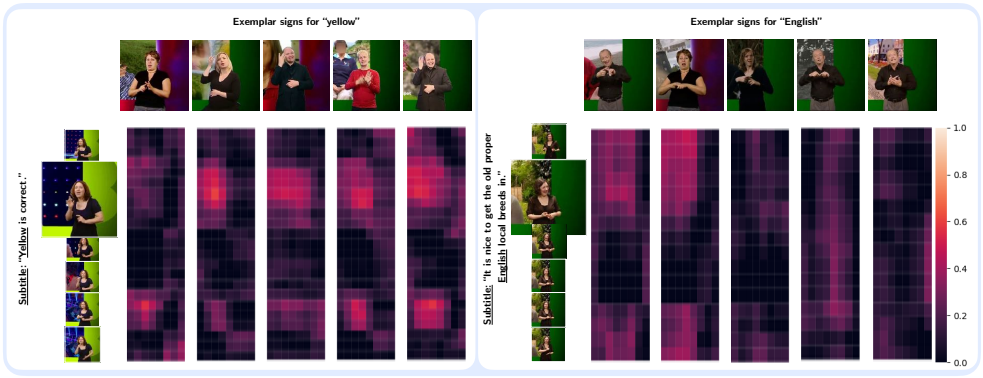


Figure 6.9: **Mining with spotting exemplars**: By comparing the score maps between a subtitle text and multiple spotting exemplars, we can temporally locate a lexical sign in a video segment. The left example illustrates how we can find the sign for 'yellow'. There are two different signs for 'yellow', where the second, third and fifth exemplars correspond to the sign used in the subtitle, and the first and fourth exemplars show an alternative sign. By using a voting method, we can count the number of exemplars with a high cosine similarity at a particular temporal location in the reference subtitle. The right example searches for the sign 'English' in a subtitle using 5 exemplars. The fifth exemplar in an incorrect spotting annotation, and has a low cosine similarity. However, with enough exemplars by different signers in different contexts, we can locate likely temporal locations of a sign in a subtitle.

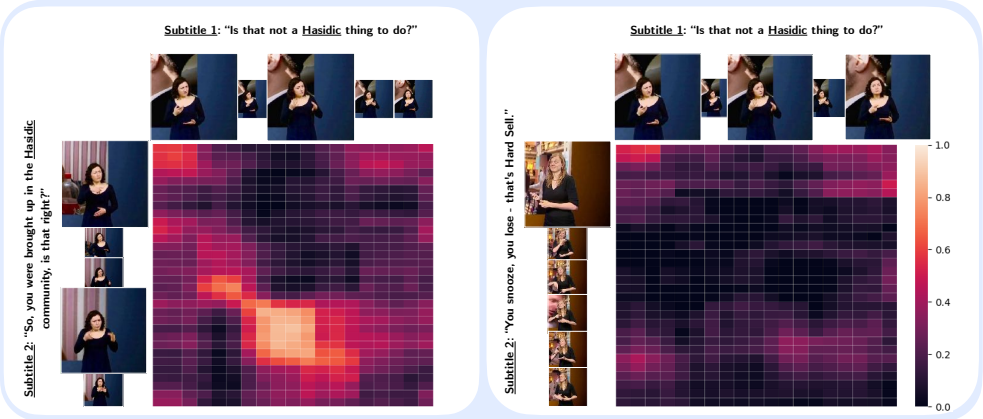


Figure 6.10: **Necessity of negative samples**: On the left, we show the cosine similarity between features of two subtitles, both of which contain the word 'Hasidic'. The cosine similarity is indeed high at the temporal intersection of both signs for 'Hasidic'; but the cosine similarity does also peak at pointing signs common to both subtitles. On the right, we show a score map for a subtitle containing the word 'Hasidic' and a subtitle without this keyword. By using the score maps of negative examples, we can identify non-lexical signs common across subtitles, such as pointing signs, and hence avoid incorrectly labeling the common lexical query sign.

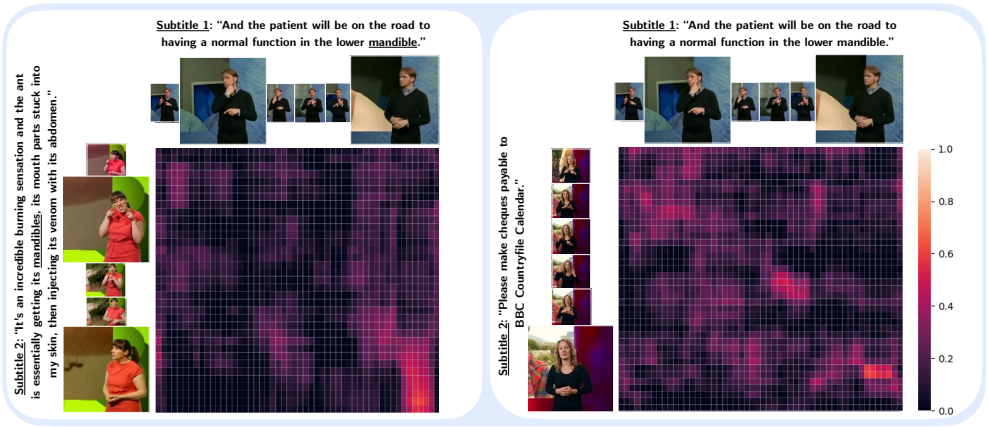


Figure 6.11: **Failure case**: On the left, we show the score map for two subtitles sharing the common word 'mandible'. However, in the first example, 'mandible' refers to a human mandible and in the second example, mandibles of an ant. The sign language interpretation of this word differs in each context, and the score map only shows strong cosine similarity when the signers are in a neutral pause position. The right score map demonstrates that this neutral position, frequent across many subtitles, can be located using negative exemplars. Using information from negative exemplars, we can avoid incorrect annotations.

6.6 . Conclusion

Progress in sign language research has been accelerated in recent years due to the availability of large-scale datasets, in particular sourced from interpreted TV broadcasts. However, a major obstacle for the use of such data is the lack of available sign level annotations. Previous methods [2, 135, 187] only found *sparse* correspondences between keywords in the subtitle and individual signs. In our work, we propose a framework which scales the number of confident automatic annotations from 670K to 5.47M (which we make publicly available). Potential future directions for research include: (1) increasing our number of annotations by incorporating context from *surrounding* signing to resolve ambiguities; (2) investigating *linguistic* differences between spoken English and British Sign Language such as the different word/sign ordering or the spatial organisation of signs; (3) leveraging our automatic annotations for sign language translation.

One of the main limitations of BOBSL is that it is interpreted sign language, which may not be representative of spontaneous sign language usage by native deaf signers. In Chapter 7, we present a new dataset, originally in LSF, with aligned French subtitles.

Part III

Returning to Non-Interpreted Data with a New Corpus

7 - Mediapi-RGB

In this chapter, we introduce MEDIAPI-RGB, an extension of the MEDIAPI-SKEL dataset presented in Chapter 2. MEDIAPI-RGB contains 86 hours of video, and is thus around 3 times larger than MEDIAPI-SKEL. Most importantly, MEDIAPI-RGB contains videos of signers, rather than only skeleton keypoints. Unlike BOBSL, presented in Chapter 4, this dataset does not consist of interpreted sign language content, rather original journalistic content in French Sign Language, and has subtitles in written French aligned to the signing. The current release of MEDIAPI-RGB is available at www.ortolang.fr/workspaces/mediapi-rgb, and can be used for academic research purposes. The test set contains 13 hours of video and the validation set contains 7 hours of video. The training set contains 66 hours of video and will be released progressively, with the full release by December 2024. Additionally, the current release contains skeleton keypoints, sign segmentation, features and subtitles for all the videos in the train, validation and test sets, as well as a suggested vocabulary of common and proper nouns for evaluation purposes. We suggest potential technological and linguistic applications for this new dataset.

The dataset release was prepared with the help of Yanis Ouakrim, Annelies Braffort and Michèle Gouiffès. I contributed to the majority of the preprocessing steps and the writing.

7.1 . Introduction

Mediapi-RGB is a large corpus of original LSF content produced by deaf journalists, available for academic research purposes. The source of the data is the French media association *Média-Pi!*,[1] the same source of the data used to create MEDIAPI-SKEL (Chapter 2). The dataset contains 1230 videos, representing a total of 86 hours of LSF with written French subtitles aligned to the signing. Relying on the subtitle timings, we temporally crop the videos into 50k sentence-like video segments with their associated translations from the subtitle text. These sign language video-text pairs can be used for numerous purposes such as training or evaluating sign language retrieval, recognition or translation models. Figure 7.1 shows a screenshot of a video from Mediapi-RGB along with its associated subtitle.

The Mediapi-RGB corpus has the same advantages of the MEDIAPI-SKEL corpus discussed in Section 2.2, in addition to two main improvements: it is significantly larger (86h vs. 27h) and includes RGB videos in addition to

[1] https://media-pi.fr/

Figure 7.1: **Mediapi-RGB source data.** The source data consists of journalistic content in LSF, along with translated French subtitles aligned to the signing content.

skeleton keypoints. Although BOBSL (Chapter 4) has more hours of sign language video (1400 hours), BOBSL contains BSL content interpreted from English, whereas Mediapi-RGB contains original content in LSF. Moreover, in contrast to BOBSL, the subtitles in Mediapi-RGB are aligned to the signing.

Using skeleton keypoints instead of the original RGB videos results in a loss of information and tends to decrease both human understanding of the sign language as well as model performance. In a human study in [190], the authors find that the level of understanding of signing represented by skeleton keypoints is between 'Poor' (2) and 'Fair' (3) on a mean opinion score scale from 1-5. Using the skeleton keypoints to generate realistic human avatars slightly improves the level of understanding, but it still remains between 'Poor' (2) and 'Fair' (3). In particular, the authors note that hand keypoints are of high uncertainty and low precision, which reduces comprehensibility of sign language. Fingerspelling, for example, relies on accurate hand shape detection.

Methods using RGB frames for action recognition as well as for sign language recognition currently tend to work better than methods which only use skeleton keypoints. In [206], the authors improve results for action recognition under the constraint of using only skeleton keypoints on the Kinetics dataset [35] with a 30% top-1 accuracy rate, in comparison to the previous state-of-the-art at 20% top-1 accuracy. This is in contrast to numerous methods using RGB frames, easily achieving over 50% top-1 accuracy. The current state-of-the-art methods of action recognition on kinetics all use RGB frames, including [186] (87%), [200] (87%), [121] (86%).

For sign recognition, [4] find that using OpenPose 2D keypoints achieves lower results (62% top-1 accuracy) than using RGB frames (76% top-1 accuracy) on BOBSL (Chapter 4). The highest performing models on the popular benchmark corpus PHOENIX14-T [74, 75] for continuous sign language recognition also all use RGB frames as input, including [86] (20.5 word error rate), [132] (22.1 WER) and [148] (24.0 WER). Nevertheless, skeleton keypoints can be used in order to improve the performance of models that take RGB frames as input, as demonstrated in [76] on the PHOENIX14-T dataset (BLEU-4 score of 0.225 without skeletons and 0.248 with skeletons) and a CSL dataset (BLEU-4 score of 0.916 without skeletons and 0.990 with skeletons).

One key issue with sign language corpora available for research is the lack of representation of native deaf signers outside of laboratory conditions. Tab. 4.2 details a list of sign language datasets available for research purposes. Most of the datasets which come from outside of laboratory conditions are from interpreted TV data, e.g. PHOENIX14-T [74], SWISSTXT-WEATHER and SWISSTXT-NEWS [32], and BOBSL (Chapter 4). Interpreters are highly skilled but they are not all deaf native signers. As discussed in Sec. 4.4.1, there are differences between interpreted and non-interpreted language [51] due to source language interference and time constraints. There is some evidence of differences between hearing and deaf interpreters [176], and differences between hearing non-native signers, deaf non-native signers and deaf signers [137].

There are very few sign language corpora made outside of laboratory conditions with original sign language production, rather than sign language interpretation. Two examples of sign language corpora produced outside of laboratory conditions are MEDIAPI-SKEL (Chapter 2) and [166], a fingerspelling dataset compiled from various online sources such as YouTube videos. For a very recent sign language translation challenge, [140] release a dataset with 19 hours of original Swiss-German Sign Language content. Mediapi-RGB aims to complement existing corpora for automatic sign language recognition tasks, by providing 86 hours of video of sign language content outside of laboratory conditions with a high representation of deaf native signers.

7.2 . Dataset Overview

In Section 7.2.1, we first provide a summary of the current Mediapi-RGB dataset release and future releases. In Section 7.2.2, we provide a comparison with the MEDIAPI-SKEL dataset, described in Chapter 2.

7.2.1 . Dataset Content and Statistics

The source data for the Mediapi-RGB corpus is from videos produced by the deaf media association *Média-Pi!*. In order to protect the economic model of *Média-Pi!*, we publish videos of information and sports programmes pro-

duced over three years ago, and these videos may only be used for academic research purposes. We are nevertheless able to partially release elements of the training set, including OpenPose keypoints [33], video features, automatic sign segmentations [154] and subtitles. Due to the time constraint of releasing the original videos 3 years after the production date, we have decided to use videos dating from 2017-2018 in the validation set, and videos from 2019 in the test set, so that models trained on other data sources can already be evaluated on Mediapi-RGB. Other videos from 2020-2022 will be progressively released until December 2024. The current release of the training set of Mediapi-RGB contains skeleton keypoints, features, automatic sign segmentations and subtitles of the Mediapi-RGB training set (as well as the validation and test sets). We also suggest a 7k vocabulary of common and proper nouns in the Mediapi-RGB subtitles for evaluation purposes.

	Train	**Dev**	**Test**	**Total**
Release date	Progressive release of RGB videos until Dec. 2024 (features and skeleton keypoints available)	Fully available	Fully available	-
# videos	950	74	206	1230
# subtitles	37651	4373	8060	50084
# hours (vid.)	66.1	7.2	12.5	85.9
# hours (subs.)	52.3	4.9	10.8	68.0

Table 7.1: Train - Dev - Test split

7.2.2 . Relationship to Mediapi-SKEL

The Mediapi-RGB dataset is intended as a replacement dataset for MEDIAPI-SKEL, rather than a complementary dataset to MEDIAPI-SKEL. This is because the training set of MEDIAPI-SKEL partially overlaps with the test partition of Mediapi-RGB. The current release of the training set of Mediapi-RGB contains skeleton keypoints for the train, validation and test sets, and so skeleton-based models can currently be trained and evaluated on Mediapi-RGB rather than MEDIAPI-SKEL. The advantages of Mediapi-RGB are that: 1) there are more hours of video (86 hours vs. 27 hours), 2) all videos are converted to 25 fps to facilitate analysis, 3) the original RGB videos are

available on the validation and test sets, and will soon be released on the training set, and 4) additional data such as OpenPose keypoints [33], features and automatic sign segmentations are available on the train, validation and test sets. The main disadvantage of Mediapi-RGB in comparison to MEDIPAI-SKEL is the fact that there are fewer signers, due to the exclusion of certain types of programmes including interviews with members of the public. A comparison of the statisics of MEDIAPI-SKEL and Mediapi-RGB can be found in Table 7.2.

	MEDIAPI-SKEL	**Mediapi-RGB**
Global statistics		
# subtitled videos	368	1230
# hours	27	86
# frames	2.5 million	7.7 million
Video statistics		
Resolution	1080p (327 videos) 720p (41 videos)	2160p (28 videos) 1080p (1182 videos) 720p (18 videos) 480p (2 videos)
Framerate	30 fps (111 videos) 25 fps (242 videos) 24 fps (15 videos)	25 fps
Average length of video	4.5 minutes	4.2 minutes
# signers	>100	>10
Text statistics		
# subtitles	20 187	50 084
Average length of subtitle	4.2 seconds 10.9 words	4.9 seconds 12.2 words
Vocabulary size (tokens)	17 428	35 599
Vocabulary size (nouns+verbs+adjectives)	14 383	27 343

Table 7.2: Descriptive statistics of our datasets MEDIAPI-SKEL and Mediapi-RGB.

7.3 . Dataset Collection

This section describes the dataset collection, as well as various post-processing steps to facilitate usage of Mediapi-RGB. We download all available videos from *Média-Pi!*'s private YouTube channel, and convert all videos to 25fps. We release these videos for the validation and test sets,

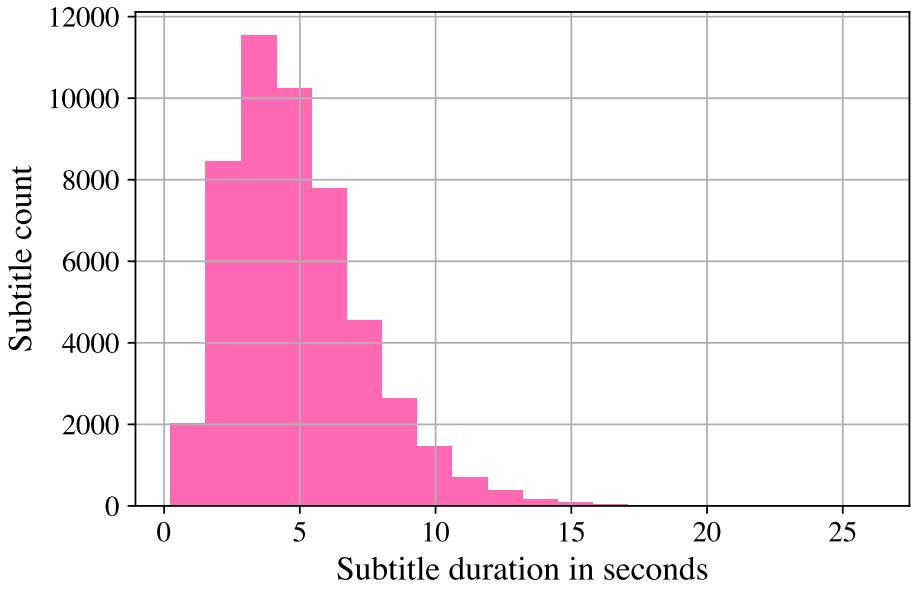


Figure 7.2: **Duration of subtitles.** The distribution of the duration of the extracted subtitle clips (without 0.5s padding).

and will progressively release the videos for the training set. We temporally crop the videos using the start and end times of each subtitle (Sec. 7.3.1) and spatially crop the signers of each subtitle text (Sec. 7.3.2). This creates video-text sentence-like pairs, with the signer centred within a square crop of 444×444 pixels at 25fps. We remove duplicates to ensure that there are no identical videos across the train, validation and test splits (Sec. 7.3.3). Although we cannot currently release the training set videos, we can release derivative products on the training set, including 2D skeleton keypoints (Sec. 7.3.4), features (Sec. 7.3.5), sign segmentation (Sec. 7.3.6), original and processed subtitles (Sec. 7.3.7), as well a proposed noun vocabulary for evaluation purposes (Sec. 7.3.8).

7.3.1 . Temporally Cropping Subtitles

In Mediapi-RGB, the subtitles are aligned to the signing. Almost all of the *Média-Pi!* videos are produced in LSF, then subsequently subtitled in written French. There are rare cases where a hearing person is interviewed in spoken language and this is interpreted into LSF. In these cases, the subtitles correspond to the original audio, but are aligned to the signing, and the audio track is removed. We temporally crop the videos using the subtitle timestamps, adding 0.5s on each side as padding. This creates sentence-like video-text pairs. The distribution of subtitle durations is shown in Fig. 7.2, and the distribution of the original video durations is shown in Fig. 7.3.

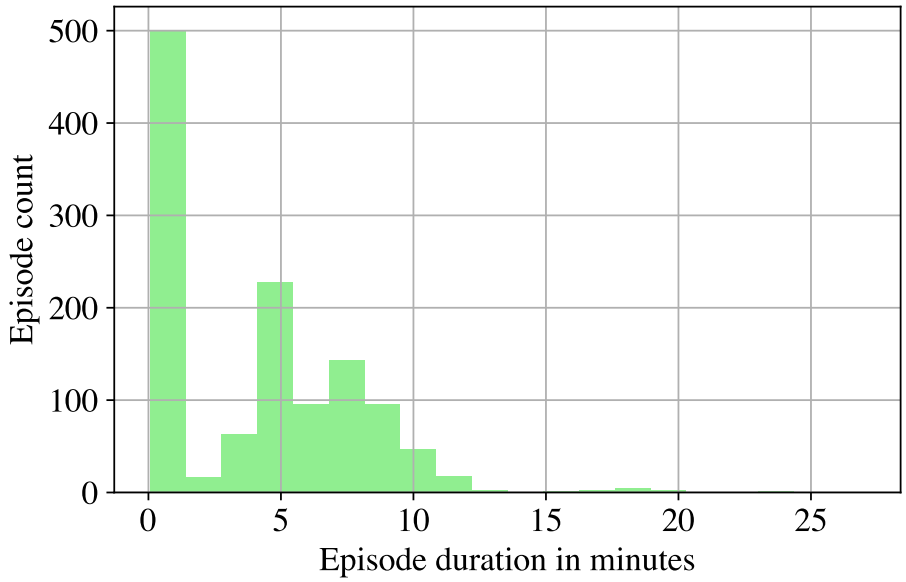


Figure 7.3: **Duration of episodes.** The distribution of the duration of *Média-Pi!* episodes.

7.3.2 . Spatially Cropping Signers

As the *Média-Pi!* videos capture signers in different positions and orientations, we automatically extract bounding boxes around the most likely signer in each of the extracted subtitle crops (Sec. 7.3.1). To do this, we follow the methodology described in Sec. 3.3 with available online code.[2] We extract the 2D OpenPose [33] keypoints of each person in the videos, omit the legs and feet keypoints, track each person between consecutive frames, impute missing skeleton keypoints using past or future frames, temporally smooth keypoints using a Savitzky-Golay filter and omit unlikely signers such as people with occluded hands, people with hands that hardly move or people in the background. In the case of multiple potential signers, we then choose the most likely signer based on a metric computed by multiplying the hand size and the variation of wrist movement of the dominant hand. We then use a static square crop around this most likely signer for the duration of each subtitle. This square crop is then resized to 444×444 pixels. Fig. 7.4 shows an example when two people are detected using OpenPose. The person on the left has static hands, so the person on the right will be detected as the signer. We note also that OpenPose fails to detect the fingers of the person on the left due to the gloves. Nevertheless, failure to detect hands due to gloves or other occlusions also indicates that the person is not the main signer.

[2] https://github.com/hannahbull/clean_op_data_sl

7.3.3 . Removing Duplicate Videos

To ensure that there are no duplicate videos across train, validation and test splits, we firstly look for subtitles with identical text. The associated videos may simply be a common phrase used in different contexts, but the associated videos with these subtitles also could be identical, due to the same video sequence being used in two videos (e.g. an advertisement video for another video, a citation or the opening sequence of a programme). To avoid biases due to duplicate videos in both the train and validation or test sets, we measure the similarity of subtitle crops with the same subtitle text.

We extract features for sequences of 16 frames using the method described in Sec. 7.3.5. For each pair of videos, corresponding to two subtitles with the same subtitle text, we extract the features at stride 4 and compute the maximum cosine similarity between each pair of features. Letting $\{\mathbf{a}_1, ..., \mathbf{a}_n\}$ be features of a video A and $\{\mathbf{b}_1, ..., \mathbf{b}_m\}$ be features of a video B, we compute:

$$S_{A,B} = \max_{i \in 1,...,n, j \in 1,...,m} \frac{\mathbf{a}_i \cdot \mathbf{b}_j}{\|\mathbf{a}_i\| \|\mathbf{b}_j\|}. \tag{7.1}$$

If the maximum cosine similarity $S_{A,B}$ in Equation 7.1 is above a threshold $T = 0.95$, we consider the two videos to be identical. If A and B are in different splits (e.g. train and test sets), we lower the threshold to $T = 0.85$, in order to prevent duplicates across splits. We then remove either video subtitle A or B corresponding to the shortest original video episode. We choose to omit the subtitle from the shortest original video episode in order to remove videos corresponding to advertisement segments of longer videos.

7.3.4 . Extract OpenPose Skeleton Keypoints

OpenPose [33] keypoints are useful inputs for many automatic sign language processing tasks. For example, skeleton keypoints were used in Chapter 3 to extract signers and to segment sign language, and have been used for making hand or face crops [94, 166], generating sign language [190, 162], or as inputs to improve recognition methods [10, 96]. We thus provide OpenPose keypoints for the body (25 keypoints), the hands (21×2 keypoints) and the face (70 keypoints) [33, 169]. The OpenPose keypoints are X and Y pixel coordinates with confidence scores between 0 and 1, where keypoints which fail to be detected or are occluded have 0 confidence score values. Fig. 7.4 shows an example of the face, body and hand OpenPose keypoint detection. There are alternatives to OpenPose keypoints, such as MediaPipe[3]. However, we choose to provide OpenPose keypoints due to their widespread usage in sign language understanding [162, 190, 4, 24].

7.3.5 . Extract Features

[3] https://google.github.io/mediapipe/

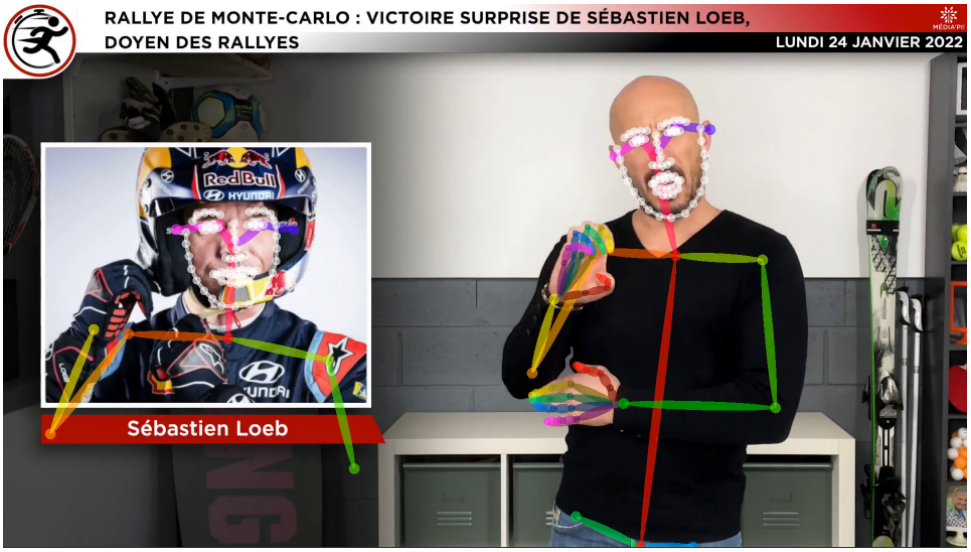


Figure 7.4: **Illustration of OpenPose keypoints.** Illustration of the face, body and hand keypoint detection using OpenPose [33].

Model	Annot. source	Num. train annot.	Vocab. size	I3D predictions (subtitle independent) Recall	IoU	Coverage
I3D	M [2]+D [135]	426K	2K	25.5	6.4	15.5
I3D	M [2]+D [135]	670K	8K	26.3	7.9	16.3
I3D	M [2]+D [135]+A [135]+P	3.56M	8K	28.6	9.5	17.6
Swin	M [2]+D [135]+A [135]+P	3.56M	8K	30.9	10.9	19.1

Table 7.3: **Comparison of video features.** Swin features outperform I3D features on the task of sign recognition. (Extension of Tab. 6.3.)

Features are a useful input for automatic sign language processing systems, because they allow models to be trained more rapidly than end-to-end systems using RGB video frames as input. We use features directly as input in Chapter 5 to align subtitles to sign language video. Features are also used in Chapter 6 to search for lexical signs using exemplars (Sec 6.3.1).

To extract features, we use the Swin transformer model [125] trained for the task of sign recognition using dense automatic annotations acquired using the methods described in Chapter 6, and also used in [153]. The input to the Swin transformer model is a temporal context window of 16 frames, and the output is a vector of features of dimension 768. We extract these features at stride 1. Due to 0.5s padding applied on each of the subtitle timestamps during temporal cropping, all of the extracted videos contain at least 16 frames. Tab. 7.3 extends upon Tab. 6.3, illustrating the improvements of the Swin model in comparison to the I3D model used in Chapter 6.

Fig. 7.5 illustrates an example of the cosine similarity between pairs of features from two video extracts corresponding to subtitles, computed using Equation 7.1. The highest value on the cosine similarity map corresponds to

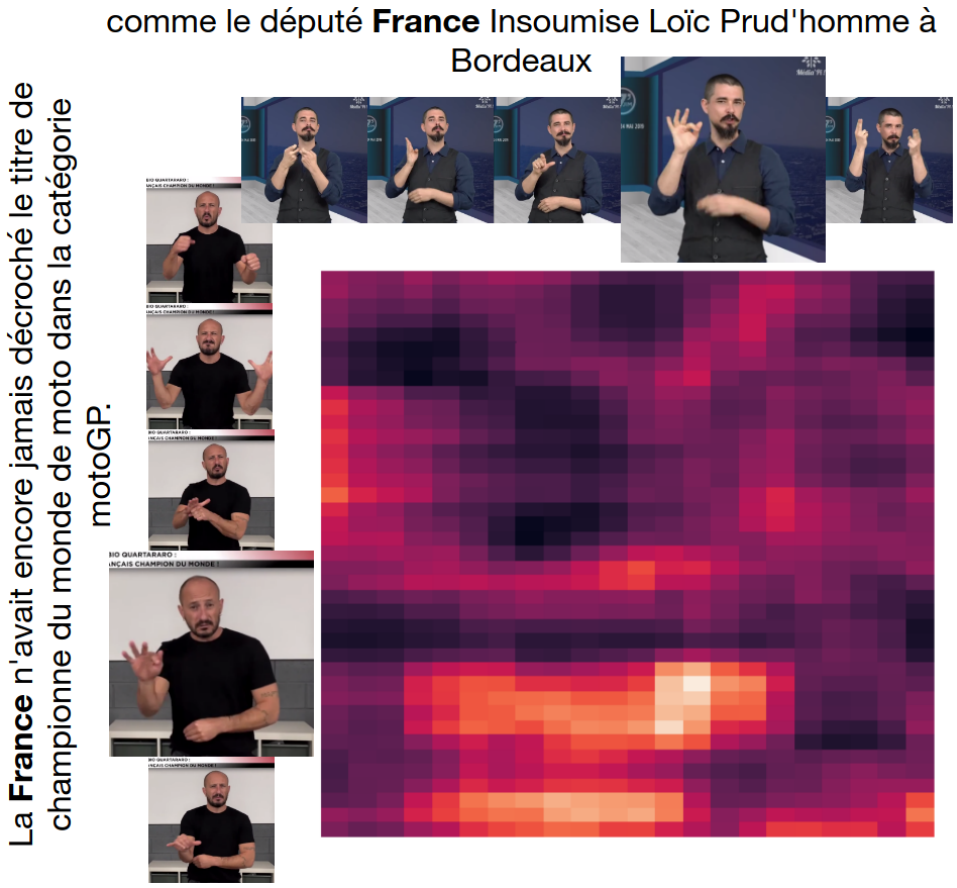


Figure 7.5: **Feature similarity.** Example of cosine similarity between subtitles containing the same word 'France'. The highest cosine similarity occurs when both signers sign 'France'. There is also a high cosine similarity when the signer on the vertical axis signs 'jamais' (never) and the signer on the horizontal axis signs 'France'.

the moment when both signers sign 'France', the common word in both subtitle texts. There is also a high cosine similarity corresponding to when the signer on the y-axis signs 'jamais' (never) and the signer on the x-axis signs 'France', due to similar hand movements, although different hand shapes.

7.3.6 . Sign Segmentation

In [155], the authors train a model to automatically segment sign glosses using annotated sign glosses from BSL Corpus [164]. This model inputs I3D features from [4] and outputs change-points between sign glosses. The model tends to recognise changes in handshape, and thus over-segments fingerspelling. We use available online code[4] to segment signs in the Mediapi-RGB

[4] https://github.com/RenzKa/sign-segmentation

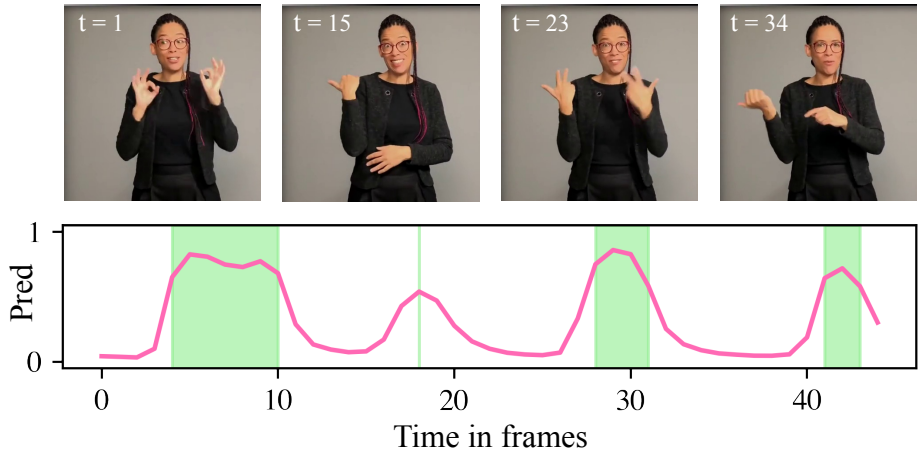


Figure 7.6: **Sign segmentation.** Example of sign segmentation on an extract of video from Mediapi-RGB. The pink line denotes the predicted output scores and the green blocks denote a binary threshold for scores above 0.5.

subtitle crops. Although trained on BSL data, this model seems to provide valuable approximations of sign gloss boundaries in LSF. Fig. 7.6 shows an illustration of the sign segmentation model on a Mediapi-RGB video. The sign segmentation model recognises transitions between signs due to visual cues such as changes in hand shape.

7.3.7 . Text Processing

We provide the raw subtitle texts for all of the videos in the train, validation and test sets. Additionally, we extract the part-of-speech of each subtitle word.[5] The total vocabulary size is 35599 and the vocabulary size of nouns, verbs and adjectives is 27343 (Tab. 7.2). The later vocabulary size is relevant for sign language analysis, as it is unlikely that other parts of speech such as prepositions, determinants and adverbs have associated lexical signs.

7.3.8 . Noun Vocabulary

We propose a vocabualary of 6939 common and proper nouns for evaluation purposes. This vocabulary corresponds to a list of nouns and proper nouns from the Mediapi-RGB subtitles (prior to episode filtering), appearing at least 5 times. Fig. 7.7 plots the frequency of the top 20 most commonly used nouns in this corpus. This vocabulary is for the purposes of evaluating weakly-supervised sign recognition with subtitle text. Certain parts of speech such as nouns and proper nouns are more likely to be present in the signing as lexical signs than parts of speech such as adverbs, prepositions or determinants. We would like to be able to evaluate models for recognising certain

[5] https://huggingface.co/gilf/french-postag-model

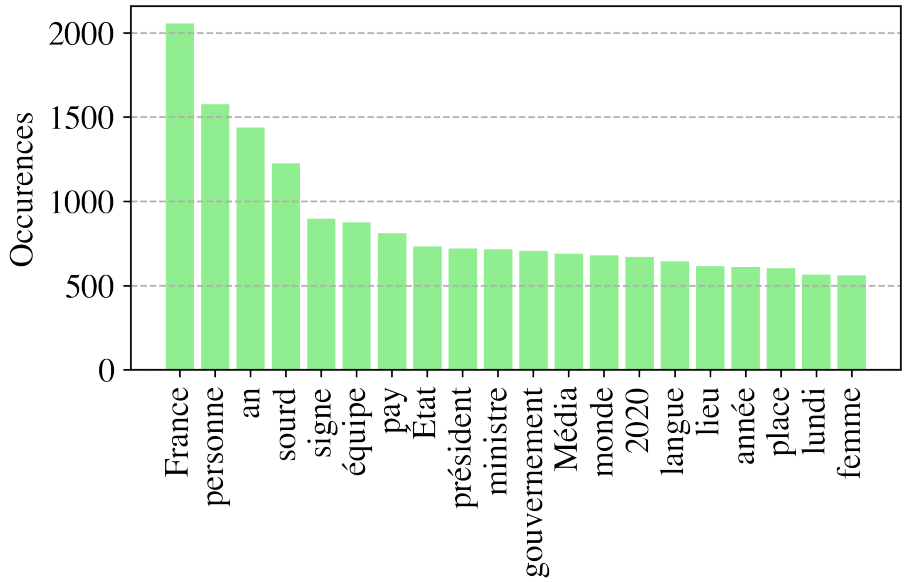


Figure 7.7: **Top 20 most frequent words in noun vocabulary.** The frequency of the top 20 noun occurrences.

sign in continuous sign language sequences without manual sign gloss annotations. The nouns and proper nouns in the subtitle text give information on the presence or absence of certain lexical signs. Fig. 7.8 shows the distribution of the number of words in each subtitle, as well as the number of words in the vocabulary in each subtitle.

7.4 . Opportunities and Limitations

In this section, we discuss opportunities and limitations of the Mediapi-RGB dataset for technological applications (Sec. 7.4.1) and for research in sign language linguistics (Sec. 7.4.2).

7.4.1 . Applications perspective

In this section, we discuss the potential of Mediapi-RGB to lead to useful applications for deaf communities. We note that there are historical examples of sign language technological applications with little consultation of the deaf community nor practical value [20, 64]. Research projects using Mediapi-RGB should take into consideration input from deaf researchers and members of the deaf community in order to ensure practical benefits and to avoid harm. *Média-Pi!* is a project created by deaf journalists to increase access to information in the deaf community. We thus note three potential applications of Mediapi-RGB for information access: anonymity, sign retrieval and automatic

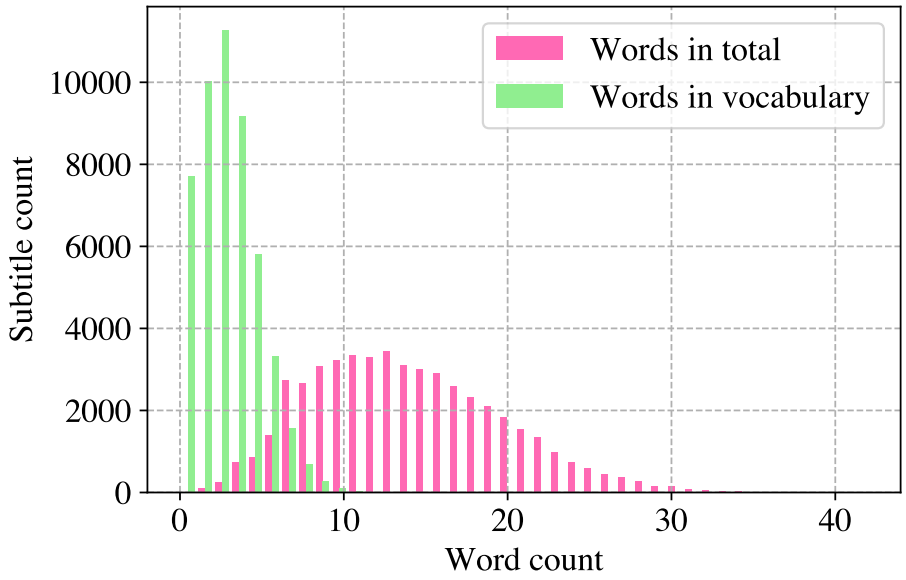


Figure 7.8: **Number of words in subtitles.** In pink, the distribution of the total number of words in the Mediapi-RGB subtitles; and in green, the distribution of the number of the nouns in our vocabulary in each subtitle (see Sec. 7.3.8 for details on this vocabulary).

subtitling.

One key characteristic of written language is the ability to share information anonymously. It is onerous to present information in sign language without also representing the identity of the signer. Solutions such as computer-generated avatars are difficult to implement. MOCAP methods for animating sign language avatars are not necessarily anonymous, as individuals can be recognised by their movements and signing style [11]. In [12], the authors discuss methods to remove identity cues from sign language production. Such methods could be combined with techniques of realistic sign generation [162, 190]. Anonymous representations of sign language can then be used to communicate factual information such as legal or administrative information, governmental documents or weather reports, or to represent signers who wish to conceal their identity.

Search engines are very efficient for finding written information based on textual queries. Searching for information in sign language using text queries or sign queries is very difficult due to the challenges of recognising signs and clustering similar topics in sign language videos without written translations. In [60], the authors present a method for searching for sign language sequences using free-form textual queries. In [95], the authors search for isolated signs in continuous sign language video on BSL Corpus [164] and

PHOENIX14-T [74, 75]. The Mediapi-RGB corpus can be used to train and evaluate models for retrieval tasks on continuous sign language videos using textual or sign queries, as the subtitles of the videos may be used as weak annotation.

Adding subtitles to sign language video improves accessibility and comprehension, but is a time-consuming task. Various ways to simplify this task are automatic segmentation into subtitle-units (as discussed in Chapter 3), automatic alignment of subtitles to video (as discussed in Chapter 5), and eventually automatic translation of sign language video to text. These tasks can be trained and evaluated using the videos in the Mediapi-RGB corpus.

Users of Mediapi-RGB should be aware of the specificities of this dataset. Models trained on other datasets may not necessarily perform well on Mediapi-RGB, and models trained on Mediapi-RGB may not generalise well to other situations. The *Média-Pi!* videos are of professional quality and are unlikely to be representative of spontaneous conversations in daily life. This can be considered both an advantage and a limitation of Mediapi-RGB. The signers are highly skilled deaf journalists producing examples of eloquent and formal LSF, but models trained on Mediapi-RGB may be less applicable to signers with lower levels of fluency or a more informal register.

7.4.2 . A sign linguistics perspective

One common critique of current large sign language corpora for automatic sign language processing is the over-representation of interpreted TV data (See 5.1 for a table of sign language corpora) [20]. Nevertheless, the differences between interpreted sign language data and non-interpreted sign language data are not well understood. Comparing Mediapi-RGB with the other available journalistic sign language content from interpreted journalistic data sources such as SWISSTXT-NEWS [32], VRT-NEWS [32] and BOBSL (Chapter 4) may provide some clues on how sign language production from deaf journalists differs from interpreted journalistic content from both hearing and deaf interpreters. Increased linguistic knowledge about this differences would help acknowledge and alleviate the biases of models trained on interpreted sign language data.

The number of signers is lower in Mediapi-RGB than in MEDIAPI-SKEL, due to the removal of numerous videos involving interviews with members of the public at events. Nevertheless, there are over 10 different signers in Mediapi-RGB, allowing for linguistic comparisons across signers.

7.5 . Conclusion

We presented a summary of the new dataset Mediapi-RGB, a large dataset with 86 hours of sign language content from deaf journalists at the French media association *Média-Pi!*. The videos are in LSF with aligned subtitles in

written French. We hope that Mediapi-RGB can be used to train and evaluate automatic sign language processing tasks, as well as contribute to research in sign language linguistics.

8 - Conclusion

In this chapter, we summarise our contributions in Sec. 8.1. We note the risks of this work in Sec. 8.2. We then outline perspectives for future work in Sec. 8.3, developing upon the contributions of this research.

8.1 . Summary of Contributions

This thesis provided large sign language datasets with subtitles for academic research purposes. Using these datasets, we developed new methods for automatic sign language understanding.

- Chapter 2 introduced MEDIAPI-SKEL, a 2D-skeleton keypoint dataset of 27 hours of French Sign Language with written French subtitles. The dataset comes from journalistic content originally in LSF, and the subtitles are well aligned to the signing. We presented three potential challenges for this dataset: sentence-like segmentation, alignment and video-text features.
- Chapter 3 developed baseline results for a new task of segmenting sign language into sentence-like units, using 2D-skeleton keypoints as input. Our model is a spatial-temporal graph convolutional neural network. We trained and evaluated sentence-like segmentation on the MEDIAPI-SKEL dataset of Chapter 2.
- Chapter 4 presented the BOBSL dataset, a collection of around 1400 hours of interpreted British Sign Language video from BBC programs, with English subtitles. The subtitles are not aligned to the signing, rather to the audio track. We detailed the different splits of the data, as well as available annotations. We also discussed the limitations of this data.
- Chapter 5 demonstrated a new method to jointly segment video and align subtitles to signing, building upon the work in Chapter 3. Our model is a transformer, inputting the query subtitle text in the encoder and inputting the video in the decoder. We trained and evaluated the model using BOBSL, as well as other datasets.
- Chapter 6 improved upon existing methods and developed new methods for dense annotation of lexical signs. We improved mouthing and dictionary spotting methods, made use of pseudo-labelling using a sign recognition model for sign spotting and proposed a novel approach for increasing annotations using sign exemplars. These methods were trained and evaluated using BOBSL.

- Chapter 7 introduced MEDIAPI-RGB, a dataset with 86 hours of RGB video from the same source as MEDIAPI-SKEL. The videos are originally in LSF, with aligned French subtitles. We also released 2D-skeleton keypoints, video features, sign segmentations and a proposed vocabulary for sign recognition tasks.

8.2 . Risks

There are risks associated with work on automatic sign language understanding. An important risk of sign language research is that the deaf community is ignored or excluded from projects, leading to either useless or harmful applications and redirecting funding from other projects with a more positive impact. There are numerous examples of such projects, such as gloves for recognising the ASL alphabet, from the computer science community [20].

Automatic sign language understanding can lead to increased surveillance of deaf communities. As sign language production in videos is generally not anonymous, even if the face is blurred, automatic methods can potentially be used to recognise and track content produced by particular individuals [11]. Sign language content can be automatically moderated or censured, or used for additional purposes such as targeted advertising.

Another risk is that the language models developed in this work learn to replicate undesirable aspects of the corpora on which they are trained. For example, despite the fact that our corpora come from reputable, professional sources, models may learn racist or sexist biases present in the data. The corpora used in this work may not be representative of the diverse range of signers and situations that occur in daily life. In particular, models trained using interpreted data may have lower performance on non-interpreted sign language videos. Some groups of signers are not well represented in our corpora, including children and elderly signers as well as minority groups. These populations may be inadvertently excluded from access to sign language technologies trained using our data due to low performance.

8.3 . Future Work

In this section, we outline ideas on how to build upon the work developed in this thesis.

Sentence-like segmentation: It would be interesting to revisit the problem of sign language segmentation in Chapter 3, but using the MEDIAPI-RGB dataset described in Chapter 7. Using RGB videos rather than skeleton keypoints would likely improve performance. It would also be interesting to train segmentation using corpora from multiple sign languages. The visual cues for

sentence-like segmentation - such as short pauses, blinks and nodding - are likely to be similar across sign languages, as suggested by the fact that even non-signers can reliably segment sign language into sentence-like units [71].

Alignment: After developing our method for dense annotation in Chapter 6, it would be interesting to revisit the problem of alignment discussed in Chapter 5. By recognising many new lexical signs from the subtitle text, it should be possible to improve automatic alignment of subtitle texts to segments of sign language video.

From dense annotation to translation: In Chapter 6, we densely annotate lexical signs. This is one step towards translation, however there remains a lot of research to understand the non-lexical signs in sign language discourse. In [9], the authors recognise learn to recognise types of non-lexical signs. It remains unclear how to consolidate the knowledge of lexical and non-lexical signs to form coherent written sentences in automatic sign language translation.

Differences between interpreted and non-interpreted data: In Chapter 2 and Chapter 7 we present two non-interpreted sign language datasets and in Chapter 4, we present a dataset with interpreted sign language. Although fluent signers can generally tell the difference between interpreted sign language and non-interpreted sign language, as well as signing by native deaf signers and non-native or non-deaf signers, there is little work on describing or quantifying these differences. Using automatic methods, it is possible to conduct large-scale experiments to quantify differences in interpreted and non-interpreted data, as well as differences between non-native, native, hearing and deaf signers. For example, perhaps mouthings or lexical signs are more frequent in interpreted sign language, or amongst certain groups of signers.

Cross-language sign features: Due to iconicity, there are many similarities across different sign languages. For example, the sign for the verb 'to eat' generally involves some sort of hand-to-mouth motion, although the hand shape may vary. Rather than collect large datasets for each existing sign language, it could be better to train a multi-language model to recognise different signs across different sign languages, and then finetune these models depending on the content. A model trained across different sign languages could learn to make fine-grained distinctions between signs, such as different mouthings or slight variations between hand shapes. Large language models such as GPT-3 [21] and XGLM [123] are trained on multiple written languages in order to have better generalisation to various applications.

Practical tools: Although perhaps not the main mission of researchers, there is much future work to making automatic sign language tools available to the public. For example, the VIA Sign Language Annotator [1] [155, 61] is a tool

[1] https://github.com/RenzKa/VIA_sign-language-annotation

to automatically segment signs and to annotate propositions of lexical signs. There are many other tools which would be relatively simple to implement, based on the work in this thesis and in other sign language research. For example, an automatic alignment tool could be useful to signing vloggers to save time when temporally aligning written subtitles to their sign language video. An automatic alignment tool could also be useful to create bilingual concordancers with aligned written language and sign language video, similar to DeepL's Linguee [2]. A dense annotation tool could be useful for creating a sign language learning application, or for making a collection of sign language videos easily searchable using text queries.

[2] https://www.linguee.com/

Bibliography

[1] N. Adaloglou, T. Chatzis, I. Papastratis, A. Stergioulas, G. T. Papadopoulos, V. Zacharopoulou, G. J. Xydopoulos, K. Atzakas, D. Papazachariou, and P. Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020.

[2] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020.

[3] S. Albanie, G. Varol, L. Momeni, T. Afouras, A. Brown, C. Zhang, E. Coto, N. C. Camgöz, B. Saunders, A. Dutta, N. Fox, R. Bowden, B. Woll, and A. Zisserman. Signer diarisation in the wild. *Technical Report*, 2021. URL https://www.robots.ox.ac.uk/~vgg/publications/2021/Albanie21a/albanie21a.pdf.

[4] S. Albanie, G. Varol, L. Momeni, T. Afouras, H. Bull, H. Chowdhury, N. Fox, R. Cooper, A. McParland, B. Woll, and A. Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv preprint arXiv:2111.03635*, 2021.

[5] S. Alexanderson and J. Beskow. Towards fully automated motion capture of signs – development and evaluation of a key word signing avatar. *ACM Trans. Access. Comput.*, 7(2):7:1–7:17, June 2015. ISSN 1936-7228. doi: 10.1145/2764918. URL http://doi.acm.org/10.1145/2764918.

[6] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, and H. Bajaj. Continuous arabic sign language recognition in user dependent mode. *JILSA*, 2, 01 2010. doi: 10.4236/jilsa.2010.21003.

[7] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Quan Yuan, and A. Thangali. The American Sign Language lexicon video dataset. In *CVPRW*, 2008.

[8] B. Bauer and H. Hienz. Relevant features for video-based continuous sign language recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 440–445. IEEE, 2000.

[9] V. Belissen, A. Braffort, and M. Gouiffès. Dicta-Sign-LSF-v2: Remake of a continuous French sign language dialogue corpus and a first baseline for automatic sign language processing. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*,

pages 6040–6048, Marseille, France, May 2020. European Language Resource Association (ELRA).

[10] V. Belissen, A. Braffort, and M. Gouiffès. Experimenting the automatic recognition of non-conventionalized units in sign language. *Algorithms*, 13(12):310, 2020.

[11] F. Bigand, E. Prigent, and A. Braffort. Retrieving human traits from gesture in sign language: The example of gestural identity. In *Proceedings of the 6th International Conference on Movement and Computing*, pages 1–4, 2019.

[12] F. Bigand, E. Prigent, and A. Braffort. Synthesis for the kinematic control of identity in sign language. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 1–6, 2022.

[13] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4462–4470, Santiago, Chile, December 2015. IEEE Computer Society.

[14] M. Borg and K. P. Camilleri. Sign language detection "in the wild" with recurrent neural networks. *ICASSP*, 2019.

[15] C. Börstell, J. Mesch, and L. Wallin. Segmenting the swedish sign language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In *Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages*, pages 7–10, 2014.

[16] P. Boyes Braem and R. Sutton-Spence. *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*. Hamburg: Signum Press, 2001. ISBN 3927731838.

[17] A. Braffort. A gesture recognition architecture for sign language. In *Proceedings of the second annual ACM conference on Assistive technologies*, pages 102–109, 1996.

[18] A. Braffort. Research on computer science and sign language: Ethical aspects. In *International Gesture Workshop*, pages 1–8. Springer, 2001.

[19] A. Braffort and M. Filhol. Sign language: Constraint-based sign language processing. In *Constraints and Language*, chapter 9, pages 191–218. Cambridge Scholars Publishing, 2014.

[20] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Ringel Morris. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *ASSETS '19: The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31, Pittsburgh, PA, USA, October 2019. ACM.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[22] P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968. IEEE, 2009.

[23] P. Buehler, M. Everingham, and A. Zisserman. Employing signed TV broadcasts for automated learning of British sign language. In *Workshop on the Representation and Processing of Sign Languages*, 2010.

[24] H. Bull, A. Braffort, and M. Gouiffès. MEDIAPI-SKEL - a 2D-skeleton video database of French sign language with aligned French subtitles. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, pages 6063–6068, Marseille, France, May 2020. European Language Resource Association (ELRA).

[25] H. Bull, M. Gouiffès, and A. Braffort. Automatic segmentation of sign language into subtitle-units. In *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*, 2020.

[26] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman. Aligning subtitles in sign language videos. In *International Conference on Computer Vision (ICCV)*, 2021.

[27] N. Camgoz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. Content4all open research sign language translation datasets. *ArXiv*, abs/2105.02351, 05 2021.

[28] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. SubUNets: End-to-end hand shape and continuous sign language recognition. In *ICCV*, 2017.

[29] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[30] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden. Multi-channel transformers for multi-articulatory sign language translation. *16th European Conference on Computer Vision (ECCV), ACVR Workshop*, 2020.

[31] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] N. C. Camgoz, B. Saunders, G. Rochette, M. Giovanelli, G. Inches, R. Nachtrab-Ribback, and R. Bowden. Content4all open research sign language translation datasets. *arXiv preprint arXiv:2105.02351*, 2021.

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[35] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[36] X. Chai, H. Wang, and X. Chen. The devisign large vocabulary of chinese sign language database and baseline evaluations. *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS*, 2014.

[37] T. Chan and W. Zhu. Level set based shape prior segmentation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1164–1170. IEEE, 2005.

[38] C. Charayaphan and A. Marble. Image processing system for interpreting motion in american sign language. *Journal of Biomedical Engineering*, 14(5):419–425, 1992.

[39] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200, Portland, Oregon, June 2011. Association for Computational Linguistics.

[40] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua. Temporally grounding natural sentence in video. In *EMNLP*, 2018.

[41] K. L. Cheng, Z. Yang, Q. Chen, and Y.-W. Tai. Fully convolutional networks for continuous sign language recognition. In *European Conference on Computer Vision*, pages 697–714. Springer, 2020.

[42] N. Cherniavsky, R. E. Ladner, and E. A. Riskin. Activity detection in conversational sign language video for mobile telecommunication. In *IEEE International Conference on Automatic Face Gesture Recognition*, 2008.

[43] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015.

[44] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[45] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision (ACCV)*. Springer, 2016.

[46] J. S. Chung and A. Zisserman. Signs in time: Encoding human motion as a temporal image. *arXiv preprint arXiv:1608.02059*, 2016.

[47] H. Cooper and R. Bowden. Large lexicon detection of sign language. In M. Lew, N. Sebe, T. S. Huang, and E. M. Bakker, editors, *Human–Computer Interaction*, pages 88–97, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75773-3.

[48] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2574. IEEE, 2009.

[49] O. A. Crasborn. How to recognise a sentence when you see one. *Sign Language & Linguistics*, 10(2):103–111, 2007.

[50] R. Cui, H. Liu, and C. Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7361–7369, 2017.

[51] D. Dayter. Collocations in non-interpreted and simultaneously interpreted english: a corpus study. In *New empirical perspectives on translation and interpreting*, pages 67–91. Routledge, 2019.

[52] R. De Beaugrande. Sentence first, verdict afterwards: On the remarkable career of the "sentence". *Word*, 50(1):1–31, 1999.

[53] J. Deng and H. Tsui. A two-step approach based on pahmm for the recognition of asl. *ACCV, Jan*, 2002.

[54] K. Desai and J. Johnson. VirTex: Learning visual representations from textual annotations. *arXiv:2006.06666*, 2021.

[55] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2019.

[57] P. Dreuw and H. Ney. Towards automatic sign language annotation for the ELAN tool. In *Proceedings of the Third LREC Workshop on Representation and Processing of Sign Languages*, pages 50–53, Marrakech, Morocco, May 2008. European Language Resource Association (ELRA).

[58] P. Dreuw, J. Forster, T. Deselaers, and H. Ney. Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[59] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[60] A. Duarte, S. Albanie, X. Giró-i Nieto, and G. Varol. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104, 2022.

[61] A. Dutta and A. Zisserman. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279, 2019.

[62] D. M. Eberhard, G. F. Simons, and C. D. Fennig. *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, 22nd edition, 2019.

[63] S. Ebling, N. Camgoz, P. Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, and M. Magimai-Doss. Smile swiss german sign language dataset. In *LREC*, 2018.

[64] M. Erard. Why sign-language gloves don't help deaf people. The Atlantic, https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/, 2017.

[65] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy" – automatic naming of characters in tv video. In *BMVC*, 2006.

[66] I. Farag and H. Brock. Learning motion disfluencies for automatic sign language segmentation. In *ICASSP*, 2019.

[67] A. Farhadi and D. Forsyth. Aligning ASL for statistical translation using a discriminative word model. In *CVPR*, 2006.

[68] I. Feinerer and K. Hornik. *wordnet: WordNet Interface*, 2020. URL https://CRAN.R-project.org/package=wordnet. R package version 0.1-15.

[69] S. S. Fels and G. E. Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *IEEE transactions on Neural Networks*, 4(1):2–8, 1993.

[70] A. Feng, D. Casas, and A. Shapiro. Avatar reshaping and automatic rigging using a deformable model. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pages 57–64, 2015.

[71] J. Fenlon, T. Denmark, R. Campbell, and B. Woll. Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200, 2007.

[72] M. Filhol, M. N. Hadjadj, and A. Choisier. Non-manual features: the right to indifference. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the manual channel. Satellite Workshop to the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 49–54, 2014.

[73] E. Flint, E. Ford, O. Thomas, A. Caines, and P. Buttery. A text normalisation system for non-standard english words. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 107–115, 2017.

[74] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. H. Piater, and H. Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey, May 2012. European Language Resource Association (ELRA).

[75] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, and H. Ney. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, 2014.

[76] S. Gan, Y. Yin, Z. Jiang, L. Xie, and S. Lu. Skeleton-aware neural sign language translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4353–4361, 2021.

[77] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

[78] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude. Motion history images for online speaker/signer diarization. In *ICASSP*, 2014.

[79] B. Gebrekidan Gebre, P. Wittenburg, and T. Heskes. Automatic signer diarization-the mover is the signer approach. In *CVPRW*, 2013.

[80] S. Ghosh, A. Agarwal, Z. Parekh, and A. Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *NAACL-HLT*, 2019.

[81] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2409–2416, 2014.

[82] G. J. Grimes. Digital data entry glove interface device, Nov. 8 1983. US Patent 4,414,537.

[83] J. Gu, H. Hassan, J. Devlin, and V. O. Li. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*, 2018.

[84] P. Guitteny. *Le passif en langue des signes*. PhD thesis, Université Michel de Montaigne-Bordeaux III, 2006.

[85] M. N. Hadjadj, M. Filhol, and A. Braffort. Modeling French sign language: a proposal for a semantically compositional system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 4253–4258, Miyazaki, Japan, May 2018. European Language Resource Association (ELRA).

[86] A. Hao, Y. Min, and X. Chen. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11303–11312, 2021.

[87] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, 2019.

[88] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[89] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li. Signbert: Pre-training of hand-model-aware representation for sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11087–11096, 2021.

[90] H. Hu, W. Zhou, J. Pu, and H. Li. Global-local enhancement network for nmf-aware sign language recognition. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(3):1–19, 2021.

[91] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017.

[92] L. Hu, L. Gao, W. Feng, et al. Self-emphasizing network for continuous sign language recognition. *arXiv preprint arXiv:2211.17081*, 2022.

[93] J. Huang, W. Zhou, H. Li, and W. Li. Attention-based 3d-cnns for large-vocabulary sign language recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(9):2822–2832, 2018.

[94] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li. Video-based sign language recognition without temporal segmentation. In *AAAI Conference on Artificial Intelligence*, 2018.

[95] T. Jiang, N. C. Camgoz, and R. Bowden. Looking for the signs: Identifying isolated sign instances in continuous video footage. *IEEE International Conferene on Automatic Face and Gesture Recognition*, 2021.

[96] T. Jiang, N. C. Camgoz, and R. Bowden. Skeletor: Skeletal transformers for robust body-pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[97] T. Jin and Z. Zhao. Contrastive disentangled meta-learning for signer-independent sign language translation. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[98] T. Johnston. Auslan corpus annotation guidelines. centre for language sciences, department of linguistics, macquarie university, sydney, australia, 2011.

[99] T. Johnston. Lexical frequency in sign languages. *Journal of deaf studies and deaf education*, 17(2):163–193, 2012.

[100] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1943–1950. IEEE, 2010.

[101] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[102] H. R. V. Joze and O. Koller. MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. In *British Machine Vision Conference (BMVC)*, 2019.

[103] M. Kaczmarek and M. Filhol. Use cases for a sign language concordancer. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages*, pages 113–116, 2020.

[104] P. Kapoor, R. Mukhopadhyay, S. B. Hegde, V. Namboodiri, and C. V. Jawahar. Towards automatic speech to sign language generation. In *INTERSPEECH*, 2021.

[105] D. Kelly, J. Mc Donald, and C. Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):526–541, 2010.

[106] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *2011 international conference on computer vision*, pages 169–176. IEEE, 2011.

[107] S.-K. Ko, C. J. Kim, H. Jung, and C. Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683, 2019.

[108] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[109] J. Kolář and L. Lamel. Development and evaluation of automatic punctuation for French and english speech-to-text. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[110] O. Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[111] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015.

[112] O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3793–3802, 2016.

[113] O. Koller, S. Zargaran, and H. Ney. Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4297–4305, 2017.

[114] M. Kopf, M. Schulder, and T. Hanke. Overview of datasets for the sign languages of europe, 2021. URL https://www.project-easier.eu/wp-content/uploads/sites/67/2021/08/EASIER-D6.1-Overview-of-Datasets-for-the-Sign-Languages-of-Europe.pdf.

[115] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[116] L. Leeson. Making the effort in simultaneous interpreting. In *Topics in Signed Language Interpreting: Theory and Practice*, volume 63, chapter 3, pages 51–68. John Benjamins Publishing, 2005.

[117] D. Li, C. Rodriguez, X. Yu, and H. Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469, 2020.

[118] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *NeurIPS*, 2020.

[119] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6205–6214, 2020.

[120] D. Li, C. Xu, L. Liu, Y. Zhong, R. Wang, L. Petersson, and H. Li. Transcribing natural languages for the deaf via neural editing programs. *arXiv preprint arXiv:2112.09600*, 2021.

[121] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[122] R.-H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings third IEEE international conference on automatic face and gesture recognition*, 1998.

[123] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021.

[124] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua. Attentive moment retrieval in videos. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[125] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[126] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[127] S. Lu, S. Igi, H. Matsuo, and Y. Nagashima. Towards a dialogue system based on recognition and synthesis of japanese sign language. In *International Gesture Workshop*, pages 259–271. Springer, 1997.

[128] J. Ma, W. Gao, and R. Wang. A parallel multistream model for integration of sign language recognition and lip motion. In *International Conference on Multimodal Interfaces*, pages 582–589. Springer, 2000.

[129] L. Meurant, M. Gobert, and A. Cleve. Modelling a parallel corpus of French and French belgian sign language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4236–4240, Portorož, Slovenia, May 2016. European Language Resource Association (ELRA).

[130] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[131] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[132] J. Min and M. Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021.

[133] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, Yokohama, Japan, June 2018. ACM.

[134] L. Momeni, T. Afouras, T. Stafylakis, S. Albanie, and A. Zisserman. Seeing wake words: Audio-visual keyword spotting. *BMVC*, 2020.

[135] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman. Watch, read and lookup: Learning to spot signs from multiple supervisors. In *ACCV*, 2020.

[136] L. Momeni, H. Bull, K. R. Prajwal, S. Albanie, G. Varol, and A. Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*, 2022.

[137] J. P. Morford and M. L. Carlson. Sign perception and recognition in non-native signers of asl. *Language learning and development*, 7(2):149–168, 2011.

[138] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan. Real-Time Sign Language Detection using Human Pose Estimation. In *ECCVW, Sign Language Recognition, Translation and Production (SLRTP)*, 2020.

[139] A. Moryossef, K. Yin, G. Neubig, and Y. Goldberg. Data augmentation for sign language gloss translation. In *MTSUMMIT*, 2021.

[140] M. Müller, S. Ebling, E. Avramidis, A. Battisti, M. Berger, R. Bowden, A. Braffort, N. C. Camgöz, C. España-Bonet, R. Grundkiewicz, et al. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, 2022.

[141] K. Murakami and H. Taguchi. Gesture recognition using recurrent neural networks. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 237–242, 1991.

[142] C. Myers and L. Rabiner. A comparative study of several dynamic time-warping algorithms for connected-word recognition. *The Bell System Technical Journal*, 60:1389–1409, 1981.

[143] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1925–1932. IEEE, 2009.

[144] E.-J. Ong, H. Cooper, N. Pugeault, and R. Bowden. Sign language recognition using sequential pattern trees. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2200–2207. IEEE, 2012.

[145] E.-J. Ong, O. Koller, N. Pugeault, and R. Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1923–1930, 2014.

[146] O. Özdemir, A. A. Kındıroğlu, N. Cihan Camgoz, and L. Akarun. BosphorusSign22k Sign Language Recognition Dataset. 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*, 2020.

[147] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[148] I. Papastratis, K. Dimitropoulos, D. Konstantinidis, and P. Daras. Continuous sign language recognition through cross-modal alignment of video and text embeddings in a joint-latent space. *IEEE Access*, 8:91170–91180, 2020.

[149] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[150] T. Pfister, J. Charles, and A. Zisserman. Large-scale learning of sign language by watching tv (using co-occurrences). In *British Machine Vision Conference (BMVC)*, 2013.

[151] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Asian Conference on Computer Vision (ACCV)*, pages 538–552, Singapore, November 2014. Springer.

[152] K. Prajwal, L. Momeni, T. Afouras, and A. Zisserman. Visual keyword spotting with attention. In *BMVC*, 2021.

[153] K. Prajwal, H. Bull, L. Momeni, S. Albanie, G. Varol, and A. Zisserman. Weakly-supervised fingerspelling recognition in british sign language videos. 2022.

[154] K. Renz, N. C. Stache, S. Albanie, and G. Varol. Sign language segmentation with temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[155] K. Renz, N. C. Stache, N. Fox, G. Varol, and S. Albanie. Sign segmentation with changepoint-modulated pseudo-labelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[156] P. Resnik and N. A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

[157] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 993–1000. IEEE, 2006.

[158] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1939–1946, 2013.

[159] M.-A. Sallandre and C. Cuxac. Iconicity in sign language: a theoretical and methodological point of view. *Lecture notes in computer science*, pages 173–180, 2002.

[160] K. P. Sankar, C. Jawahar, and A. Zisserman. Subtitle-free movie to script alignment. In *BMVC*, 2009.

[161] P. Santemiz, O. Aran, M. Saraçlar, and L. Akarun. Automatic sign segmentation from continuous signing via multiple sequence alignment. *ICCVW*, 2009.

[162] B. Saunders, N. C. Camgoz, and R. Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020.

[163] A. Schembri, J. Fenlon, R. Rentelis, S. Reynolds, and K. Cormier. Building the British sign language corpus. *Language Documentation & Conservation*, 7:136–154, 2013.

[164] A. Schembri, J. Fenlon, R. Rentelis, and K. Cormier. British Sign Language corpus project: A corpus of digital video data and annotations

of British Sign Language 2008–2017 (Third Edition). 2017. URL http://www.bslcorpusproject.org.

[165] Z. S. Sehyr, N. Caselli, A. M. Cohen-Goldberg, and K. Emmorey. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277, 2021.

[166] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu. Fingerspelling recognition in the wild with iterative visual attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5400–5409, 2019.

[167] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2984–2991, 2013.

[168] F. Shipman, S. Duggina, C. D. D. Monteiro, and R. Gutierrez-Osuna. Speed-accuracy tradeoffs for detecting sign language content in video sharing sites. *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 2017.

[169] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, Venice, Italy, July 2017.

[170] O. M. Sincan and H. Keles. AUTSL: A large scale multi-modal Turkish Sign Language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020.

[171] Speech Group at Carnegie Mellon University. CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2014.

[172] A. Sridhar, R. G. Ganesan, P. Kumar, and M. M. Khapra. Include: A large scale dataset for indian sign language recognition. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[173] T. Stafylakis and G. Tzimiropoulos. Zero-shot keyword spotting for visual speech recognition in-the-wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–529, 2018.

[174] R. Stamp, A. Schembri, J. Fenlon, R. Rentelis, B. Woll, and K. Cormier. Lexical variation and change in British Sign Language. *PLoS One*, 9(4): e94053, 2014.

[175] T. E. Starner. Visual recognition of american sign language using hidden markov models. Technical report, Massachusetts Inst Of Tech Cambridge Dept Of Brain And Cognitive Sciences, 1995.

[176] C. Stone and D. Russell. Interpreting in international sign: decisions of deaf and non-deaf interpreters. 2011. URL http://hdl.handle.net/2436/624146.

[177] M. Sundermeyer, H. Ney, and R. Schlüter. From feedforward to recurrent lstm neural networks for language modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):517–529, 2015.

[178] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[179] R. Sutton-Spence. Mouthings and simultaneity in British Sign Language. In *Simultaneity in Signed Languages: Form and Function*, pages 147–162. John Benjamins Publishing Company, 2007.

[180] R. Sutton-Spence and B. Woll. *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

[181] S. Tamura and S. Kawasaki. Recognition of sign language motion images. *Pattern recognition*, 21(4):343–353, 1988.

[182] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1471, 2014.

[183] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *Proceedings of International Conference on Multimedia Retrieval*, 2014.

[184] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Book2Movie: Aligning video scenes with book chapters. In *CVPR*, 2015.

[185] V. C. Tartter and K. C. Knowlton. Perception of sign language from an array of 27 moving spots. *Nature*, 289(5799):676, 1981.

[186] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. 2022.

[187] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman. Read and attend: Temporal localisation in sign language videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[188] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[189] O. Veksler. Star shape prior for graph-cut image segmentation. In *European Conference on Computer Vision*, pages 454–467. Springer, 2008.

[190] L. Ventura, A. Duarte, and X. Giró-i Nieto. Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941*, 2020.

[191] V. Viitaniemi, T. Jantunen, L. Savolainen, M. Karppa, and J. Laaksonen. S-pot–a benchmark in spotting signs within continuous signing. In *Proceedings of the 9th international conference on Language Resources and Evaluation (LREC 2014), ISBN 978-2-9517408-8-4*. European Language Resources Association (LREC), 2014.

[192] V. H. Vo, E. Sizikova, C. Schmid, P. Pérez, and J. Ponce. Large-scale unsupervised object discovery. *Advances in Neural Information Processing Systems*, 34, 2021.

[193] C. Vogler and D. Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 156–161. IEEE, 1997.

[194] C. Vogler and D. Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 363–369. IEEE, 1998.

[195] U. von Agris, M. Knorr, and K. Kraiss. The significance of facial features for automatic sign language recognition. In *8th IEEE International Conference on Automatic Face Gesture Recognition*, 2008.

[196] C. Wang, W. Gao, and S. Shan. An approach based on phonemes to large vocabulary chinese sign language recognition. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 411–416. IEEE, 2002.

[197] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, pages 431–445. Springer, 2014.

[198] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021.

[199] W. Wang, Y. Huang, and L. Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019.

[200] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[201] R. B. Wilbur and A. C. Kak. Purdue RVL-SLLL American sign language database. *School of Electrical and Computer Engineering Technical Report, TR-06-12, Purdue University, W. Lafayette, IN 47906.*, 2006.

[202] G. Wojtanowski, C. Gilmore, B. Seravalli, K. Fargas, C. Vogler, and R. Kushalnagar. 'Alexa, can you see me?' making individual personal assistants for the home accessible to deaf consumers. *The Journal on Technology and Persons with Disabilities*, page 130, 2020.

[203] B. Woll. The sign that dares to speak its name: echo* phonology in British Sign Language (BSL). *The Hands are the Head of the mouth: The Mouth as Articulator in Sign Languages, P. Boyes Braem & R. Sutton-Spence (eds.), 87–98*, 2001.

[204] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *AAAI*, 2019.

[205] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, USA, June 2016.

[206] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[207] H.-D. Yang, S. Sclaroff, and S.-W. Lee. Sign language spotting with a threshold model based on conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 31(7):1264–1277, 2008.

[208] R. Yang and S. Sarkar. Detecting coarticulation in sign language using conditional random fields. In *ICPR*, 2006.

[209] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[210] K. Yin and J. Read. Better sign language translation with stmc-transformer. In *COLING*, 2020.

[211] Y. Yuan, T. Mei, and W. Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *AAAI*, 2019.

[212] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, and C. Gan. Dense regression network for video grounding. In *CVPR*, 2020.

[213] J. Zhang, W. Zhou, and H. Li. A threshold-based hmm-dtw approach for continuous sign language recognition. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 237–240, 2014.

[214] H. Zhou, W. gang Zhou, W. Qi, J. Pu, and H. Li. Improving sign language translation with monolingual data by sign back-translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[215] H. Zhou, W. Zhou, Y. Zhou, and H. Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[216] Z. Zhou, V. W. Tam, and E. Y. Lam. Signbert: A bert-based deep learning framework for continuous sign language recognition. *IEEE Access*, 9: 161669–161682, 2021.

[217] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, 2016.