M1 Computer Science 2018-2019
Internship Report:

Sign Language Processing

Student:
Chloé Paris

Professor:
Jong C. Park
Referent:
Jung-Ho Kim

ENS DE LYON

KAIST

August 26, 2019

Contents

1. Introduction

The main focus of this internship was *sign language processing*. It was an incredible opportunity to learn about many aspects of the research world:

- Discovering a research lab in a foreign country:
  - Getting a better understanding of the research process around the world;
  - Learning from researchers with a different approach and educational background;
  - Settling into a new research and cultural environment;
- Learning about the full process of research:
  - Studying the previous findings in the field and current research angles;
  - Finding a new research direction;
  - Constantly dealing with setbacks and redefining the research plan;
- And what is inherent to the process:
  - The paperwork;
  - The academic rules (conditions to producing publishable work, ethical reviews);
  - The meetings and presentations, reporting on the progress;
  - The collaboration and mutual help between lab members.

1.1 Research lab

The host institution was the Korean Advanced Institute of Science and Technology (KAIST) in Daejeon, South Korea. Its NLP lab has been working for many years in a variety of subjects in the Natural Language Processing field, each member focusing on one particular subject. These subjects include processing language for:

- mental illness detection;
- statement reliability assessment;
- emotion classification from dialogues;
- biomedical text mining;
- event extraction;
- argument generation;
- sign language processing.

In particular, Jung-Ho Kim, the PhD student I worked with, is currently working on Sign Language processing, and Professor Park, the supervisor for my internship has been publishing papers related to Sign Language processing for more than a decade.

Since the lab is in Korea, their first interest is Korean Sign Language (KSL). For technical reasons, notably the existing data set, they have been working on other sign languages, especially German Sign Language (DGS) but never on French Sign Language (LSF).

1.2 Subject

Being a student in both Computer Science and Sign Language interpreting, it is only natural to try and build a bridge between these two domains. Languages have always been a passion of mine and technology offers us, with computers, a sophisticated tool to analyze and process languages in a way human beings cannot (and the reverse is also true, computers can't comprehend languages in the way human beings do), which makes computer science the perfect complementary companion for anyone interested in languages.

Thus, I was eager to learn about all the techniques that exist to automatically process languages, focusing primarily on sign languages because they hold an additional dimension: where spoken languages are constrained to linearity in time because it is not possible to produce two different sounds at the same time, gestures are not.

1.2.1 Motivation

Apart from personal interest and the fact that analyzing languages proves useful at least in the teaching world, automatic sign language processing is beneficial to the people using sign language for several reasons.

Sign language is the only way for deaf and hard of hearing people to have full accessibility to the world around them. It is discriminatory not to allow them the same access both *to* sign language and to anything *in* sign language. The amount of information, human connection and food for thought that they are denied because of a hearing impairment is huge, inhuman and unjustifiable considering the simple teaching and using of sign language would relegate this handicap to a mere difference.

One praiseworthy but definitely *not* stand-alone direction to remedy the hearing impairment this day is the tentatives to make them more like hearing people. Though it is great to want to give them better access to things that are sound-oriented, it is not currently working well enough to ignore sign languages.

It is important to note that (contrary to popular belief), *even in their written form*, spoken languages are *not* the key to accessibility for deaf people. They can help sometimes but hearing impairment implies partial to no access to spoken languages, even with the help of the best technology, hard of hearing children and adults are not exposed to the same quantity, nor the same quality of audio input. Thus, many of them do not develop sufficient dexterity in the local spoken language to use it as a means of accessibility, even in the written form. And even though deaf children are taught written French in France (for instance), because there is no research on the teaching method (this idea is very well explained in this book: [Séro-Guillaume, 2008]), 80% of them are considered illiterate (with a very specific form of illiteracy, close to the difficulties that dyslexic people encounter; they tend to know words, but not be able to make sense sentences).

Sign language processing is a great tool to remedy the lack of accessibility for deaf people: they have a truncated access to the news, culture, education, then sign language translation, avatar production and better diffusion are the ways to go. Avatar production would a great tool to offer deaf people anonymous expression (a thing that any hearing people gets to do without noticing and that is not a possibility for deaf people). Creating a written form of sign languages using the more dynamic approach that computers offer (because it seems that pen and paper are too limited for sign languages) would allow for a greater production of learning material, a greater evolution, maturing (in the sens that sign languages are "younger" languages in their state of development compared to spoken languages, mainly because of the lack of written form) of sign languages (by adding a new dimension to them, just like written and oral forms of spoken languages differ), and also new ways to communicate, produce art, takes notes, connect and be part of the world.

As mentioned, sign languages processing also holds promises of better diffusion of the language because it would allow the gathering of more information and the creation new kinds of resources for people willing to learn it: deaf people, people with other disabilities (autism, speech impediment, aphasia) and their close relatives but also more exposure for anyone, meaning better communication in life for the afore mentioned people, and for those who learn, it is good for their brain, for open-mindedness as well as global communication.

All in all, it holds interesting and very broad challenges for computer science research and any result is a step in the right direction to add material to research in the linguistics field.

After many brainstorming sessions and reading scientific articles, several research directions emerged.

1.2.2 Translation between French and Korean Sign Languages (both ways)

No research seems to have been conducted on the translation between two sign languages. In particular, for my internship, the two most relevant languages were Korean and French Sign Languages (KSL and LSF). France and Korea are almost as far away culturally as two countries can be on this earth, so choosing them also made sense because these two sign languages would be as far away as possible from each other (since mutual influences were kept at a strict minimum for two sign languages on this planet).

The action plan for this was to study and formalize the differences between the two languages in order to use them in the creation of a translation tool.

The translation would be focusing on gloss writing of KSL and LSF, i.e. writing the names of the signs (in English, Korean or French) in the order they are used to represent a sentence in sign language (each sign is conventionally written between brackets). Thus, no video recognition is necessary for this research.

1.2.3 Focusing on the visual aspect

The translation part is focusing on gloss and I knew from the beginning that this would imply mistakes and limitations that did not feel right. I also wanted to discover more about the image processing part and work on the visual aspect, because, without it, it is not sign language and because of a side project of mine on sign language processing that made me aware of the necessity to learn about this aspect as well in order to really understand the sign language processing field.

While reading articles to learn more about the subject as a whole, I felt inspired and passionate (this is a field that is close to my heart and the articles kept giving me new things to think about and opening my mind about it). After a while, I was able to extract two research directions that seemed to make sense and had one thing in common: they were about the visual meaningfulness of sign languages.

I wanted to analyze sign language data to extract what was visually meaningful in it, in a way I wanted to answer in details the question that many hearing people with no knowledge of sign language stumble upon when they see a sign language interpreter working from speech to sign: "Why do I feel like I understand the sign language even though I don't know it?".

To me the answer lies in two separate elements of sign languages:

- the vocabulary being inspired by real world actions and shapes when possible;
- the inherent grammatical use of mime, role taking and impersonation.

In terms of application of the results, they would open the way to formalizing a more comprehensive way to write gloss (it is easily understandable that a mime would be complicated to note in a gloss, but if we extracted a list or more likely a pattern, we could define a way to annotate them), they might open possibilities to simplify the translation from one sign language to another (depending on how constant they are internationally) and allow the creation of tools for non signers to become more efficient non verbal communicators.

2. State of the art

Any scientific work needs to be preceded by a review of the state of the art. Without it, the researcher would just be repeating already implemented work, especially since the techniques are now so developed in every single subject.

When my team was researching existing work about sign language processing during the year for our integrated project (creating a writing tool for sign languages), we did not find very advanced papers so I thought that sign language processing was just not a very active field in computer science. Thanks to this internship, I found out that this was not true at all. Being guided towards some relevant articles, I kept finding more and more papers on any subject surrounding sign language processing. These articles were enlightning on the various possibilities, made me think much further but were also quite frustrating (either in the way they seemed to neglect key aspects of sign languages or their title was too promising for the actual results), but frustration is a good seed for motivation.

Here are some of the most significant and latest publications on the subject, illustrating the state of the art at the beginning of my internship.

2.1 Machine Translation

One of the first subjects that come up when dealing with natural language processing is machine translation. There is Neural Machine Translation (NMT) and Statistical Machine Translation (SMT).

2.1.1 Oral languages

The research is way more advanced with languages of wider currency that present extremely large and varied corpora. Sign languages are not languages of wider currency, but because it is the same discipline, it is worth looking into what exists for spoken languages. We all know the quality and limits of machine translation (and actually everything in natural language processing) for spoken languages because we experience it almost everyday on the internet. So I will not get into too much detail about these because information is extremely easy to find on the subject.

The usual method is sequence to sequence (seq2seq) learning with neural networks ([Sutskever et al., 2014]). Sequence to sequence means that the data set is made of sentences (sequences) in the two languages, paired by identical meaning. This way avoids word for word translation (everyone knows it does not work).

Many recent papers have shown that injecting domain specific expert knowledge can improve machine translation. In terms of natural language processing, linguistic analysis is the way to go, and that is what grammars are for [Hockenmaier and Steedman, 2007].

2.1.2 Sign Language

I talked very briefly about oral languages because this was just a reminder that this research is what guides us when dealing with the exception that sign languages are. What is more interesting is the state of the art in the subjects that are more specific to sign language:

2.1.2.1 Gloss

Sign language translation is often based on gloss writing of the sign language. This allows researcher to simulate a written language. However it has drawbacks [Michael Erard, 2017]: not only is the gloss not totally compatible with the multi-dimensional aspect of sign language (see Annex for more details), it fails to acknowledge a crucial fact (sign language is purely oral, not written) and encourages researchers to deal with an oral expression as if it was written language, when the two instances are fundamentally different.

Some paper try to deal with these questions and how better to consider sign language as a whole instead of focusing only on manual features[Filhol et al., 2014]. This paper created a model called *AZee* to describe sign language without the usual restrictions.

2.1.2.2 Neural Networks

Neural Networks[Cui et al., 2017]Recurrent convolutional neural networks for continuous sign language recognition by staged optimization;

This is just one example. There are many papers, each describing its own neural network architecture, all of them with similar performances. But the thing is, when they announce in the title that they do continuous sign language recognition, they always just do a restriction of it. give examples

2.1.2.3 Grammar

Many papers have discoursed on the use of grammars specifically in the case of sign language processing. Some authors developed ways and methods to model sign languages specifically.

[Filhol and Falquet, 2017] is one of them. They present a way to build grammars and create linguistic input for the production of sign language (with an avatar).

[Hadjadj, 2017] is a thesis on the subject of making a model for French Sign Language using semantic information.

2.2 Video processing

2.2.1 Image recognition

Some very useful tool are implemented for sign language recognition.

OpenPose[Cao et al., 2018] is not specifically written for sign language recognition, it extracts the coordinates of reference points (keypoints) on the body, face and hands for people in the picture, producing what looks like a skeleton (see illustrations in Annexe C). It is exploited a lot in works on sign language recognition because it bypasses skin segmentation problems and other issues and directly outputs a JSON file with all relevant data for sign language recognition.

[Book and Fisher,] is an example of a method that uses OpenPose. It presents a supervised learning framework comprised of CNNs (convolitional neural networks) and RNNs (recurrent neural networks) whose goal is ASL recognition from raw and head-on videos of a single person performing one sign. They extract frames from the videos, use OpenPose to extract the keypoints and the use the RNN to predict what sign is being performed. They only train on ten different ASL words (with 20 videos of the same signs to get rid of one-time specificities) and perform correct predictions on 93% of never seen before videos. This performance is very high, but it is only ten different signs. The approach is very interesting but can hardly be generalized at a greater scale by lack of data sets (they created the 200 videos that they used for this).

Though OpenPose is a very well-rounded tool, it has a more general purpose and some more specialized tools exist that provide better results on more specific tasks.

One of these tasks is facial expression recognition. OpenFace[Baltrusaitis et al., 2018] is a very interesting tool for facial expression recognition from 2D pictures or videos and I will talk about it more later in the report. As for 3D analysis, [Tarnowski et al., 2017] is a work on the recognition of seven emotional states (neutral, joy, sadness, surprise, anger, fear, disgust). They tested both a k-NN classifier (3-NN to be exact) and MLP neural network. They trained the neural network using back propagation algorithm with conjugate gradient method and evaluated their results both in a subject-dependent angle (separating the data according to the people appearing in it) that gave an average result of 90% accuracy for the MLP and 96% accuracy for the 3-NN and in a broader subject-independent manner that gave an average result of 95.5% accuracy for the MLP and 75.9% accuracy for the 3-NN. They also rated the emotions by difficulty of recognition and found that the two most difficult ones were sadness and fear (they were often confused respectively with neutral and surprise). However they nuanced their results saying that they were probably affected by the small number of facial expressions used.

2.2.2 Continuous time

Going one step further from research that focuses on single-word videos, some papers explore the possibility to deal with videos containing sequences of signs (sentences). One option is to use deep learning.

That's what [Camgoz et al., 2017] does. They work in a sequence to sequence fashion, proceeding first with alignment and then recognition. The novelty in their approach is that they decompose the chore by using subunits: they introduce series of specialized expert systems that they call "SubUNets" and implicitly perform transfer learning between tasks. In doing so, they imitate in a way how human beings learn.

2.2.2.1 Dynamic Time Warping algorithm

The Dynamic Time Warping algorithm is interesting because it offers the possibility to align in time videos that were initially following different rythms, which is one problem when processing sign language videos.

One researcher goes much further in exploiting this algorithm and some of her publications are about sign language processing, and, in particular, the non manual features of sign language.

In one of her most recent publications on the subject [Kacorri et al., 2016], Kacorri starts from non-manual expressions in American Sign Language to select a specific examplar using the Dynamic Time Warping method with good results.

In another one [Kacorri and Huenerfauth, 2016], the Continuous Profile Models is used to identify a trace of the performance, with better result according to deaf users and metric for animations.

2.2.2.2 Sign separation

When dealing with sentences in videos, one useful ability is the one to partition the sentence into a sequence of words (find the separation points between the signs).

One paper on sign language recognition that tackles this issues is [Mocialov et al., 2017] They segment the video stream using LSTMs for automatic classification of the derived segments. The accuracy of the segmentation is $> 80\%$.

Another version of this article explains the end to end version of their work: [Mocialov et al.,]

2.3 Noted difficulty

Usually in the natural language processing field, very large data sets are required to draw results from. However, in the case of sign language processing, many authors complained about the lack of data and had to settle on using poorly suited data or creating their own (i.e. very small) data set. That explains also why the results were sometimes frustrating: very good ideas had to be twisted to be able to obtain measurable results considering the limitations in data. Authors also noted that in most cases, when there was data, it only contained isolated signs and the data sets containing sequences were even fewer.

3. KSL - LSF translation

3.1 Context

Sign languages are the most natural way of communication for deaf and hard of hearing people. In France, the language in use is the French Sign Language (LSF) and in South Korea, it is the Korean Sign Language (KSL).

In fact, to the surprise of many hearing people, sign language is not universal, each country having its own sign language (and sometimes even several of them), simply because sign languages appear and grow in the exact same way spoken languages do: through interactions among geographically close people. However, like Esperanto for the global hearing community, there exists an international sign language (ISL). Just like its spoken counterpart, ISL is not widely known in the signing community, and though it is more in use than Esperanto, it has not reached the effectiveness that the English language has attained for global communication, a point that was confirmed during my stay in Korea through many attempts to communicate with local signing people.

That is why we wished to implement a translation system between two sign languages, namely the French and Korean Sign Languages (KSL and LSF) for obvious reasons.

Moreover, despite looking for similar tools, it seems they have yet to exist. We could foresee that there would be more to this kind of translation than simply applying the same methods that work quite well when not dealing with two sign languages.

The final goal is a two-way translation system between these two languages.

However, initially, the focus is on translating from LSF to KSL because we will need reliable data in the source language in order to test the translation result in the target language. And while we have access to both signing communities (French and Korean) to review the resulting sentences, up to this point, only LSF can reliably provide us with starting point sentences (because of my expertise in the language and deaf matters, which allows me to find reliable resources and check the content if needed).

3.2 What makes it unique

This is machine translation, which means we can learn from what exists in this field. However, it has some unique qualities that stem from the fact that the two languages are sign languages and it creates additional challenges and new possibilities and thus a necessity to explore new options as well.

One challenge is the lack of data in both languages. Another is the fact that sign languages are less "mature" than other languages, they don't have a written form, and orality in languages is a lot more chaotic than writing, they also have less vocabulary, forcing signers to find strategies to communicate in spite of a lack of signs.

The new possibilities lie in the fact that sign languages have more similarities to one another than spoken languages do and this is what we will try to make good use of in our approach.

3.3 Method

3.3.1 In preparation

In order to find the best approach to this translation, I spent time thoroughly reading various articles on machine translation and sign language processing, trying to relate them to the task at hand.

Meanwhile, in order to know the material we wanted to work on, we spent a lot of time studying Korean Sign Language from two books that I cannot cite here because any writing on them was in Hangeul (Korean writing). The first one was addressing KSL learners from a beginner level. It introduced basic vocabulary, sentences for everyday situations and information on KSL use and other cultural pillars for the Korean Deaf community. The second one was written for Korean Sign Language interpreters in training and more advanced level students. It contained a much wider range of KSL sentences and vocabulary as well as some explanations.

Studying these books was always done concurrently with note taking on underlying structures of the language and an effort to compare it to my knowledge of the rules and use of French Sign Language.

We also looked for all the existing data sets and possibilities to gain access to them. I showed the reliable LSF resources I knew and it was confirmed that these would be a good starting point but the academic process required us to check with the owners of these before using them and access to new, very large and game-changing KSL data was expected in a short time.

3.3.2 Initial hypothesis

From my theoretical knowledge of sign languages around the world and the comparison of LSF and KSL, I settled on a first hypothesis:

> *In order to translate from one sign language to another, if one only translates the lexical signs while keeping the word order and the grammatical elements unchanged, then result will be correctly understood and possibly perceived as quite natural.*

This hypothesis can be understood by refering to the description of sign language in the introduction. According to the established notions this would mean keeping the syntax and grammar and translating the vocabulary.

This hypothesis was to be improved with the progress of our work and the findings we would make along the way but it seemed fitted as a starting point. The first milestone from there was to test the hypothesis. I had to find a way to do so.

3.3.3 Testing the initial hypothesis

In order to test the hypothesis, I created 17 sentences in French. These sentences had to be questions that could be answered unambiguously in preferably one word or at least a few words. The intended process being to first translate them into LSF and then from LSF to KSL using the hypothesis in order to survey the Korean signing community, we needed their answers to be minimally impacted by factors such as not knowing the answer to the question or not being good at expressing themselves. The only determining factor in the correctness of the answers had to be their understanding of the question. To assure this, I kept the French Deaf people I know in mind, trying to find questions that I knew for sure they would all be able to answer correctly without even thinking about it.

The designed questions are the following:

- What season comes after summer? → autumn
- What can you watch in general in a movie theater? → movie
- What is the thing that is saved in banks? → money
- How many is 2 plus 3? → 5
- Who is the current president of the United States? → Trump
- What is the most important food to make a sandwich ? → bread (added items are possible)
- How many months are there in a year? → 12
- What is the word to describe someone who cannot see? → blind (this question was deleted because considered a sensitive subject)
- What is the word to describe someone who speaks two languages? → bilingual
- In relation to you, who is the father of your father? → grandfather
- What do babies drink except water? → milk
- Alice stayed at Bob's from 10 am to noon and Bob is not Alice's doctor. Did Alice miss her 10h30 doctor's appointment? → yes
- When Bob and Alice are together, he keeps smiling. Does Bob like Alice? → yes
- Alice gave Bob an apple at 9am and he ate it 3h later. When did Bob eat the apple? → noon
- If I don't work, would my boss be bothered? → yes
- If I don't work, wouldn't my boss be bothered? → yes, he would
- What has 5 fingers but no skin and no bone ? → I'll let you guess this one[1]

The sentences were rated by difficulty (to later evaluate the answers with more nuance), translated into LSF and then transposed into KSL by applying the hypothesis. The survey would also let people give their opinion on how easily understandable the question was to them, how natural it sounded and how better to formulate it. This was supposed to help us elaborate a more sophisticated hypothesis. Indeed, while studying KSL, I noticed that there were some differences, in particular:

- Word order inversion:
  - LSF usually puts the pronoun after the noun ([brother]+[my]) while KSL does the opposite ([my]+[brother]);
  - LSF usually puts the adjective after the noun ([apple]+[red]) while KSL does the opposite ([red]+[apple]);

  this kind of word inversion was possibly harmless and that is why it was interesting to not correct these "mistakes" in the first attempt at a translation and see what the reactions would be. Indeed, in LSF the above mentioned rules only apply in the minds of people who know the rules, but actual users of the language do not apply them, resulting in a seemingly orderless syntax, and it was possibly the same with KSL, which would make these inversions useless.

[1] it's a glove

- Some sentences presented in KSL did not make any sense to me as a French person and thus the translation into LSF would be a challenge;
- and some expanding or compressing seemed to be happening when one notion didn't have a sign in one language but did in the other (for instance, there is a sign for [big brother] in KSL but LSF uses the phrase [brother]+[big]).

It would be interesting to find the precise rules to apply while still keeping the hypothesis mostly intact but it would not necessarily be possible to keep it so simple and that was what the survey was supposed to tell us. If it was possible to keep the hypothesis, with just some adjustments, then it would make for a higher quality translation system relying on what has been found between sign languages: that they are very close to one another because of their visual nature (which entails a singular way of thinking) and only differ in vocabulary (especially for more abstract things).

I kept reading articles to know more about machine translation in the particular case where one of the languages is a sign language while waiting for the survey to be approved. However, I learned that, because it would potentially reach Deaf people, who are considered to be vulnerable subjects, the survey had to be reviewed by ethical commissions (from both institutions) before we could use it and we had to write down the project to submit it.

The ENS approved it but no information from the Korean ethical commission has come my way (application to *KAIST IRB's 4th Regular Review* was made on July 10^{th}). All I know is that we haven't started the survey yet (there is still hope though, the internship won't be over until the end of August, and after that we want to keep in touch and keep making progress in order to maybe publish something in the end). We had to move on to other directions without getting the desired answers on this matter.

3.3.4 New direction

Because we did not have access (or more precisely we were waiting for the authorization to get it) to the feedback from the Korean Deaf community, or to any Deaf community in the world (since there was still the obligation to get approved by the Korean ethical commission for any feedback from any deaf person), the only language we could assess in the way we wanted was LSF. The problem now lied in the production of KSL sentences that we could tentatively translate into LSF. For this, we went back to the two KSL books in order to find as many example sentences as possible in KSL.

We used the book *Introduction to KSL interpretation* and drew KSL - Korean sentence pairs considering various grammatical rules to have as broad of a sample as possible (simple, complex and compound sentences; various types of sentences, tenses, and negations). We found 163 of them (cf "KSL to LSF" tab) and took note of both the meaning of the sentence and the gloss of the KSL.

A two-step program was then applied to these sentences:

- I translated the meaning into French and LSF;
- I translated the KSL gloss into LSF applying the hypothesis.

3.3.4.1 Comparison result

The first step was meant to give us a clear comparison between the two languages.

The second step was there to evaluate the hypothesis: I was able to separate the sentences in several categories:

- perfectly understandable and actually constitutes a completely valid LSF sentence (1);
- perfectly understandable but some changes could be applied to make it seem more natural in LSF (1);
- somewhat understandable (.5);
- ambiguous (0);
- not understandable as is (0).

Attributing a score of 1 to the first two categories, .5 to the third one and 0 to the two last ones, I counted how many sentences were a satisfying translation using the hypothesis. Transforming the result into a percentage, it appears that **the hypothesis is good enough for 67.6% of the sentences**.

3.3.4.2 A few words about this result

This percentage (around 60%) appears quite often when talking about sign languages:

- in sign language interpreting school, we were repetitively told (and I read some memoirs that did indeed compute this percentage) that numerous studies found that deaf people can understand about 60% of what is translated by a professional sign language interpreter (for a variety of reasons, including the lack of knowledge of their own language -since most of them learn it later in life and are never taught anything about it, unlike French hearing kids who study French for more than a decade in school- or the lack of knowledge of the subject -because accessibility for deaf and hard of hearing people is a huge issue and thus they do not get nearly the same exposure as hearing people do to a variety of subjects- and also more technical considerations all related to the fact that they need to be able to properly see the interpreter at all times and keep focus watching and listening -which is also more difficult to do through the visual canal and even harder because they are not used to listening for long periods of time because of the usual lack of accessibility);

- I also heard and read several times that two signers from two European countries speaking two distinct European sign languages would be able to understand about 60% of each other's sign language on the first encounter (because the visual way to say things is somewhat unique).

The first point means that, in some way, producing a translation tool with a good performance in 67.6% of the time is unfortunately actually already outperforming the daily best-case-scenario understanding that deaf and hard of hearing people get. However, because we are looking to make an actual improvement in accessibility for deaf and hard of hearing people, and anything less than what hearing people get still equates discrimination, we should not settle for 67.6%, especially since such a simple hypothesis lead to a quite good result, that means we can do better by pushing further, and using linguistic analysis is the right way to produce a good translation system.

With the second point, it is clear that this simple hypothesis is already enough to outperform the mutual understanding that two foreign deaf people can reach without actually learning one another's language (especially since we are dealing with two culturally very different countries -France and Korea-, unlike European countries that have more common points with one another, and whereas 60% understanding is actually what it feels like to me when meeting foreign deaf people, with Korean signers, before I started learning Korean Sign Language, it felt more like 10% understanding).

The other thing to mention about this number is that I have some reservations about its actual accuracy (it might underestimate the efficiency of the hypothesis) and they only grew while progressing further into the work. I will talk about it more in depth later in this report, but my main concern is the quality of the sample sentences. They might not be a good representation of the real Korean Sign Language because the book we drew them from seems to be somehow adapted to hearing learners of KSL, thus deforming the language to make it easier to learn for Korean speaking students. We have no way to access enough of the real KSL to evaluate the difference it might make and my discussions with the Deaf KSL teacher on the matter were not admissible to refute what is written in a book. Still, to me, it felt wrong, and I kept trying to learn about the real KSL (and succeeding but not enough to become an expert or for my protestations to be admissible).

3.3.5 Grammar

In natural language processing, one needs to either have a lot of data to learn from or a precise grammatical understanding of the language (and its formal definition) in order to obtain results that make any sense. To better understand the transformations that had to occur to translate accurately between the two languages and prepare the implementation of the translation tool, we decided to formally define the grammars of the two languages and compare the parsed results (from the defined grammar). The concept of a formal grammar was first developed by Chomsky ([Chomsky, 1956], [Chomsky and Lightfoot, 2002]).

3.3.5.1 Preliminary observations

Based on the sample sentences and other knowledge, we first established some observations about the two languages.

Characteristics of KSL:

- Word order: SOV (e.g., *[son]+[teacher]+[become]*)

- No 'be' verbs (e.g., *[he]+[student]*)

- Fusion of S+V and O+V (e.g., *[sun(rising-movement)]* and *[pass-the-exam]*)
- Omitting unnecessary words (e.g., *[I]+[dormitory]* and *[woman]+[friend]+[meet]*)

Characteristics of LSF:

- Date + Place + Protagonists + Action + Consequences (e.g., *[yesterday]+[place]+[France]+[son]+[teacher]+[become]+[happy]*)
- Simple sentence: SOV (e.g., *[son]+[teacher]+[become]*)
- No 'be' verbs usually (e.g., *[he]+[student]*)
- Possessive pronouns and adjectives come after the noun: context before details (e.g., *[brother]+[my]+[younger]+[smart]*)
- Question word always at the end of the question (e.g., *[summer]+[then]+[what?]*)
- Precision on action state: right after the verb (e.g., *[not], [not yet], [never happened], [very soon]*)
- Possible fusion of S+V and O+V (e.g., *[sun(rising-movement)]* and *[eat-apple]*)
- Omitting unnecessary words (e.g., *[apple]+[eat-apple]*)

3.3.5.2 Context-Free Grammar (CFG)

The repository for this work can be accessed with the following gitlab link.

3.3.5.2.1 What is a CFG? A context-free grammar (CFG) is a type of formal grammar where rules can be applied regarless of context. It is built with four elements:

- a finite set of *terminal* symbols;
- a finite set of *nonterminal* symbols (disjoint from the previous set);
- the *start symbol* S, a nonterminal symbols;
- a finite set of production rules that are pairings of one nonterminal symbol with a sequence of terminal and non terminal symbols.

The sentences of the language can then be derivated from S with successive applications of the rules until they are only filled with terminal symbols.

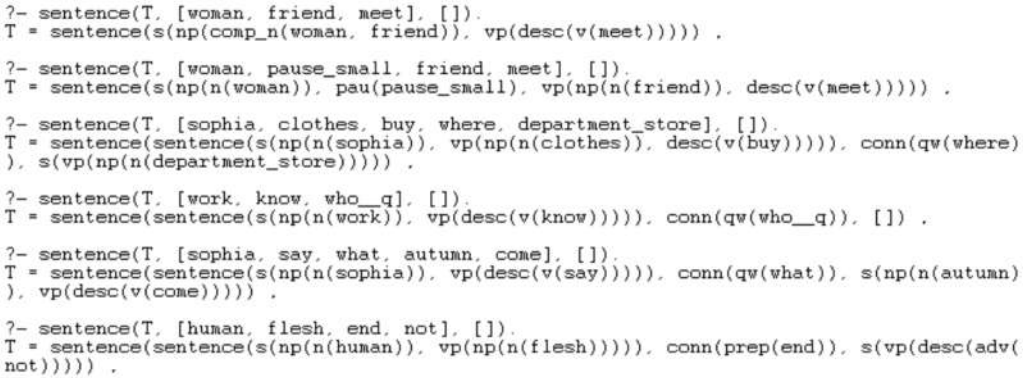
3.3.5.2.2 CFG implementation With Jung-Ho Kim, we wrote the CFG rules for KSL and LSF (he did the KSL and I did the LSF). They can be found in the *hand-written* folder of the repository.

From this, the next step was to write the actual parser:

- Grammar: Context-Free Grammar (CFG);
- Language: Prolog (SWI-Prolog ver. 8.0.3);
- Progress: tested on all sentences for both languages;
- Folder: *cfg* folder of the repository.

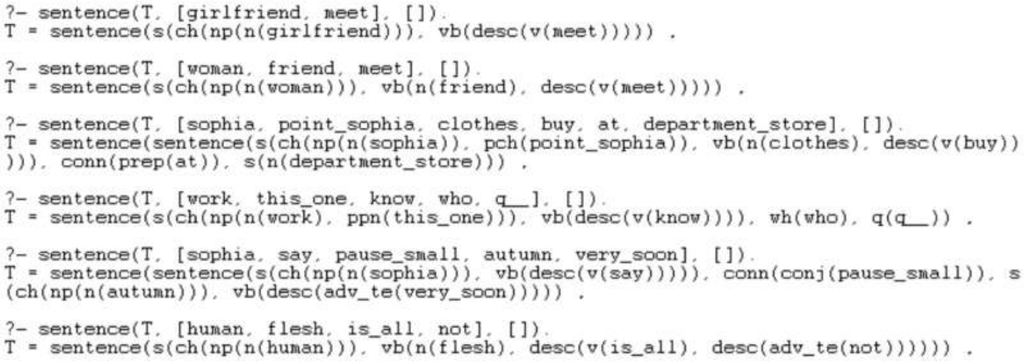
3.3.5.2.3 Results For illustration, here are six sentences (you can see from the unusual feeling of the sentences another reason why I have reservations about them: many are not the kind of sentences people usually say, instead they look exactly like what they are, sentences from a language learning book) and the parsing results on them can be seen in the Figures 3.1 and 3.2:

- *Meet my girlfriend.*
- *A woman meets her friend.*
- *Sophia bought clothes at a department store.*
- *Does anyone know this work?*



```
?- sentence(T, [woman, friend, meet], []).
T = sentence(s(np(comp_n(woman, friend)), vp(desc(v(meet))))) .

?- sentence(T, [woman, pause_small, friend, meet], []).
T = sentence(s(np(n(woman)), pau(pause_small), vp(np(n(friend)), desc(v(meet))))) .

?- sentence(T, [sophia, clothes, buy, where, department_store], []).
T = sentence(sentence(s(np(n(sophia)), vp(np(n(clothes)), desc(v(buy))))), conn(qw(where)
), s(vp(np(n(department_store))))) .

?- sentence(T, [work, know, who__q], []).
T = sentence(sentence(s(np(n(work)), vp(desc(v(know))))), conn(qw(who__q)), []) .

?- sentence(T, [sophia, say, what, autumn, come], []).
T = sentence(sentence(s(np(n(sophia)), vp(desc(v(say))))), conn(qw(what)), s(np(n(autumn)
), vp(desc(v(come))))) .

?- sentence(T, [human, flesh, end, not], []).
T = sentence(sentence(s(np(n(human)), vp(np(n(flesh))))), conn(prep(end)), s(vp(desc(adv(
not))))) .
```

Figure 3.1: KSL grammar parsing



```
?- sentence(T, [girlfriend, meet], []).
T = sentence(s(ch(np(n(girlfriend))), vb(desc(v(meet))))) .

?- sentence(T, [woman, friend, meet], []).
T = sentence(s(ch(np(n(woman))), vb(n(friend), desc(v(meet))))) .

?- sentence(T, [sophia, point_sophia, clothes, buy, at, department_store], []).
T = sentence(sentence(s(ch(np(n(sophia)), pch(point_sophia)), vb(n(clothes), desc(v(buy))
))), conn(prep(at)), s(n(department_store))) .

?- sentence(T, [work, this_one, know, who, q__], []).
T = sentence(s(ch(np(n(work), ppn(this_one))), vb(desc(v(know)))), wh(who), q(q__)) .

?- sentence(T, [sophia, say, pause_small, autumn, very_soon], []).
T = sentence(sentence(s(ch(np(n(sophia))), vb(desc(v(say))))), conn(conj(pause_small)), s
(ch(np(n(autumn))), vb(desc(adv_te(very_soon))))) .

?- sentence(T, [human, flesh, is_all, not], []).
T = sentence(s(ch(np(n(human))), vb(n(flesh), desc(v(is_all), desc(adv_te(not)))))) .
```

Figure 3.2: LSF grammar parsing

- *Sophia said, "Oh! here's autumn".*
- *A human is not made of flesh.*

From the parsing results and my theoretical knowledge of LSF linguistics, these results seemed quite satisfying. I felt really happy with how the grammar turned out for the parsing of the sentences. However, for the purpose of building a translation tool, this was not sufficient.

According to the theory of interpretation ([Lederer and Seleskovitch, 2014]) that I studied in ESIT (Ecole Superieure d'Interpretes et Traducteurs), linguistic translation does not make for the best kind of translation (they instead advised a two-step approach: first, extract the meaning, forget the sentence in the source language; second, produce a sentence in the target language with the meaning in mind) and I was afraid we would go in the wrong direction.

When I learned about Combinatory Categorial Grammars and their specificities, it was reassured that this was a type of grammar that would allow to keep going in the right direction.

3.3.5.3 Combinatory Categorial Grammar (CCG)

The next step was converting the grammars from CFG to CCG. This would allow have the benefits of controlling the range of non-manual expressions using *context* and exploiting them as grammatical information such as types of tenses (*meet/MOUTH_PA* in KSL where the labial information implies a past action) and sentences (*go/?* where eyebrows indicate that it is a question and what type of question), and spatial information ($change_{a \to b}$ where the placement of the sign informs as to what is subject and what is object in the sentence).

The comparison of the two CCG grammars was supposed to lead to the validation or improvement of the hypothesis, followed by the implementing of a translation system.

CCG is a new type of grammar, prefered to others nowadays for natural language processing.

While reading [Blackburn, 2005], the similarity between the formal representation and the way it allows the computer to deal with language on the one hand and the recommended translation process for human beings (involving what is called "déverbalisation", which means keeping the meaning and forgetting about the wording) was clear. This made me think that CCG would allow the computer to parse a sentence into a logical proposition that only remembered the "idea" (i.e. the protagonists and objects involved and how they interacted) and forget the linguistical information.

In order to understand how it was supposed to work, I did the exercises proposed in the book, that let me take on the role of the computer and write the logical propositions from the sentences. The concept seemed to work easily and was quite powerful.

Another book [Hockenmaier and Steedman, 2005] offers another approach to understanding CCG. I tried to learn how to actually write a CCG, but could not manage to do it and asked for an example on KSL. Unfortunately, no example came my way and, a while later, we realized that CCG was not the right move because it did not have a translation purpose, contrary to synchronous CFG.

So, I learned about this tool but I have no idea how to actually use it and would like to learn more because it seemed full of interesting promises.

3.3.5.4 Synchronous CFG

While waiting to find a way to use CCG, I started studying synchronous CFGs and writing one for KSL - LSF translation.

3.3.5.4.1 What is synchronous CFG? Synchronous CFG is the tool based on CFGs that allows the generation of pairs of related strings instead of single strings (the rules are the same ones as for CFGs but paired up). Many situations that requires the specifcation of a recursive relationship between two languages can be dealt with using synchronous CFG.

Their first application was oriented toward programming languages but their use has since then been extended to natural language processing for machine translation ([Wu, 1997], [Yamada and Knight, 2001], [Chiang, 2005], [Yamada and Knight, 2012]).

An example from [Chiang, 2007] will help illustrate synchronous CFGs:

$$\mathbf{X \rightarrow (yu\ X1\ you\ X2,\ have\ X2\ with\ X1)}$$

The two languages here are Chinese and English. This means that a Chinese sentence of the form "yu [...] you [...]" can be tranlated in English to "have [...] with [...]" where the two subphrases represented by "[...]" appear in opposite order (and are recursively translated as well).

3.3.5.4.2 Why not use it? Unfortunately, I was told that the last paper that used synchronous CFG was dated from 2013, which made it outdated and not fit for future hopes of publication.

3.4 Taking stock

This is all that has been done for now on the translation subject. I would have liked to get to write a neural network, I had studied many of the ones presented in the articles in order to be as ready as possible for this part, unfortunately we did not get this far. We will continue working together on this from a distance once the internship is over but I have some reservations about this work that I will explain in the conclusion.

4. The visual aspect of sign languages

4.1 L'iconicité des langues des signes

This concept was developped by Christian Cuxac[Cuxac, 1993], linguist and professor in *Sciences du Langage* in Paris XIII University. It is heavily used in French literature in the field of sign language linguistics.

Definition: ***Iconicité***

The property for a sign, a signifier or any linguistic form produced to resemble or recall the meaning, the signified, the associated concept.

That is to say that, to the eye of anyone, sign languages contain elements that are visually meaningful.

There are two types of such elements: the visually meaningful *lexical* signs and the visually meaningful grammatical elements.

4.1.1 From sign language vocabulary to universal gestures?

4.1.1.1 What are visually meaningful *lexical* signs?

Lexical signs that *make sense* visually can be better explained with examples: for instance, the sign for [eat] or [drink] in any sign language immediately make sense for anyone who doesn't know anything about sign language.

4.1.1.2 Why are they worth mentioning?

Visually meaningful lexical signs are interesting in terms of global communication and body language.

They logically tend to be similar from one sign language to another.

That is to say, for each word, if we could find commonalities or constants between many sign languages in the way the associated sign is executed, then we could deduce a body language gesture that anyone would immediately understand and memorize without effort. And hopefully, we could do this for a large set of vocabulary and extract a list of body languages gestures for a better communication (with foreigners, with children, with deaf and hard of hearing people, with people with a language impairment and any other situation where oral communication is made more difficult). This is one research direction that seemed particularly interesting to me.

To illustrate this idea, consider the word "rainbow".

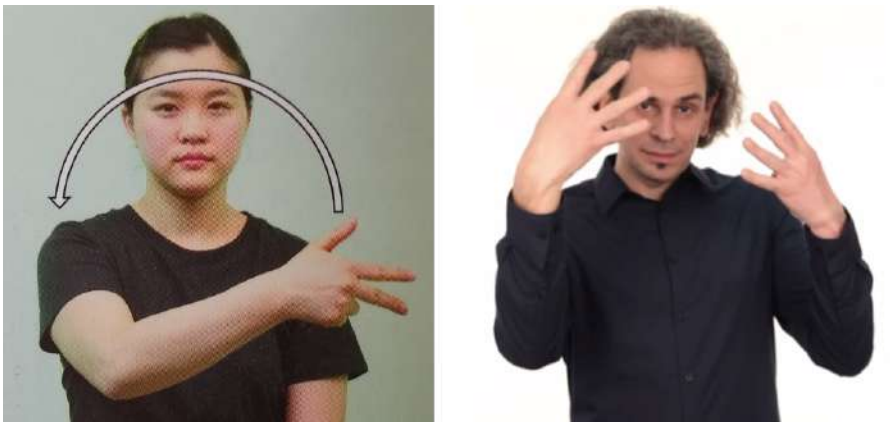


Figure 4.1: On the left, "rainbow" in KSL, picture extracted from KSL beginners book; on the right, "rainbow" in LSF, picture extracted from a video from Elix

As illustrated in Figure 4.1, KSL and LSF signs share similarities such as the movement, the placement and the orientation of the palm. The hand shapes differ, but it is actually explained by the fact that, in this case, the hand shape in LSF is meant to be descriptive (visually meaningful) whereas in KSL it is the number 7 (even though it looks like a 3 to a French person, it is a 7 in the very sophisticated one-handed counting system in KSL) because they say that there are seven colors in a rainbow. If we take more sign languages into consideration (which can be done by researching "arc en ciel" in the Spread the Sign dictionary), it appears that placement and movement are always identical and the hand shapes are often identical to the LSF one. In conclusion of this example, maybe anyone can enrich their body language by effortlessly adding "rainbow" in the way LSF does it and for instance, once they are sitting in a train and notice a beautiful rainbow, they can share it by gesturing through the window to their loved one who is standing on the platform.

In terms of machine translation, the fact that a lexical sign is visually meaningful, does not usually make a difference. After all, it is a sign like any other. However it makes a difference in the case of verbs because sign language dictionaries usually only show the standard verb (like the infinitive form in French dictionaries) but this standard form is almost never actually used because the verb is modified to adapt to the nouns they depend on.

An illustration of this is the verb [eat]. The video in Elix is the standard version. But the sign that is actually executed in sentences will almost look like a mime of actually eating the precise thing in question.

For instance, "I ate an apple" will be signed [apple]+[eat(apple)], where [eat(apple)] has the handshape that one would have holding an apple and looks like the person is taking one bite out of the apple. Meanwhile, if a meal was eaten, the gesture will either look like someone is holding a fork and eating, or show the chapsticks (index and middle fingers) depending on the context. And the possibilities are almost endless.

The difficulty with this is that it enlarges the vocabulary enormously, and no data is actually gathered for this. While working on gloss for the translation, it will not be a real problem, but for a translation system to have any use with sign language, it needs to be able to deal with the visual variations of the verbs, and no currently existing data set is large enough to let the machine learn this on its own.

4.1.1.3 Comparing Sign Languages to find universal gestures

Goal: Deduce an elaborate *universal gestures* list for global communication.
Resources:

1. Spread The Sign: `www.spreadthesign.com`.
   "An international dictionary that aims to make all sign languages of the world accessible. This pedagogic self-learning tool is free to use for everyone in the world."
   Sign Languages: Arabic (Syria), Bulgarian, Chinese, Croatian, Czech, English (x4), Estonian, Finnish, French, German (x2), Greek, Icelandic, International, Italian, Japanese, Latvian, Lithuanian, Polish, Portuguese (x2), Romanian, Russian (x2), Slovak, Spanish (x5), Swedish, Turkish, Ukrainian, Urdu.
2. Elix: `www.elix-lsf.fr`
3. KSL dictionary: `http://sldict.korean.go.kr`

Process:

1. Define a *distance*[Mikolov et al., 2013];
2. Apply video recognition, keep relevant information[Cao et al., 2018];
3. Compute distance between: all words with same meaning from Spread the Sign, KSL vs LSF words from both dictionaries and words of different meaning to compare (but this would be a heavy computation);
4. Define a threshold;
5. Merge signs of the same meaning when appropriate;
6. Survey hearing people on the result.

I worked on the definition of the distance for such a task and found that separating the problem according to each *parameter* (as defined in Annexe) is well suited for this. This leads to the necessity to define a distance for the various possibilities for each parameter. For instance, in the case of placement, distance could be the euclidian distance in space at time t. This implies to extract the information for each parameter and work on them separately. If we consider the handshape for instance, I tried to find a way to extract this information ([Dilsizian et al., 2014]). I studied neural networks that do so. Helpful resources exist, such as this paper [Panteleris et al., 2018] that presents a 3D hand tracking tool used for sign language processing. The idea from this was to apply transfer learning, however once again the lack of annotated data stopped me. Considering another parameter that is facial features, the extraction of this information is a work in progress and is explained later in the report.

Another difficulty is the difference in the execution time for two different signers. We have to align the videos before comparing them and there are two possibilities to do this: either by using the Dynamic Time Warping algorithm [Berndt and Clifford, 1994] or by only extracting the information at specific given times in the video (for instance, beginning state, intermediary state, final state) but the latter would require finding said times, which is no easy task.

4.1.2 What about *non-lexical* signs?

4.1.2.1 What are they?

We are talking here about *non-lexical* gestures that are necessary to the syntax: for instance, the sign for [computer] would not be understood by someone who has never seen it before but in many sentences involving this sign, the signer will add a gesture that consists in mimimg opening a laptop computer and hitting a keyboard because it is grammatically needed. This is not a lexical sign, it only has a grammatical function, like an ending variation to a verb in French, but it is clear for anyone looking that the signer is talking about using a computer.

Sign language users know to use this kind of "visual signing" to make their speech clearer, not only to other experienced signers but even more so to less experienced people, total strangers and foreign signers who use a different vocabulary.

They even push it to the extreme in an art form called *Visual Vernacular* (VV) where stories are told without any lexical item (which makes it understandable to the same degree for any signer around the world).

4.1.2.2 Why are they worth mentioning?

Usually, we work with gloss to write sign languages, but this kind of signs are *extremely difficult to write down in a gloss* because they can be anything and are often improvised on the go (see in the conclusion the part about the fact that sign languages being purely oral make them not suited for machine translation).

The second aspect is the fact that they are simply incompatible with online dictionaries for translation.

Context:

- Challenging for Sign Language Processing (SLP);
- Challenging for gloss writing (usually, the way we write it, gloss is incomplete);
- No comprehensive list yet, establishing one would be interesting both for linguistics and sign language teaching;
- Interesting to study to improve SLP, SLT and especially SL ⇋ SL translation.

Process:

1. Break sign language video into a sequence of signs (LSTM neural networks [Sundermeyer et al., 2012] hold the answer to this[Mocialov et al., 2017]);
2. Compare each sign with dictionary database (this needs Dynamic Time Warping and a way to ignore individual signer variations and grammatical declinations applied to the signs);
3. Keep the ones not found in the dictionary;
4. Apply analysis (not yet defined precisely).

After learning about useful techniques and reading potentially related articles, I abandonned this research direction. It was very informative to think about and led me to discover new aspects of the research spectrum but it seemed impossible because any candidate algorithm would have a disqualifyingly long execution time because of the second step (comparing each video of the sentence to tens of thousands of vocabulary videos and doing this with a lot of sentences to draw conclusions).

Despite the lack of comprehensiveness, human analysis of sign language sentences seems much more productive to find answers on this particular subject than the use of computers.

5. Extraction of facial features

5.1 Motivation

As explained in the introduction, one of the essential parameters in sign language is the facial expression. However, many approaches to sign language processing disregard this parameter. Actually, we focused only on some facial expressions because it was a first step, but the research should be extended to all non-manual features of sign language.

Extracting non-manual features from sign language videos is crucial for many reasons:

- they are essential *grammatical* constituants of the language and thus ignoring them is similar to processing spoken languages but deciding to disregard all vowel sounds;
- extracting them and reproducing them in the target language is part of the action plan for the sign to sign translation system;
- for any analysis of sign language, they bring their own set of information;
- analysing extracted non-manual features in sign language might lead to new discoveries about the range of non verbal expression that the body can produce (which could be helpful to design tools to help people communicate better, whether they interact with people who speak a different language or live with some sort of disability that might make verbal communication more difficult for them, and for people who might have trouble interpreting body language, a better understanding would help them learn more).

5.2 Data set

As for many things in language processing, the first question is that of the data sets that we can use. Machine learning seemed to be the best approach for this. Unfortunately, there was no well annotated data set for this task. I was asked to work on a data set called *Phoenix 2014*, which is made of short German Sign Language videos (in the form of folders of pictures that are the frames). It consists in roughly 10000 weather forecast videos with one single, well-framed signer, separated in three subsets (called test, dev and train, the latter containing almost 8000 videos). All the pictures are in extremely low quality (so low that even by hand I was unable to label most of them -the ones I tried, drawn at random- for eyebrow position because the wrinckles that would give clues ([Zhang and Tjondronegoro, 2011]) did not appear at such a low quality, the eyebrows themselves were sometimes to thin to be visible or hidden by hair, head rotation or a hand, or the signer seemed to have unmovable eyebrows) and each video is very short (less than 200 frames). This data set is labelled for other things but nothing relating to non-manual features, so I was asked to write an algorithm to label them so that we could evaluate the improvement to an already created translation system between German Sign Language and German. And I was asked very specifically to write a conceptually simple algorithm, basically just computing some distances and drawing a conclusion for them, both because it was just a side task and because there was not really any other possibility without labelled data.

5.3 Facial features

First, a brief description of the main facial features that we wanted to recognize is needed. The following list contains the most meaningful information (in regard to sign languages) that is found in the face:

- eyebrows (raised / frowned / resting);
- eyes (open / closed);
- mouth (resting (closed) / o-shaped / stretched (smile for instance) / teeth biting bottom lip);
- cheeks (resting / aspirated / inflated);
- tongue.

Some more precise categories can be found in [NIKL, 2018].

5.4 OpenPose

OpenPose ([Qiao et al., 2017], [Cao et al., 2018]) is a skeleton extraction software. Given a picture (or by extension a video) featuring one or several people, it recovers the positions of many reference points on the body and the face of the people. In particular, the Figure 5.1 shows the reference points on the face.

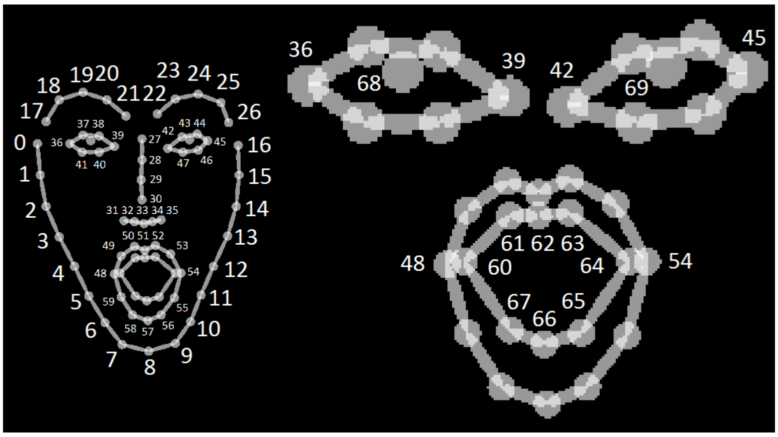


Figure 5.1: The reference keypoints on the face that OpenPose can retrieve.

OpenPose has 2D and 3D extraction features, but given the data set only 2D can be retrieved.

The first step was to apply OpenPose on the whole data set and extract the facial keypoints in order to use them in the facial feature extraction algorithm.

5.5 Eyebrow position extraction

5.5.1 Algorithm

The code to this algoritm can be accessed following this gitlab link, namely in the *eyebrows.py* file.

The main idea of the algorithm is to compute the distance between the eyebrow and the eye for each frame of the video, find the minimum and the maximum of this value throughout the video and conclude for each frame that the eyebrow is frowned if the distance is close enough to the minimum, raised if the distance is close enough to the maximum and resting otherwise.

The algorithm contains the following steps:

- Transform the data from the original OpenPose output to two vectors V_x and V_y such that keypoint k is $K(V_x[k], V_y[k])$. The whole process (from reading the file to extracting V_x and V_y) can be done using $V_x, V_y = \text{norm}(\text{read\_file}(f))$.

- In order to deal with potential issues with the head being turned to the right or the left, after some tests, it appeared that it made more sense to only keep the data from the eye of the side that is turned toward the camera. This eye is selected by comparing the distance $36 - 39$ and the distance $42 - 45$ and keeping the larger one. Then we compute the middle of these two points (because it gives us an invariant center for the eye, contrary to the pupil, we cannot move the corner of our eyes) and compute its distance to keypoint 19 (resp. 24). Since these 2 points are supposed to be more or less vertically aligned (this assumption might be an issue for instances where the head is turned in such a way that they are not vertically aligned at all), the possible rotation of the head to the right or left does not impact this distance too much and they are in theory the 2 most relevant points when trying to figure out whether the eyebrows are raised, frowned or resting.

- Assuming one folder contains the frames from one video only, we compute said distance for each frame. Everytime, the value is inserted in a sorted array of distances and we keep track of the order.

- We compute the threshold values using the minimum and maximum distances, the value in between ($m = \frac{\text{min+max}}{2}$) and a coefficient (that can be changed but it seems that its optimal value would be around 0.65). The two threshold values are $T_1 = \frac{\text{coef}\times\text{min}+m}{1+\text{coef}}$ and $T_2 = \frac{\text{coef}\times\text{max}+m}{1+\text{coef}}$. To conclude, for each computed distance d, if $d < T_1$, then the eyebrows are frowned, if $d > T_2$, the eyebrows are raised and otherwise, they are resting.

5.5.2 Accuracy

The script *sample.py* lets the user enter the number of frames and outputs 15 random numbers. It was used to draw frames at random from the folders once they were chosen. Selected frames were then labelled as accurately as possible by hand (the difficulty as mentioned being that the label was not always easily known, even for the human eye) in JSON files. Then the algorithm was used to label them in a second JSON file.

The script *accuracy.py* compares the 2 results on the 4 chosen folders (folders were picked at random but the selection was repeated until one that was easier to label for the human eye was found, and many of them did not seem to work). It outputs percentages (one for each folder individually and then the global one).

Please note that I was reprimanded for my pretentious use of the words "at random" which were not totally provable. I understand the issue with this, however it was only supposed to be an intermediary testing in order to decide whether this approach was good enough or not. Besides, the accuracy testing can be replicated (or improved), thus I do not believe that my method was wrongful as an intermediary step.

5.5.3 Results

Four folders from the data set were tested, the results are the following:

On 01December_2011_Thursday_tagesschau-3473:
The 15 chosen frames are [116, 134, 6, 130, 2, 22, 183, 177, 12, 72, 81, 179, 128, 51, 13].
The result is 14 correct out of 15, ie **93.333...%** accuracy.

On 15March_2011_Tuesday_tagesschau-3326:
The 15 chosen frames are [28, 5, 6, 32, 37, 69, 38, 13, 23, 26, 63, 60, 21, 65, 18].
The result is 10 correct out of 15, ie **66.666...%** accuracy.

On 27November_2011_Sunday_tagesschau-5149:

The 15 chosen frames are [12, 86, 64, 74, 10, 67, 82, 99, 21, 105, 25, 54, 113, 43, 79].
The result is 11 correct out of 15, ie **73.333...%** accuracy.

On 03March_2011_Thursday_tagesschau-7064:
The 15 chosen frames are [116, 134, 6, 130, 2, 22, 183, 177, 12, 72, 81, 179, 128, 51, 13].
The result is 12 correct out of 15, ie **80%** accuracy.

In total, this is 47 correct out of 60, meaning 78.333...% accuracy.

A few words on the results:

- Difficulties with the dataset: as mentioned, even by hand it is very difficult to annotate some (many) frames from the dataset. Indeed the pictures are very small (so we cannot see in most cases the wrinkles that would help one determine the facial expression, which makes it very difficult to validate or invalidate the result), usually not very precise, some signers have hair hiding their forehead (or punctually hands hinding it) or their eyebrows are not obviously visible. All of this is a problem for the human trying to annotate the frames as well as for the openpose software, which, from what I have seen of the results, is approximately correct in most cases but not precisely correct, and small variations might change the results of this algorithm (because the variations between frowned / resting / raised could be the same size as the mistakes of the openpose software on these pictures). To check this, I have tried the algorithm on much better quality videos of my own and I found that the results seemed better.
- As a whole, I would say that the results of the algorithm are much lower than the ones shown earlier, because I voluntarily excluded folders that I was not able to manually annotate (not seeing the face well enough, or signer who did not seem to move their eyebrows at all), but when running the algorithm, it is even more confused than I am.
- As to what could induce errors in the algorithm:
  - 3D to 2D: the distances would make more sense in 3D. On 2D pictures, they are distorted. I tried to choose the keypoints whose results would be the least affected by the 2D-ism of the picture. However, maybe introducing some other considerations might allow us to get better results.
  - Rotation of the head:
    * Left to Right: if this is the only rotation, it should not be a problem because the points are vertically aligned;
    * Other directions: the head can move in any direction, and apparently it does so most of the time. Might be interesting to try and get information about this from the keypoints but this is really hard because there aren't so many points
  - Obstacles for openpose or simply bad results of openpose: there is not much to do about that, except for taking the confidence score from openpose into account;
  - Lack of eyebrow movement from the signer throughout the whole video: usually this should not happen, but it seems to happen a lot in the dataset. In this case, since the distance only shows slight variations, and the answer should be constant, the algorithm will still propose three results on the various distances. One solution would be to define a minimum variation range to consider (depending on the size of the head on the picture of course) before supposing it is changing. But there are several issues: personal variations (in the proportions of the head as well as the range of movement of the eyebrows), still not knowing if their is only one position which one it is and the case where there are only 2 positions (for instance the eyebrows are never raised) or actually 3 with limited range);
- This algorithm does not make any use of the temporal evolution between the frames. Instead, each frame is studied almost independently, except for the fact that the threshold is computed using the minimum, maximum and middle values from all the frames, but the ordering is completely forgotten and it should not be because the evolution holds crucial information.
- Possible room for improvement: the statistical tools. I tried to use the mean, the median and settled on the middle value between the min and the max. However, smarter use of statistical tools for the choice of the thresholds might help improve the results.

I reported on these results and was told that I had not spent enough time looking into previous work for this specific task, which explained the problems I encountered. This was the only time that I tried to do some work without researching previous work before (because I was told to do so), and even though they did not appreciate it, it felt enlightening to me, because I discovered the real issues behind the subject by doing so.

5.6 OpenFace

The mistake in direction showed that OpenPose is outdated for the purpose of extracting facial features from 2D videos. Indeed, OpenPose in this context only outputs 2D keypoints, which means that an approach based on distances was impeded by any movement of the head. Besides, even though we do not have an annotated dataset to train a neural network to recognize facial features, this kind of work has already been conducted and we could benefit from the results instead of starting from scratch.

This is where OpenFace ([Baltrusaitis et al., 2018], [Baltrušaitis et al., 2016], [Amos et al., 2016]) comes in.

What is particularly interesting with OpenFace is that they already do some facial feature extraction with what they call Action Units (see Figure 5.2). They not only detect the presence of these Action Units but also the intensity at which they appear in the face.

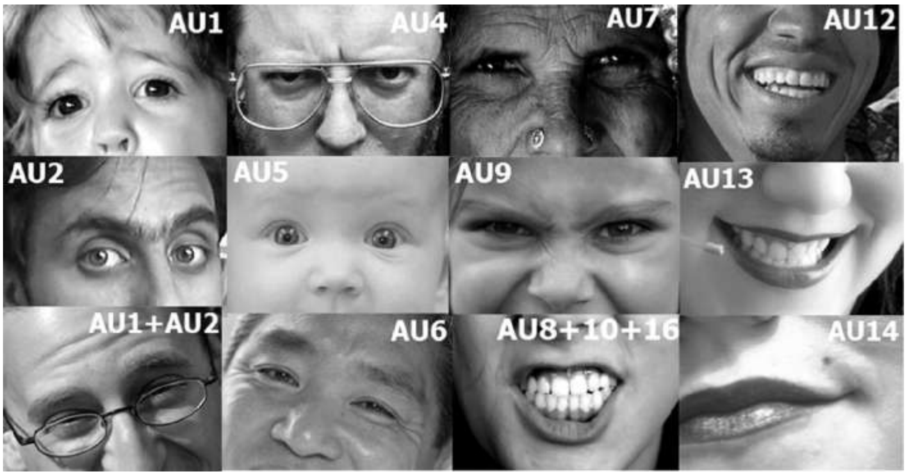


Figure 5.2: Example of facial expressions recognized by OpenFace

Once again, I wrote a script to apply OpenFace to the entire data set. In the results, the CSV file created for each video was the interesting part and in particular, the Action Units (AUs) mentions. In order to later add this as annotation for the already existing neural network (for German Sign Language to German translation), I wrote a python script extracting the wanted data, which can be found in the file called *ann.py* in the same gitlab link.

6. Conclusion

This internship was an amazing opening for me to get familiar with the topic of sign language processing. I did not get results but I learned and discovered a lot of things, gained a method and an overall view of the subject (which is a very effective tool to have to make more progress in the initiated research and to start in new research directions), experienced having to be very autonomous and adapt to the practicalities (waiting for data sets that never come, being stopped in a direction for lack of an authorization and having to improvize a new one). I learned about myself as well, how reading articles can sparkle enthusiasm to learn about newly formed questions, how witnessing promise unfulfilled can trigger frustration, which in turn ignites motivation. And the adventure goes on because they want me to keep working with them and I will be very happy to do that, even more so because it will be pressure-free.

When I arrived, they kept telling me that 3 months (and I stayed more than that) would be way too short given the subject. And I understood that "too short" was because I would not be able to publish anything in this time, which was their goal. Reaching the end of my internship, I agree that 3 months was too short for the very broad subject of "sign language processing": in this time, I have merely been able to grasp the stakes. Unfortunately, I think that the "publishing" goal that drives research as I have seen it in this internship is counter-productive in the very specific field of sign language processing because it pushes researchers to bend the research subjects to be able to use the data that they have. It happened to us during my internship, and it seemingly happend to many researchers from what I can see in their publications.

Besides, researchers usually try to learn about the languages they are working on and they do understand a lot about them but I have yet to meet someone working on sign language processing who uses the language and interacts with the community on a daily basis and I believe that it makes a real difference (which is why I started a Master's degree in sign language interpreting in the first place). Being a user of the studied language assures that the researchers keeps the big picture in mind and remembers the real motivation behind sign language processing. I think it also helps with targeting goals in a different way: for the research team at the lab, the only research direction that seemed to make real sense was KSL - LSF translation. We can justify this (and we did) by saying that the deaf community would benefit from it but, actually, deaf people are way more comfortable learning foreign sign languages than hearing people are (it takes signers about 3 days to be comfortable communicating in a new sign language). What's more, deaf people are used to fighting hard to be able to communicate and they developed skills for it, so global communication is not so much of a problem for them. It might be more pressing to make French content accessible in LSF than to make KSL content accessible in LSF (especially considering that KSL content that would follow the conditions to be translated by our translation tool basically does not exist). But because we needed to be able to make relatively quick progress and KSL - LSF translation still makes sense even though it should not be the priority, that is the main direction that we chose. Also a problem in this was the approach that we used, disregarding the nature of sign language, we wrote gloss of sign languages instead of using videos. This allows us to focus on one precise part of the problem, it is true, but everytime I had to write gloss, it felt really wrong because I knew, no matter how much we tried to include specificities into the gloss, that I had to cut out part of what made the essence of sign language. Researchers know about the language but they do not know the language, I suppose that is why they can't feel that gloss is *not* sign language, it merely captures elements of it. I did not know KSL when I arrived so I did not have such strong opinions about this language, but trying to learn it and analyzing the gloss that we had, it felt just as wrong as the LSF gloss. I tried to keep in mind that this was probably due to lack of cultural knowledge on my part but once I had learned enough KSL to discuss the material with the KSL teacher, he confirmed my impression: what we were using as a baseline for translation (and what they were using to teach KSL) was actually not KSL, it was a modified and truncated version created to make hearing Korean-speaking learners more comfortable. I actually started by asking "Could we say it like that?" and producing LSF syntax and grammar with KSL vocabulary, basically applying my first hypothesis for the translation, and he would answer "that's the way deaf people would say it, yes, but I have to follow the guidelines to teach hearing students here", which is what makes me think that the first hypothesis is really worth exploring.

This internship convinced me that, for the greater good, researchers in sign language processing need to take a step back, stop running forward into potential dead ends, take a look at the big picture and rebuild the basis where they skipped it.

The first obvious thing is that sign languages have no written form for now (there are tentative writings but none of them is used so it does not count). They can't keep treating they like spoken language that have a writing: orality is a very special case (and very chaotic) and that's all there is with sign language at the moment.

So the number one priority for now would be to actually create a writing for sign languages (fortunately, I started this project this year with a team of classmates and now that I have learned about the many tools and ways to deal with sign language, I am better equipped to make progress in the project), let it be used by the community and in time a standard written form will appear and be better suited for many sign language processing projects, and the language as a whole while know a new growth from this.

Not far behind is the necessity to make, gather and annotate data, but first actually prepare and make a list of possible uses for the data so that it can be annotated as comprehensively as possible (I tried to make the best of the existing data as explained in Annexe B). Once again this is a step that takes years to be fulfilled to the extent that machine learning requires.

Bibliography

[Amos et al., 2016] Amos, B., Ludwiczuk, B., Satyanarayanan, M., et al. (2016). Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6.

[Baltrušaitis et al., 2016] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

[Blackburn, 2005] Blackburn, P. (2005). Representation and inference for natural language: A first course in computational semantics.

[Book and Fisher,] Book, D. and Fisher, G. Signnet: A neural network asl translator.

[Camgoz et al., 2017] Camgoz, N. C., Hadfield, S., Koller, O., and Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084. IEEE.

[Cao et al., 2018] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*.

[Chiang, 2005] Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

[Chiang, 2007] Chiang, D. (2007). Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228.

[Chomsky, 1956] Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

[Chomsky and Lightfoot, 2002] Chomsky, N. and Lightfoot, D. W. (2002). *Syntactic structures*. Walter de Gruyter.

[Cui et al., 2017] Cui, R., Liu, H., and Zhang, C. (2017). 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.

[Cuxac, 1993] Cuxac, C. (1993). Iconicité des langues des signes. *Faits de langues*, 1(1):47–56.

[Cuxac and Pizzuto, 2010] Cuxac, C. and Pizzuto, E. A. (2010). Emergence, norme et variation dans les langues des signes: vers une redéfinition notionnelle. *Langage et société*, (1):37–53.

[Dilsizian et al., 2014] Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. N. (2014). A new framework for sign language recognition based on 3d handshape identification and linguistic modeling. In *LREC*, pages 1924–1929.

[Filhol and Falquet, 2017] Filhol, M. and Falquet, G. (2017). Synthesising sign language from semantics, approaching" from the target and back". *arXiv preprint arXiv:1707.08041*.

[Filhol et al., 2014] Filhol, M., Hadjadj, M. N., and Choisier, A. (2014). Non-manual features: the right to indifference. In *6th Workshop on the Representation and Processing of Sign Language (LREC)*.

[Gianni et al., 2019] Gianni, F., Collet, C., and Lefebvre, F. (2019). Modèles et méthodes de traitement d'images pour l'analyse de la langue des signes.

[Hadjadj, 2017] Hadjadj, M. (2017). *Modélisation de la Langue des Signes Française: Proposition d'un système à compositionalité sémantique.* PhD thesis.

[Hockenmaier and Steedman, 2005] Hockenmaier, J. and Steedman, M. (2005). Ccgbank: User's manual.

[Hockenmaier and Steedman, 2007] Hockenmaier, J. and Steedman, M. (2007). Ccgbank: a corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396.

[Kacorri and Huenerfauth, 2016] Kacorri, H. and Huenerfauth, M. (2016). Continuous profile models in asl syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2084–2093.

[Kacorri et al., 2016] Kacorri, H., Syed, A. R., Huenerfauth, M., and Neidle, C. (2016). Centroid-based exemplar selection of asl non-manual expressions using multidimensional dynamic time warping and mpeg4 features.

[Lederer and Seleskovitch, 2014] Lederer, M. and Seleskovitch, S. (2014). *Interpréter pour traduire.* Les Belles Lettres.

[Michael Erard, 2017] Michael Erard (2017). Why sign-language gloves don't help deaf people. [Online; last accessed June 21, 2019].

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

[Mocialov et al.,] Mocialov, B., Hastie, H., and Turner, G. Towards automated sign language to written language translation.

[Mocialov et al., 2017] Mocialov, B., Turner, G., Lohan, K., and Hastie, H. (2017). Towards continuous sign language recognition with deep learning. In *Proc. of the Workshop on the Creating Meaning With Robot Assistants: The Gap Left by Smart Devices.*

[NIKL, 2018] NIKL (2018). Research and building of the ksl corpus. *Seoul, National Institute of Korean Language.*

[Panteleris et al., 2018] Panteleris, P., Oikonomidis, I., and Argyros, A. (2018). 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.

[Qiao et al., 2017] Qiao, S., Wang, Y., and Li, J. (2017). Real-time human gesture grading based on openpose. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE.

[Sundermeyer et al., 2012] Sundermeyer, M., Schlüter, R., and Ney, H. (2012). Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association.*

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Séro-Guillaume, 2008] Séro-Guillaume, P. (2008). *Langue des signes, surdité & accès au langage.*

[Tarnowski et al., 2017] Tarnowski, P., Kołodziej, M., Majkowski, A., and Rak, R. J. (2017). Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184.

[Wu, 1997] Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.

[Yamada and Knight, 2001] Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 523–530.

[Yamada and Knight, 2012] Yamada, K. and Knight, K. (2012). Syntax-based statistical translation model. US Patent 8,214,196.

[Zhang and Tjondronegoro, 2011] Zhang, L. and Tjondronegoro, D. (2011). Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, 2(4):219–229.

Annexes

Annexe A: Explanation of sign language notions for context

As mentioned above, sign languages fundamentally differ from spoken languages. Instead of using successive sounds to produce meaning, sign languages rely on simultaneous visual information (then combined in time just like sounds).

1 Decomposition of signs into parameters

One widely used way to classify the visual information provided in signs is with what linguists call *parameters*. Parameters were primarily theorized by a French linguist called Christian Cuxac ([Cuxac, 1993]; [Cuxac and Pizzuto, 2010]) and have been used consistently in every field (namely both linguistics and computer science) since then. There are traditionally five simultaneous parameters that evolve in time to build a sign language expression:

- the shape of each hand;
- the orientation of each hand (the direction of the palm);
- the placement of the hands;
- the movement;
- and, traditionally, the facial expression, which, in turn, contains many other parameters (shoulder and head movement, body inclination and rotation, eyebrow, cheeks, mouth movement, eye direction and shape, emotion).

2 The structure of a sign language sentence

Everyone knows that spoken languages sentences are build with several elements:

- the word order is called *syntax*;
- the modifications that affect words to make them fit in the sentence and provide additional meaning are called *grammar*;
- bits of additional information lie in the *lexical* choices, the *vocabulary* being there to let us know what characters, objects and actions are involved.

These three elements (syntax, grammar, vocabulary) are still present in sign language but it is crucial for the rest of this report that the reader gets an idea of what they look like in sign language sentences.

- The *syntax* is different and not as fixed but there is still a prefered way to order the signs;
- The *grammar* lies in some facial expressions (frowned or raised eyebrows indicating a respectively open or closed question replace the inflection in spoken form and interrogation mark in written form as well as phrasings such as "est-ce que" in French), noding or shaking of the head (to either affirm or negate the statement), sign alterations (see Figure 6.1) and the insertion of gestures that are not part of the vocabulary but improvised and visually meaningful;
- the *vocabulary* is the set of standardized signs that the signer has at their disposal, dictionaries exist (such as Elix for LSF), but signers can also draw from the local spoken language, either by spelling the word with the local manual alphabet or articulating the word. If it is an object, they may also represent its shape in the air following the sign language rules for this technique or even impersonate the object or the character they are talking about, and in these cases (description or impersonation) it falls back into the *grammar* category.

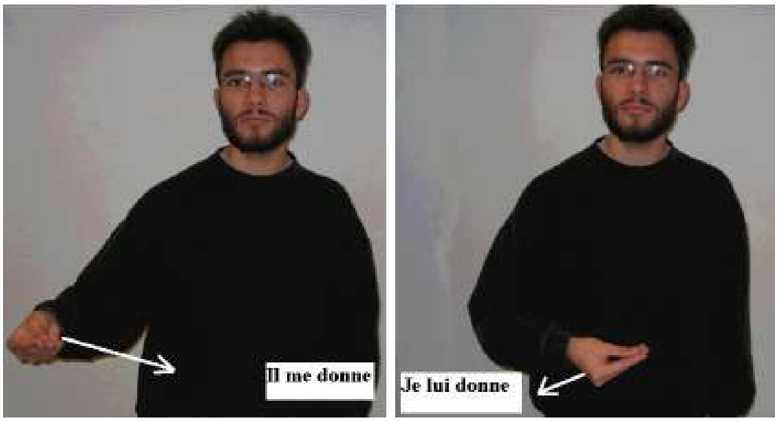


Figure 6.1: An example of a verb changing its spacial execution for grammar purpose, the verb in question is "to give", on the left "he gives me", on the right "I give him", both assuming "he" has been abstractly placed on the right of the signer previously, extracted from [Gianni et al., 2019]

Annexe B: Additionnal work (data sets)

1 Crawling the web to extract data sets

Because we needed data sets and some were on the internet, I learned how to retrieve data from the internet and wrote tools in advance that I would be able to adapt quickly to any need once I obtained the authorization to use this data.

I learned how to use scrapy and prepared tools for the two main websites containing relevant corpora: Elix, the LSF dictionary and Media Pi the magazine in LSF. The spiders can be found in the following gitlab depository.

In the *mediapi* folder, the main spider is *pi.py*, it crawls through the entire website, retrieves the links to all the videos and stores them in a CSV file along with some pieces of information: the category, the title, the date, a description, the associated tags, the link to the video and the link to the signed title of the video. The other spiders execute only subtasks of this one (the files work independently). All the spiders for the Media Pi website require a user name and a password, because the website requires authentification. I was able to test them because I had an account, but cannot share this confidential information. Any research institution could easily use the spider to retrieve the data if access was granted to them.

As for the Elix website, the easy part was the absence of user authentification, but the structure of the website made it more complicated to use. The website contains what is called a "videotheque" that contains some videos but not all of them, the crawler for this part of the website is in the file called *elix-videotheque-scraper.py*. Alongside, there is a "research" function and any word we ask for will lead to a list of words, each of them possibly accompanied by a definition in French, a link to definition in LSF and a link to one or several videos of the sign. The file *elix-definition-scraper.py* lets the user chose one word and returns the definition videos of the research results. The file *elix-sign-scraper.py* lets the user chose one word and returns the sign videos of the research results. The file *elix-research-scraper.py* lets the user chose one word and returns both the definition and sign videos of the research results. The folder *word-lists* contains lists of words and a spider that can research the video results for all the words in the list. The folder *elix-country-scraper* works on the list of every country in the world and the associated capital city. It extracts the signs of the countries when they exist and some other information, actually downloads all the country-name videos from the website and creates a pdf file containing all this information (including the videos). The purpose of this one was to learn to extract the information all the way to the end of the process.

2 Creating a data set

In parallel, I also started working on the creation of a LSF data set of my own, because I figured that, if I owned it, at least there would be no legal problem in using it, and someone needs to start somewhere at some point if we hope to get more data sets someday. However, this is a very long and tedious task and it is far from being enough data even after more than three months of daily additions.

Annexe C

The following pictures illustrate the OpenPose output:

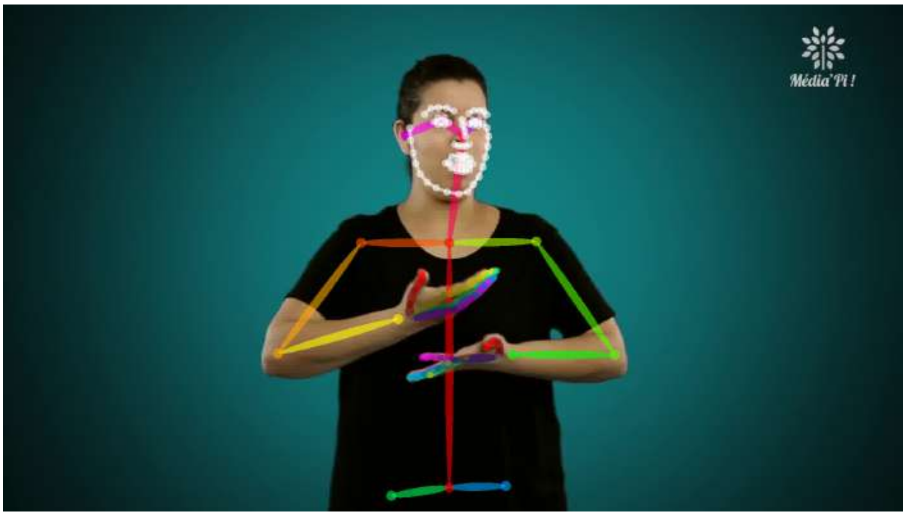
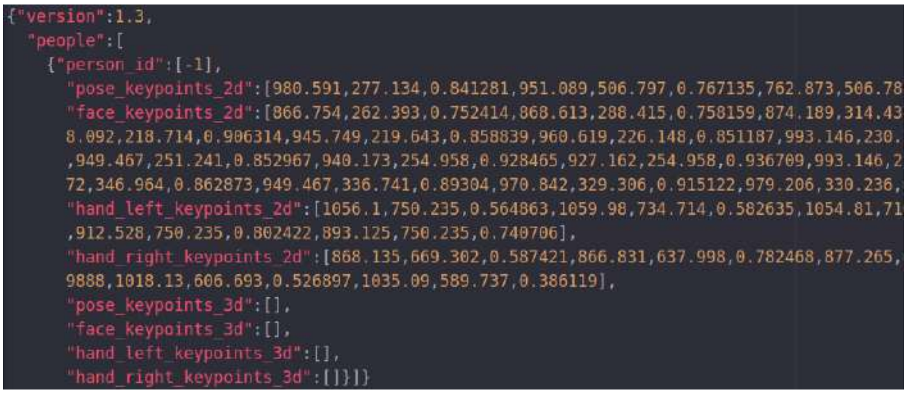


Figure 6.2: Picture extracted from LSF video found on the Media Pi website



```
{"version":1.3,
  "people":[
    {"person_id":[-1],
      "pose_keypoints_2d":[980.591,277.134,0.841281,951.089,506.797,0.767135,762.873,506.78
      "face_keypoints_2d":[866.754,262.393,0.752414,868.613,288.415,0.758159,874.189,314.43
      8.092,218.714,0.906314,945.749,219.643,0.858839,960.619,226.148,0.851187,993.146,230.
      ,949.467,251.241,0.852967,940.173,254.958,0.928465,927.162,254.958,0.936709,993.146,2
      72,346.964,0.862873,949.467,336.741,0.89304,970.842,329.306,0.915122,979.206,330.236,
      "hand_left_keypoints_2d":[1056.1,750.235,0.564863,1059.98,734.714,0.582635,1054.81,71
      ,912.528,750.235,0.802422,893.125,750.235,0.740706],
      "hand_right_keypoints_2d":[868.135,669.302,0.587421,866.831,637.998,0.782468,877.265,
      9888,1018.13,606.693,0.526897,1035.09,589.737,0.386119],
      "pose_keypoints_3d":[],
      "face_keypoints_3d":[],
      "hand_left_keypoints_3d":[],
      "hand_right_keypoints_3d":[]}]}
```

Figure 6.3: Extract of the JSON formatting of the keypoints.