Alignment Data Base for a Sign Language Concordancer

Marion Kaczmarek, Michael Filhol

Alignment Data base for a Sign Language Concordancer

Marion Kaczmarek, Michael Filhol
Université Paris-Saclay, CNRS, LIMSI
Orsay, France
kaczmarek@limsi.fr, michael.filhol@limsi.fr

Abstract
This article deals with elaborating a data base of alignments of parallel French-LSF segments. This data base is meant to be searched using a concordancer which we are also designing. We wish to equip Sign Language translators with tools similar to those used in text-to-text translation. To do so, we need language resources to feed them. Already existing Sign Language corpora can be found, but do not match our needs: working around a Sign Language concordancer, the corpus must be a parallel one and provide various examples of vocabulary and grammatical construction. We started with a parallel corpus of 40 short news and 120 SL videos , which we aligned manually by segments of various length. We described the methodology we used, how we define our segments and alignments. The last part concerns how we hope to allow the data base to keep growing in a near future.

Keywords: Sign Language, Concordancer, Computer-assisted translation

1. Introduction

Sign Languages (SL) convey meaning with hands and body gestures, facial expressions, and eye gaze. Each of these is a part of the language grammar. SLs are languages in their own right, even though they still are considered minority languages with very low visibility, if not none. In 2008, the CRPD (Convention on the Rights of Persons with Disabilities) adopted by the United Nations emphasised the right of people with disabilities to fully access any type of information or communication. This has led to an increasing need for translated content in SL, whether for display in public spaces, public transportation, healthcare or broadcasting.

Sign Language interpretation and Sign Language translation are not to be confused. The classical way to distinguish the two is to say interpretation concerns oral languages while translation concerns written texts. Concerning SL, as they do not have an editable written form, the major difference is the possibility to post-edit the translation. Translators can rework their translations many times, until they reach what they believe to be the best account of the meaning in the target language. Interpreters however can only provide one version of their translation and once it is delivered, it cannot be edited (Filhol & Tannier, 2014).

The number of professional SL translators in France is very low despite the recent creation of a diploma. And the few working SL translators are not equipped with Computer Assisted Translation (CAT) software tools like those used by text-to-text translators, because none of them support SL. CAT software eases the translator's tasks (Koehn, 2009), by providing him with an integrated working environment along with dedicated tools (translation memories, concordancers, terminology management tools for example).

We are currently working on designing CAT software dedicated to SL. This paper presents the steps for the creation of a SL concordancer, which is an essential module of all text-to-text CAT software.

2. Sign Language concordancer

To learn more about the SL translation process, we conducted studies with professional SL translators (Kaczmarek & Filhol, 2019). We organised a brainstorming session and invited them to express their insights about their everyday practices and the obstacles they thought to often encounter. This session was followed by a free discussion between the participants, where they could discuss previously mentioned ideas and potential solutions to their needs. The scarcity of SL

resources was often mentioned, as well as the need for a tool to enable reuse of prior work.

To complement those results with more objective observations, we also filmed translators at work. We set them up to work in pairs, which led them to discuss their current task. It allowed us to clearly identify their needs and the problems they encounter while working. The analysis showed that the translators could use the help of software tools such as NLP features, or automated encyclopedic searches. However, they have access to very few resources in SL, most of them not being easily searchable (content not referenced properly, dictionaries with limited queriable vocabulary), and also timeconsuming to consult (browsing through hour-long videos of SL content, searching for pictures to find ways to describe concepts with no standard signs).

Both subjectively and objectively, it appears that a bilingual concordancer with French and LSF would be a useful tool to the profession. But Sign Languages not having an editable written form, video is the most common way for signers to store messages. This simple statement already raises problems concerning the adaptation of CAT tools for SL, as videos are neither editable nor queriable, and also imply memory-consuming storage.

A turning point in text-to-text translation appearing with CAT software, is the translation memory (TM) feature (O'Hagan, 2009). It stores pairs of source and target segments, saved from the translators' work. Each of these pairs creates an *alignment*. If either segment is encountered later, or any text close enough to justify a match, the TM will suggest the counterpart segment in the stored pair, and the translator is free to accept, modify or reject the suggestion. The corrected segment will in turn create a new alignment added to the TM, which increases its accuracy. To this date, it is unavailable for SL. We wish to investigate what equivalent could be tested for SL.

Our suggestion here is to design a SL concordancer to keep the benefits from the TM. A concordancer is, regardless of the languages, a software tool allowing to search through corpora and to list the occurrences of a queried word. When it comes to translation, bilingual concordancers are used. The query is done in one of the languages of the pair, and results are provided with the aligned equivalent tagged in the other. Such tool allows translators to look up words or expressions *in context.* For example, given a list of matches for a query, similarity of the results shows the consistency of a translation in different contexts. On the contrary, a variety of results highlights the impact of the context on the translation.

Using a French–LSF parallel corpus, and since LSF cannot be queried easily in our case, we would query the written French part with text input, and provide examples in context of prior text–LSF translations. The alignments are pairs made of an identified segment of the text, and its translation in the corresponding video identified with time tags. Such alignments are stored as a data base which is built by the users themselves. Multiple segments may therefore overlap. When a query pulls up a match contained in more than one segment, the concordancer answers with the smallest in length (tightest annotation around the match), and the aligned video segment as always. If it has not been aligned, but still previously translated, the concordancer can answer by displaying the entire videos corresponding to the texts in which the query can be found, since the parallel content can be viewed as an alignment in itself, although often too big to be directly useful.

To allow the development and testing of a first prototype, we manually built a first data base of alignments. The next part presents the methodology used.

3. Methodology

3.1 Data

Bilingual concordancers are based on parallel translated content, so to elaborate our data base, we need a corpus of translations. Written French and French Sign Language are our working languages

for this study.

Concerning French Sign Language (LSF), a few corpora can be found. The Dicta-sign corpus (European project, 2010) contains on various types of discourse, but mostly dialogues and spontaneous discourses. Even if it is annotated, the annotations provided are of linguistic nature, and made of glosses which can hardly be considered translation in any case. It contains four different SLs. For the French part, larger chunks were also translated and aligned. An LSF interpreter watched the corpus, recorded himself interpreting the signs to French, then subtitled the video by typing his own interpretation. This is not quality parallel data because it has been made in one-shot. As explained in the introduction, this is more an interpretation than a translation.

We used the only usable French–LSF parallel corpus of forty short news texts, of three to five lines each in a journalistic style filmed using two cameras for a front and a side view for a total of 120 videos ("40-brèves" corpus, which you can find on the online OrtoLang platform[1]).
Here is an example of a short text contained in French:

Quelque 200 personnes pourraient avoir été tuées mercredi matin par un glissement de terrain provoqué par des pluies abondantes dans l'île indonésienne de Java, a annoncé à une source policière.

In English: *Around 200 people may have been killed Wednesday morning by a landslide caused by heavy rains on the Indonesian island of Java, a police source said.*

In this corpus, each text is translated by three different professional SL translators, and the resulting translations of an average 30-second duration. Each video already provides us with an alignment (2-3 lines texts can be aligned with 30 seconds videos).

Figure 1 : A screen shot of the video set-up

3.2 Alignments

The goal here was to build smaller alignments than the total ones. In other words, we had to segment the texts into smaller segments and paired them with their translation, in variety of sizes. We had to sort our text segments and only keep those which were suitable to align in such way that the content of both the text and the video segment is the same in order to produce parallel data. We cannot for example, align an adverb in the text with only the facial expression of the signer because the video segment would include other, e.g. manual, features which would not be covered by the text segments. By definition, an alignment implies that both segments are equivalent in meaning (translations of one another). This can be a simple lexical match (sign with no added feature aligned

1 https://www.ortolang.fr/market/corpora/40-breves

with literal word equivalent) or longer clauses on either side.

The alignments had to be made manually because there is currently no way to do it automatically, unlike text-only alignments. To process our data, we chose to use Brat for the texts and Elan for the videos. The steps are detailed below.

3.2.1 Annotation of text segments

The aim was to create an important variety of segments to align them and populate the data base. As we had three different signers, we were free to segment the texts in three different ways, and vary the spans for a given expression. The segments were chosen based on their lack of standard signs, or on variety of translations proposed by the signers: figures of speech, idiomatic expressions or grammatical phenomena. But we also included random segments provided we could align them with a counterpart segment.

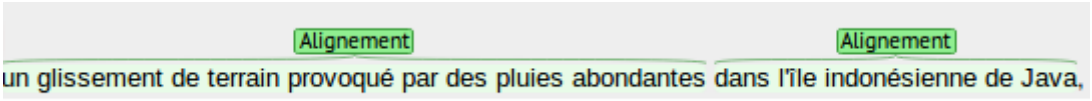

Figure 2 : The source text processed in Brat

We used Brat for this task, because it creates for each selected segment a unique key, which upon export, is characterised by the indexes of its first and its last characters in the source text (considering plain text, with no formatting character). Those will be used in the processing of the videos below.

```
T1      Alignement 15 63        Quelque 200 personnes pourraient avoir été tuées
T2      Alignement 64 72        mercredi
T3      Alignement 83 142       un glissement de terrain provoqué par des pluies abondantes
T4      Alignement 148 153      l'île
T5      Alignement 170 174      Java
T6      Alignement 176 208      a annoncé à une source policière
T7      Alignement 121 142      des pluies abondantes
```

Figure 3 : Example of exported data from Brat

3.2.2 Annotation of video segments

Elan is an annotation software which allows to create tracks and tags to annotate a video. We only needed one track to report the keys. For each text segment, the first step was to watch the corresponding videos to find their translations. Once found, we create an annotation on our track, covering the duration of the translation and label it with the segment's unique key taken from our Brat data. The exported data from Elan therefore consisted in lines displaying for each unique identification key the corresponding time tags on the video, as shown on the figure below.

```
0.05    5.11    T5
5.14    7.56    T6
7.66    9.69    T3
9.74    12.58   T1
12.61   14.93   T2
15.04   20.0    T4
```

Figure 4: Data exported from Elan

Now that we have identified both the text segment and its LSF translation in the corresponding video, we bring them together to create the alignment itself. Each alignment is stored in the database in the following format:

<TxtID, start pos., length, VidID, start time, duration>

Where:

- TxtID is the identification code of the text, ranging from 1A to 1T and from 2A to 2T. This code allows to retrieve them in their own storage space.
- Start pos. is the position of the first character from the text segment, in the text.
- Length is the number of characters of the segment.
- VidID is the unique identification of the video in their own storage space, and allow to retrieve them.
- Start time is the time tag corresponding to the beginning of the segment in the video.
- Duration is the duration in seconds of the segment's translation.

1T	87	101	1T-JP	21.450000	9.660000
1T	189	69	1T-JP	31.200000	10.800000
1A	0	49	1A-OC	14.290000	3.710000
1A	49	9	1A-OC	9.250000	1.040000
1A	68	60	1A-OC	1.530000	1.460000

Figure 5: Alignments as stored in our data base

3.2.4 Metadata files

Apart from the alignment file itself, we keep more information about the texts and videos used in metadata files. The aim is to keep trace of complementary information which could be useful to display in the concordancer, or apply filters to its results. The metadata contains information such as titles, authors or sources and topics.

4. Implementation and open tests

Our current database is made of the 120 alignments from the 40-brèves corpus, plus 343 manual alignments of finer grain in the same texts and videos, i.e. a total of 463 alignments. As we later explain, we will keep making this number grow.

We implemented the concordancer to test this data base, following the design details, to be published elsewhere (Kaczmarek & Filhol, 2020). A screenshot of a result page, which appears after a text query, is given in figure 6. The query is highlighted in yellow in the text and the user is presented with its context. Similarly, the video player shows the matched segment in yellow on the time bar, the user being free to adjust right and left context duration (buttons A and B in the figure), or even watch the entire video.

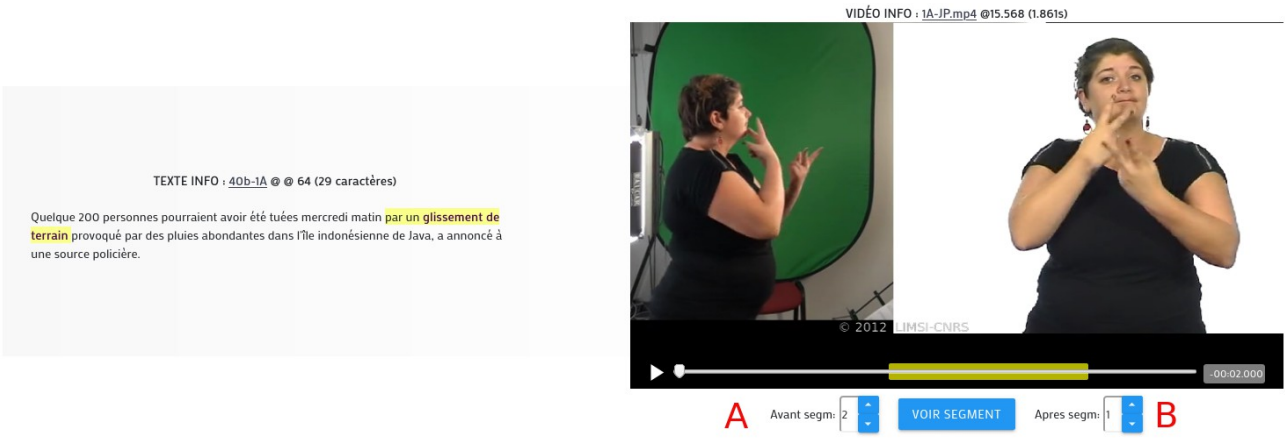

Figure 6: results page screenshot

We provided public access to this interface to allow professionals and academics to test it. It is available at https://platform.postlab.fr.[2] We invited translators to engage in testing for this project. We are collecting their feedback, in an iterative process to converge on the most adequate kind of tool for them to use in their everyday practices. Any reader interested in contributing to this feedback is welcome to e-mail us.

5. Future work and conclusion

We based ourselves on a written French/LSF parallel corpus to build an alignment data base. This data base can be searched using the concordancer we designed. With fewer than 500 entries, it is small for a concordancer resource. We will keep expanding it, but wish to open contribution to professionals and the academic audience as well, by making our concordancer a collaborative platform.

To allow this, our next planned step is the addition of a feature allowing users to add new content. We are currently developing this annotation function, whereby users select a text and a video in our library (one being the translation of at least a part of the other), select text and a video segment, and saving the alignment.

This being the first works concerning CAT & SL, and also concerning the elaboration of such data base, we hope to draw interest on the topic. SL are still seen as minority languages. Assisting translation would result in more visibility and by extension, maybe more consideration and recognition.

As an almost undocumented language, LSF resources are rare and those which can be found are oriented towards the research community and not the speakers – or signers – themselves. The corpus which we used to build our database can benefit professionals, the research community, and other domains, such a SL linguistics, SL teaching and interpretation training.

6. Bibliographical References

A. Balvet, C. Courtin, D. Boutet, C. Cuxac, I. Fusellier Souza, B. Garcia, M.-T. LHuillier, and M.-A. Sallandre.2010. The creagest project: a digitized and annotatedcorpus for french signlanguage (lsf) and natural gestu-ral languages. *In Proceedings of the International Language Resources and Evaluation Conference (LREC),Malta.*

M.Filhol and X. Tannier, 2014: Construction of a French-LSF corpus, Building and Using comparable corpora, *Language Resources and Evaluation Conference*, Island.

Kaczmarek and M. Filhol, 2019: Assisting Sign Language Translation: what interface given the lack of written form and spatial grammar *In proceedings of Translating and the Computer 41, London 2019 p. 83- 93*

M. Kaczmarek and M. Filhol, 2020: Use cases for a Sign Language concordancer, to be published *in proceedings of SignLang workshop 2020*.

P. Koehn, 2009 : *A Process Study of Computed Aided Translation,* Kluwer Academic Publisher.

S. Matthes, T. Hanke, J. Storz, E. Efthimiou, A.-L. Dimou,P. Karioris, A. Braffort, A. Choisier, J. Pelhate, and ́E.S ́af ́ar. 2010. Elicitation tasks and materials designed fordictasign's multi-lingual corpus. *In proceedings of the 4th LREC Workshop on Representation and Processingof Sign Languages: Corpora and Sign Language Technologies, Malta.*

2 To test it, registration is necessary. Please send us an e-mail following the registration so that we can grant the rights needed to access it.

M. O'Hagan, 2009: Computer-aided Translation (CAT) in Baker, Mona/saldanha, Gabriela (eds), *Routledge Encyclopedia of Translation Studies, London and New York, Routledge* p.48-51