Use cases for a Sign Language Concordancer

Marion Kaczmarek, Michael Filhol

Use cases 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 treats about a Sign Language concordancer. In the past years, the need for translated content into Sign Language have been growing, and is still growing nowadays. Yet, unlike their text-to-text conterparts, Sign Langauge translators are not equipped with computer-assisted translation software. As we aim to provide them with such software, we explore the possibilities offered by a first tool: a Sign Language Concordancer. It includes designing an alignments database as well as a search function to browse it. Testing sessions with professionals highlights relevant use cases for their professional practices. It can either comfort the translator when the results are identical, or show the importance of context when the results are different for a same expression. This concordancer is available online, and aim to be a collaborative tool. Though our current database is small, we hope for translators to invest themselves and help us to keep it expanding.

Keywords: Sign Language, Concordancer, CAT

1. Introduction

Translation is part of our world, and Sign Languages (SL) should be no exception. As far as France is concerned, the Law for Equal Rights and Opportunities, Participation and Citizenship of Persons with Disabilities published in 2005 recognizes French Sign Language (LSF) as a fully-pledged language, and as a « language of the Republic, the same way as French ». This law means that any public place must be able to welcome deaf people (either by forming their staff to SL or by means of professional SL interpreters), but also every piece of information provided (videos, written documents, or audio announcements). Helped by the CRPD in 2008 which puts emphasis on the right of people with disabilities to fully access information, the need for SL translated content is still growing. To try and fit the needs, a master degree in French/French Sign Language translation and mediation was created in 2011. However, there are still very few professional sign language translators.

And those few translators are not equipped with tools as the other translators can be. Indeed, no current Computer Assisted Translation (CAT) software is able to support SL translation.

Our aim is to provide CAT software dedicated to SL translation. The concordancer referred to in this article is a part of a bigger project. This paper, and our previous studies, are based on French Sign Language (LSF) as working language.

2. Brief state of the art

As our goal is to specify CAT software for SL translation, we first needed to learn more about it. To do so, we conducted studies involving professional SL translators and interpreters, including brainstorming sessions and observing them at work to analyse their practices. (Kaczmarek & Filhol, 2019). The results highlight their needs and the most common problems encountered, such as the scarcity of SL resources, the time spent looking for them as they are not always well referenced, the need for context and encyclopedic knowledge… Most of the identified steps taken in the process of text-to-Sign translation could benefit from already existing tools if they were able to support SL. We also pointed out what the major differences are between text-to-text translation and text-to-sign translation, sorted in four categories: no written form for SL, a *principle of linearity*, need for encylcopedic knowledge and CAT tools adaptation issues arising from the previous categories. (Kaczmarek & Filhol, 2019)

On the other hand, the major innovation brought by CAT software is Translation Memory (TM). It allows the translator to store prior work and reuse it later. Once a segment of source text is translated, the source-translation pair is stored in memory. When the translator encounters a similar one, the TM automatically suggests the prior translation. It can be shared with colleagues or even provided by the client himself. The is a time saving tool which has had a great impact on the everyday practices of translators, evolving from to translating from scratch to mostly post-editing TM entries and suggestions.(Lagoudaki 2006; O'Hagan 2009).

A concordancer is, regardless of the languages, a search engine which can look through corpora and list each and every occurrences of a queried word. When it comes to translation, bilingual concordancers are used. The query is done in a source language, and results are provided with an aligned translation in target language, in our case LSF. Such tool allows translators to look up words or expressions in context, to determine how common they might be or with which style of discourse they are more often associated with.

No such thing currently exist for SLs. The next paragraph explores how to create one.

3. Designing a concordancer

SLs do not have editable written forms, so video is the most common way to keep trace of it. This brings a problem when it comes to the adaptation of a TM tool for SL, as chaining video extracts from previous translations alone would result in an unacceptable translation.
As previously said, TM stores alignments, in other words pairs composed of two text segments, where one is the the translation of the other for two given languages. This data can be searched with concordancer.

This is why we elaborated a SL concordancer to keep the benefits of the TM. The alignments consist in pairs made of a segment of the source text, and its SL translation identified in a video with time tags. Such alignments are stored in a database built by the users themselves, and which can be shared just like a TM.

3.1 Alignments database

The creation of such database is the main topic of another article (Kaczmarek & Filhol, 2020) in which you can find more details. The next paragraph explains briefly its key points.

As there is currently no automatic way to produce text-SL video alignments, we built the first database ourselves by aligning manually. We used a French–LSF parallel corpus of forty short news texts, of three to five lines each in a journalistic style ("40 brèves", which has ISLRN).

Each text was translated by three different professional SL translators, and the resulting translations filmed using two cameras for a front and a side view, for a total of 120 videos of an average 30-seconds duration. We used it because it is a parallel corpus of short translations , so that each video already provides us with a useful alignment (the 30-seconds video can be aligned with the 2-3 lines text). As the videos are shorts, those are still interesting alignments to include in the database.

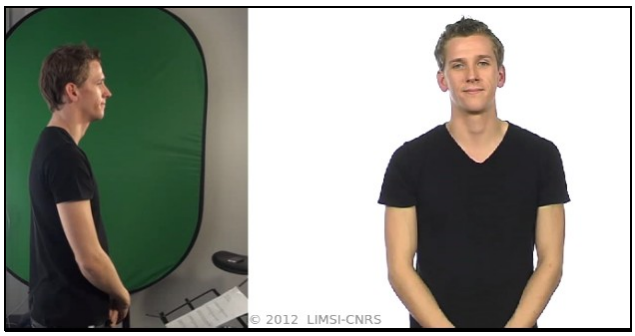

Figure 1 : A screenshot of the video set-up

We chose a few smaller segments : words, idiomatic expressions, grammatical phenomena or figure of speech. As we had three different signers, we were free to vary the spans for a given expression.

The identified expressions were also chosen either based on their lack of standard signs, or on variety of translations proposed by the signers. For each text segment identified, we search for its translation in the associated video and extract the corresponding time-tags. The selected segments are also suitable for aligning in terms of simultaneity. We cannot for example, align an adverb in the text with only the facial expression of the signer.

Alignments are stored in a database with the following format:

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

- 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 number of the first letter from the text segment, in the source text.
- Length is the length of the segment, in a matter of caracter string, spaces included.
- VidID is the unique identification of the video in their own storage space, which allows to retrieve them.
- Start time is the time tag corresponding to the beginning of the SL translation in the video.
- Duration is the total duration of the segment's translation.

3.2 Search function

The search algorithm looks for matches for the string entered as a query, yet brings up as an answer only the shortest segment which contains it. A query is usualy an exact match, plus you can:

- Use the modifier # at the end of a word to replace any suffix. You can add a number to limit the number of letter added. Searching for "person#" would bring up person, persons, personal, personality… and so on. However searching for "person#2" would bring up only words starting with person- and ending with only 2 more letters maximum, such as person, persons, personal…
- Use the ### modifier to split a query. Place between two words, it means "with at least one word here". For example: "Roughly ### participants" can be answered with any kind of numeral value between Roughly and participants. However, "Roughly ###1 participants" will only be matching with segments containing two to twelve as a number. "Roughly one participants" will be excluded because of the plural in "participants". This case can be solved using the "Roughly ###1 participant#" formula.

If the expression queried by the user has been previously aligned, the concordancer answers with the smallest span found. If it has not been aligned but still previously translated, the concordancer answer with the video of the entire source text that contains the query. If it has never been translated before, the concordancer cannot answer the query.

The concordancer itself is currently available on-line, at the following address: platform.postlab.fr . You can either test it using a public test account (please e-mail us to get the login information), or create your personal account on the website. If you chose to do so, please e-mail us as well, so we can promote your account with the corresponding user rights.
If you are interested to contribute to the database, please take contact with us. We are currently collecting feedback from our users, which may lead to a later communication.

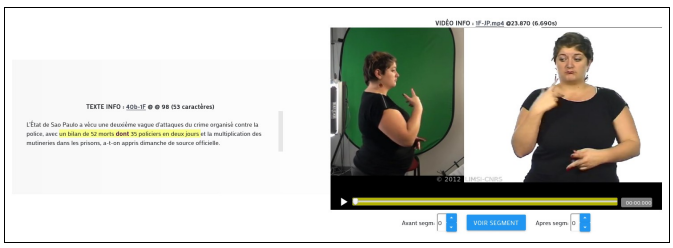

Figure 3 : The result page format

The figure above is a screen shot of the concordancer, showing one of the items matched for the

query *dont* which means “whose”, “of which” or “including”. On the left is the entire source text in which the query was matched. The segment aligned appears in yellow, and the exact query is in bold. The line above the text gives information about it: the TxtId, start pos. and length of the segment, as well as a link to open the entire text in a new tab. On the right, the video appears centered on the segment (which also appears in yellow on the time line). The title above provides the same kind of information as for the text, and a link to open the entire video in a new tab. The video is looping on the segment, and the buttons below allow the user to add left or right context (in seconds) around the segment.

As mentioned by the professional translators and interpreters during our prior studies, LSF resources are rare. And those rare resources are often badly documented or sometimes hard to access. In addition to the convenience that such tool can bring to the translators’ everyday practices, it can also be an opportunity to explore the language in a unique way. To an extent, this tool could have a certain use in teaching not only LSF, but also in teaching translation and interpretation methods by displaying in a very readable way a list of examples, counter-examples, as well as unique constructs to think about in class. Following are 3 examples of phenomena we did observe while working on our database, which could help either SL learners or SL linguists to understand better the language, to speak it or describe it. Those are also typical use cases for a text-to-Sign translator. During a previous workshop, we presented some professional translators with texts to translate into Sign Language. We built those texts around idiomatic French expressions or complex semantic or syntactic rules. When asked which part of the text they would most likely search for in a concordancer, those three examples were among the most identified ones.

4. Use cases

4.1 No standard sign, yet common form

Some frozen French expressions do not have standard signed equivalent. It is the case for the French expression *fin de non recevoir,* which can be translated in English by “refusal to consider one’s request”. But the fact that there is no standard way of signing it does not mean that it is untranslatable. When searching the concordancer for *fin de non recevoir*, three results came up. The three translators worked alone on their translation, still they chose a similar construction to translate this expression. The sign used here is the one for “to reject”. Their facial expressions are also similar, as well as the spacial construct they are using. Overall, the entire source text is translated in a quite similar way, meaning the context has only a low impact on the translation choices.

Figure 4 : Three ways to translate *fin de non recevoir.*

Three times the same construction seems trustworthy even if there is no standard sign. This kind of results can comfort the translator, either in his choice of translation, or encourage him to reuse the same construction in his own work.

4.2 No standard sign, and no common form

Here again, the French expression *mis à mal*, which means “suffering from a negative effect of something or someone”, or “to be harmed”, does not have an out-of-context equivalent is SL. Still, we can find three examples of its translation in our database, and the three of them are different. The sentence here was “Hopes for peace si Sri Lanka are once again dashed after a major military offensive against Tamil rebellion”.

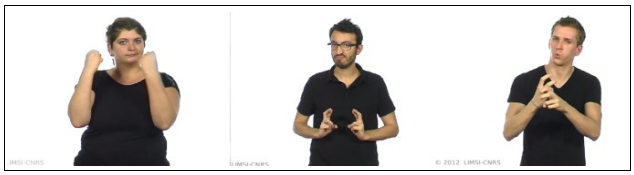

Figure 5: Three ways to translate *mis à mal.*

On figure 5, the signer on the left uses the sign for “break”, and the one on the right the sign for “difficult”. The signer in the middle uses an iconic structure based on the French sign for “hope”, which he signs falling down crumbling. Here, the way of translating is more influenced by the context than in our first example.

Three different results for the same expression translated from the same source text. This kind of results allows us to see how different matches can be in context. This is very useful for professional translators to build on what has already been done, but also for the learners to better understand the finer points of the language.

4.3 Cause/effect relationship

The concordancer is also a interesting way to observe grammatical phenomena such as this one. Cause/effect relationships can be translated in many ways depending on two things: the translator’s choice and the context. The translator is free to use the sign for “then”, as the signer on the right does on figure 6. The two ohters made the choice of using iconic structures to depict the event: an underwater earthquake occurs and causes a tsunami (*un tsunami causé par un séisme sous-marin* in French). For the left and middle signers, the cause/effect relationship lies in the order of events and in the facial expressions, as well as in the dynamics of their speech. There is a specific transition time between the two events mentioned, and their signs.

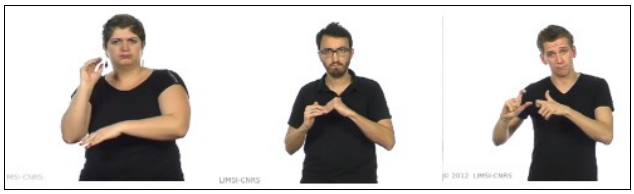

Figure 6: The ways of translating *causé par.*

Figure 7: Translation for *provoqué par.*

On Figure 7, the sentence to translate was “a landslide caused by heavy rains”. Both of them picture the waterlogged ground, and then the landslide itself. The screen shots were taken right before the landslide part, and we can see that the cause/effect relationship relies here again on the timeline of the vents but also in the facial expressions, and the very short transition time between the two, with raised eyebrows and chin as if to call for the viewer’s attention.

5. Conclusion

Designing a SL concordancer first implied to build an alignments database. We are aware that our first database is rather small, but hopefully it will keep expanding. The examples detailed comfort the relevance of a SL concordancer as a tool to equip the translators. The very first feedback we received showed enthusiasm and interest in our work. The first prototype for the concordancer is fully

working, and we are now waiting for some more specific feedback about the function itself, in an iterative process to converge on the most adequate kind of tool for them to use in their everyday practices.

We are now working on a alignment function, which would allow our users to create their own alignments database in an easy way. Based on pairs of text and signed video displayed alongside (where one is the translation of the other), the user can select a segment in the text and identify the corresponding part in the video using start-stop tags. The alignment created this way is then stored in the user's database, which is available in the search function. He can later consult his work, and report any problems or needs for databases modification. We hope that the translators who helped us during our studies will continue to invest themselves in this project, but also that others will join.

6. Bibliographical References

M. 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), Elaborating an Alignments Database for a Sign Language Concordancer, *to be published in proceedings of LREC 2020.*

E. Lagoudaki (2006) Translation Memory Survey, *Translation Memory systems : Enlightening users' perspective. Imperial College London.*

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

UN General Assembly, *Convention on the Rights of Persons with Disabilities : resolution / adopted by the General Assembly*, 24 January 2007, A/RES/61/106