Construction of a French–LSF corpus

Michael Filhol, Xavier Tannier

LIMSI–CNRS
B.P. 133, 91403 Orsay cedex, France
michael.filhol@limsi.fr, xavier.tannier@limsi.fr

Abstract

In this article, we present the first academic comparable corpus involving written French and French Sign Language. After explaining our initial motivation to build a parallel set of such data, especially in the context of our work on Sign Language modelling and our prospect of machine translation into Sign Language, we present the main problems posed when mixing language channels and modalities (oral, written, signed), discussing the translation-vs-interpretation narrative in particular. We describe the process followed to guarantee feature coverage and exploitable results despite a serious cost limitation, the data being collected from professional translations. We conclude with a few uses and prospects of the corpus.

Keywords: Sign Language; text–video parallelism; elicitation

1. Motivation for a French–LSF corpus

Sign languages are part of the less-resourced languages of the world, which means that very little data is available, and indeed linguistic knowledge all together remains limited. The Sign linguistics field has reached no agreement comparable to the more or less stable theories describing a language like French or English. Significant matters such as where and how—and even whether—to draw a line between the language construction layers, e.g. lexicon and syntax (which though not definitely do more obviously appear in written languages), remain open questions.

As for any such language, one can hardly hope to find sufficient data on a specific language feature without building an elicited corpus beforehand to serve the study. For French Sign Language (LSF), a few accessible corpora are available (LS-COLIN, 2000; Matthes et al., 2010; Balvet et al., 2010), but the community is still strongly confronted to the data limitation. Moreover, in our context of automatic or assisted translation, we felt we required not only Sign Language data, but language data for both French and LSF, in view of comparing linguistic features and structures between the two languages.

The DEGELS corpus (Braffort and Boutora, 2012) would be a little closer to our needs than SL-only data, as it involves two languages, namely spoken French and LSF. It is a comparable audio-visual corpus built for a comparative study of gestures in vocal and signed languages in face-to-face communication. To our knowledge, the only bilingual data available including written French is the feed of written news items selected and reduced from the AFP newswire, published daily on WebSourd's[1] website together with their equivalent version in LSF (cf. fig. 1). The signed version is translated, signed and recorded by professional French-to-LSF translators.

However, the WebSourd data is intended for short-term online viewing, not for academic research. Besides the data collection problem requiring that we save the few videos daily with no control on the contents, the videos come in a lossy Flash encoding format, which is a problem when analysing finer details such as the direction of the eye gaze. A better geometric and time resolution would be a requirement for any thorough study on such feature. Also, it is important for corpora to enclose relevant meta-data, for example on the informants' connection with the language to enable regional variation awareness, Sign Language (SL) still not being well documented in that respect. These were enough reasons to motivate us to build a reference corpus joining written French and LSF, for academic research and sharing.

Figure 1: WebSourd's website with the daily list of news items

2. Problems with text–SL parallelism

The oral (live production) nature and the oral-only status (no written form) of SL together have significant consequences on the way one can address translation.

First, when working with text for both source and target languages, the translator is enabled to produce a first wording of the source meaning, and work from it iteratively. Alternatively interpreting both texts, he can modify the target translation until its distance in meaning and effect to the source is satisfactorily low. This convergence process is

[1] A company providing accessibility services to the deaf public. http://www.websourd.org

to us what defines translation. It contrasts with captioning and interpretation, whether live or consecutive, which only allow one shot for output delivery. The "one shot" criterion we define here brings a contrast to the common use of the terms, where *translation* is written and *interpretation* is oral. Of course, a reason is that such distinction is not applicable to SL as no written form exists for the language, but we also think that the two activities are different in nature, and that both are possible in SL: translation if the output can be reworked on (refilmed for example); captioning and interpretation otherwise.

Because of corpus shortage, some projects have made use of interpreter's data as parallel data for translation research (Forster et al., 2012). The problem with such use is precisely that interpreted recordings are one-shot deliveries, i.e. not reviewed, not corrected. While interpretation services remain crucial and the best solution for accessibility and to enable cross-language discussions, their one-shot property makes them subject to undetected mishearings and source language bias. For example, simultaneous interpretation will at least have to follow the sentence-level chunking of the source, which is not necessarily appropriate in the target language. To avoid this bias, we make the claim that building a French–LSF parallel corpus must allow the to-ing and fro-ing between the text and the signed output, i.e. in our terms requires a translation process.

But a second problem exists when translating to the oral modality: the result eventually needs to be memorised and delivered by heart. Regardless of how prepared the output is, video capture of the translation requires that the signer performs it live, from the beginning of the message to its end. Use of a white board with personal notes behind the camera or allowing segmented production are possible tools to cope with somewhat longer texts and avoid omissions, but this is essentially a problem to which no real solution is known yet.

For now, we choose to translate texts that are short enough to remain within the limits of memorised productions, thus clear of hesitations and not requiring post video edition. In this way, our corpus tends to be a fully parallel corpus. But, as we have just seen and because of unavoidable memorising, perfect parallelism is arguably unreachable. Moreover, the community of researchers interested in corpus parallelism usually include chunk, sentence or lexical alignment, which does not apply well here. In this sense, our corpus is not a fully parallel one. This classification problem already emerged in an earlier paper where Segouat and Braffort (2009) attempted to categorise existing SL corpora. For these reasons, we prefer to situate our corpus somewhere between a comparable and a parallel set.

3. Preparing for the corpus

WebSourd textual documents are short summaries of AFP newswire articles. They contain one or two sentences for an average of 39 words. They normally describe the five 'W's of the reported event: *what*, *when*, *where*, *who* and, as much as possible, *why*. For example:

(1) "Abidjan, la capitale économique ivoirienne, était à nouveau paralysée mercredi, pour le troisième jour consécutif, par des jeunes partisans du président Laurent Gbagbo qui tiennent de nombreux barrages dans la plupart des quartiers, rendant la circulation quasiment impossible." *(Abidjan, the economic capital of Ivory Coast, was again paralysed on Wednesday for the third consecutive day, by young supporters of president Laurent Gbgabo, barricading most of the town districts and almost blocking the traffic.)*

News items were judged the ideal genre for our purpose, for different reasons:

- the domain is not restricted, the news reporting about events in virtually all contexts;
- the language is standard (no grammatical errors), and meant to be concise (no bloat or repetition) and unambiguous;
- productions involve times, places, protagonists and events, with clear relationships between them, which typically triggers heavy use of signing space, a SL specificity requiring scientific attention;
- our lab had worked with the AFP newswire feed in different projects, so we could benefit from local expertise and systems.

Our goal being to provide a corpus of reference translations, we have used the professional service of native deaf translators whose SL performance is acknowledged by the community. Professional service being costly (and currently about 10 times more by the word into SL than into a written language), it is important to select the source material and control redundancy in a way that limits noise but not linguistic use cases. A point was made to work on real-life text excerpts to avoid any fake language intrusion in the source. Hence, we decided to select a set of 40 articles among the textual news archive from WebSourd, and for cross-informant comparison, have each one signed by 3 different informants (translators). The way we chose the texts is one of the main points of our contribution, and presented in the remainder of this section.

First, we restrained the domains of linguistic features to appear, to give us a chance of building a model of a language subset. Otherwise, we would barely have collected a list of positive examples with too few of each feature to enable generalisation. However, to avoid all texts to look alike and lead our informants to guess too much of what is being analysed because of a too narrow focus, we chose four elements of focus, related to events and temporality. This choice was partly due to the fact that we already had expertise on time expressions and events from prior work in text analysis (Moriceau and Tannier, 2014; Arnulphy et al., 2012), which gave us background on the related theoretical aspects as well. Also, results on the expression of time in SL had been published[2] and showed a relevant space mapping of time anchors on all spatial axes (vertical, sagittal and horizontal left-to-right), dictated by certain semantic criteria.

The four studied features, non mutually exclusive in a single article, are the following:

[2] many referenced by Fusellier-Souza (2005)

- Event precedence: one happening before, just before or after another or a date;

  (2) "Un éleveur français de 62 ans, Christophe Beck, enlevé il y a un peu plus d'un an au Venezuela et dont le nom était depuis tombé dans l'oubli, a regagné la France dimanche, cinq jours après sa libération contre rançon par ses ravisseurs." *(Christophe Beck, a French 62-year-old breeder kidnapped in Venezuela just over a year ago [...], came back to France on Sunday, five days after being released for a ranson.)*

- Durations of events or of periods separating/preceding/following events;

  (3) "Un homme de 25 ans qui voulait braquer un bureau de Poste à Limay (Yvelines) a retenu jeudi pendant trois heures cinq personnes en otages avant d'être tué par la police." *(A 25-year-old man who was trying to hold up a post office [...] took five people hostage during three hours, before the police finally killed him.)*

- Causal relationships between events;

  (4) "Au moins 525 personnes ont été tuées en Indonésie par un tsunami causé lundi par un séisme sous-marin, selon un nouveau bilan annoncé mercredi par le gouvernement." *(At least 525 people have been killed in Indonesia by a tsunami caused Monday by an underwater earthquake, according to [...])*

- Repeated—or repetition of—events.

  (5) "Le lancement de la navette Atlantis est prévu ce mercredi à 16h29 GMT de Cap Canaveral (Floride), après avoir été reporté trois fois depuis le 27 août à cause d'orages puis de la tempête tropicale Ernesto." *(The launch of space shuttle Atlantis is expected Wednesday [...], after being cancelled three times since August 27 because of [...])*

All these relations were marked as true only if made explicit. For instance, causal relations are difficult to distinguish from simple precedence of events in the case of expressions such as "suite à" *(following)* or "à la suite de" *(in the wake of)* , which often require pragmatic or expert knowledge. Too strong an ambiguity would impair the comparison of the three collected translations, as different productions could either be imputed to different ways of expressing the same relation or to different understandings of the relation by our informants. As our resources are limited, we cannot afford this ambiguity.
To guide our selection, we listed a number of semantic criteria to discriminate items in each category, for example:

- whether or not the date of each event is made explicit in the text;
- whether or not the duration between two related events is made explicit in the text (*three days after...*);
- for a repeated event, whether the number of repetitions is made explicit (*three times*) or not (*again*).

The idea behind these criteria is that we will elicit more different structures in SL if we cover more qualitative semantic distinctions. Though the language still remains little documented in that respect, its iconic power in concision undoubtedly makes its underlying system sensitive to semantic variations more than merely to syntax.
Finally, a set of 40 texts was chosen to balance all the criterion values and create a sample as representative as possible. Table 1 provides more information concerning the distribution of studied phenomena in the corpus.

4. Filming and editing

We then proceeded to the actual corpus capture, arranging studio sessions to collect the translations of the same text material by three different translators, totalling 120 signed clips and an hour of signing. To enable movement analyses on all three axes, we filmed the informants from both facial and side views.
Along the same lines as the choice of real-life texts and to avoid any undesired discomfort which might inflict on their language, we did everything to place the translators in their usual set-up. They discovered the news in the morning, and would be left to work on it until they signed it in their own studio. The only notable difference was the additional side-view camera, which they were aware of, and the requirement to clap their hands, arms extended with horizontal palms, before every translation for video synchronisation purposes.
Once synchronised, the two camera views were rendered side by side in a single video file, as shown in the still picture of figure 2. The resulting corpus is a set of 40 news texts in French and 120 LSF video files totalling 1 hour of elicitated signed production, where each text is translated by 3 informants. The whole contents will be made available during the year, together with their respective text equivalents.

Figure 2: Still shot of a corpus video file

5. Uses and prospects

The first usage of a corpus, especially in the case of almost undocumented languages, is to be searched for patterns which may lead to the establishment of rules, together forming a grammar. The video part of this corpus has actually just served to formalise a rule system for past event

Date of the event	
Date is made explicit	Date is not made explicit
30	44

Gap between two events	
Gap is known and precise	Gap is fuzzy or unknown
4	5

Repeated events		
First occurrence of a repeated event	Other occurrences	
	Number is known	Number is unknown
4	5	5

Table 1: Number of occurrences of each criterion in the corpus. The total is higher than 40 because several events can be described and several phenomena can occur in the same document.

chronologies (dates, precedence and durations) in LSF (Filhol et al., 2013).

As for its bilingual property, an immediate prospect of this corpus lies in machine translation research, a domain on which several efforts have been summarised in Morrissey (2008). From the statistical point of view of course, data of this size will not be helpful enough to fully train a translation or a language model. Furthermore, statistical learning will normally need pre-aligned bitexts, whereas the video nature of the translated part (unsegmented and continuous stream of pixels) and the non-sequential syntax (simultaneity) of Sign Language together make this difficult. Thus even big enough such type of corpus may not serve the approach.

However, this corpus can be very useful for text-to-SL machine translation evaluation, whether based on statistical learning or on linguistic rules. Translations not being unique, we must rule out a simple comparison between the corpus data and the system's output, but such corpus can serve as a validation by positive comparison of similar output. Also, the fact that we have three productions for every text can help elaborate new metrics with a philosophy similar to BLEU, a typical score measure of statistical text-to-text translation systems based on edit distances to a set of human reference translations (Papineni et al., 2002).

Future work is required to address longer texts in bilingual corpora involving a Sign Language, especially when a parallel status is desired. We propose to work the other way around and build a corpus from signed production as input translated into text. This would allow the iterative process of translation to rather apply on the text, and indeed guarantee that no bias from the text is carried into the sign discourse, by design.

6. References

B. Arnulphy, X. Tannier, and A. Vilnat. 2012. Automatically Generated Noun Lexicons for Event Extraction. In *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CicLing 2012)*, volume 2, pages 219–231, New Delhi, India, March.

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 annotated corpus for french signlanguage (lsf) and natural gestural languages. In *Proceedings of the International Language Resources and Evaluation Conference (LREC)*, Malta.

A. Braffort and L. Boutora. 2012. Degels1: A comparable corpus of french sign language and co-speech gestures. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

M. Filhol, M. N. Hadjadj, and B. Testu. 2013. A rule triggering system for automatic text-to-sign-translation. In *Sign Language translation and avatar technology (SLTAT)*, Chicago, IL, USA.

J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. 2012. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

I. Fusellier-Souza. 2005. L'expression de la temporalité en langue des signes française. *La nouvelle revue AIS*, 31.

LS-COLIN. 2000. http://www.irit.fr/ls-colin. Project website (final report available).

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

V. Moriceau and X. Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In *Proceedings of the 9th International Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May.

S. Morrissey. 2008. *Data-driven Machine Translation for Sign Languages PhD Thesis*. Ph.D. thesis, Dublin City University, Dublin, Ireland.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

J. Segouat and A. Braffort. 2009. Toward categorization of sign language corpora. In *Building and Using Comparable Corpora, LREC*, pages 64–67, Singapore.