Elicitation protocol and material for a corpus of long prepared monologues in Sign Language

Michael Filhol, Mohamed Nassime Hadjadj
LIMSI, CNRS, Université Paris Saclay
Rue John Von Neumann, 91400 Orsay, France
{michael.filhol, hadjadj}@limsi.fr

Abstract

In this paper, we address collection of prepared Sign Language discourse, as opposed to spontaneous signing. Specifically, we aim at collecting long discourse, which creates problems explained in the paper. Being oral and visual languages, they cannot easily be produced while reading notes without distorting the data, and eliciting long discourse without influencing the production order is not trivial. For the moment, corpora contain either short productions, data distortion or disfluencies. We propose a protocol and two tasks with their elicitation material to allow cleaner long-discourse data, and evaluate the result of a recent test with LSF informants.

Keywords: Sign Language, elicitation, discourse

1. Purpose and motivation

Sign Language (SL) corpora have long been scarce resource, which impeded linguistic studies as access to observable language productions was limited until this millenium. At the same time the growth of SL recognition and spread of Sign linguistics increased the need for properly recorded data. Also then, light digital cameras together with video viewing, editing and annotation software became commonplace, and cheap. Consequently, a lot more data became available to SL research and several projects were even almost dedicated to produce reference corpora, in a variety of genres. We will be referring to them in this paper.

However, the nature of recorded Sign still makes it more difficult to collect, use, browse, share. It often weighs over 200 times more storage space than its equivalent in text, is not yet searchable in its primary form, and is almost impossible to anonymise, which makes any publicity subject to legal requirements. So while corpus studies have undoubtably been facilitated by a growing number of data collections in the past two decades, finding an appropriate corpus and gaining access to it if any exists is often still a problem when starting a SL study.

In our recent work, we have been studying camera-facing short news reports in French Sign Language (LSF), to model a neutral reference language genre without limiting possible topics. Now we would like to study longer discourse of the same genre to look at higher-level discourse/rhetoric operations.

Arguably, this is an idealised version of the language since it is not the way SL is used most of the time. Nevertheless it captures the well existing idea of canonical fluent discourse, i.e. constructed discourse with none of the following:

- interference from external channels, e.g. dialogue interruptions, reaction to feedback signals;
- disfluency, e.g. hesitations, backtracks, etc.

Avoiding interference justifies the use of camera-facing monologue, as opposed to conversation setups. Avoiding disfluency requires prepared discourse, as opposed to spontaneous. In other words the final production is intended to match a planned result, not built on the fly. Ensuring this allows to afford confidence that observed articulations are a product of intended—hence assumed correct—language usage, and not that of a reparation strategy or filler for a pause for thought.

2. Existing material and properties

Various corpus projects were conducted, of different genres, using different setups and produced with more or less specific purposes. They are now numerous enough to prevent exhaustivity (though not always easily accessible), but this section presents a few contrastive major examples, relevant for comparison with our objectives.

Many corpus projects have a dialog task setup. Dialogue is generally more ecological in the sense that it captures language in its most used and living form, but does not allow to model canonical discourse-level structures.

Often though, one party is (or plays the part of) a listener only, merely asking a few questions while avoiding to interfere with the discourse. The "BSL corpus"[1] (Schembri, 2008), the "corpus NGT"[2] (Crasborn and Zwitserlood, 2008) and some tasks of the DictaSign corpus (Matthes et al., 2010) fall in this category. This does open a window on longer monologues, but nevertheless keeps the dialogue feedback channel open, which inflicts on the dynamics of the production. And a bigger problem for us yet is the lack of discourse preparation, as those tasks are generally elicited narrations, organised and produced spontaneously.

A few resources contain prepared elicited discourse, like the story-telling tasks in LS-COLIN (Cuxac et al., 2002), the joke task in the DGS Korpus (Nishio et al., 2010). In LS-COLIN, informants were given the assignment and elicitation material in a separate room prior to standing in the studio, and could take the time to prepare their production. As far as the documents describe the collection process, no material was visible during the recording though. By

[1] BSL = British Sign Language
[2] NGT = Dutch Sign Language

contrast, the entire DictaSign corpus was collected with a helper screen on which the task elicitation material was available. It did not display informants' preparation notes (the corpus does not contain prepared discourse), but it is an example of corpus collection making use of inline visual support.

For our own work on (short) news reports, we created a corpus "40 brèves" that matched our requirements of prepared camera-facing monologue, inspired by the WebSourd® internet site in its time[3]. The resource (Filhol and Tannier, 2014) is a set of 120 videos (30-second average duration, face and side views), prepared by professional translators from short news items written in French. We note that they chose not to use visual support while signing, except for occasionally long proper names written on a whiteboard to support inline fingerspelling[4].

3. Problems

Our recent studies have therefore mostly been based on videos of prepared captures which are short in length. To enable studies of higher-level discourse operations such as rhetoric argument building, semantic sectioning and long-term contextualising, we developed a need for both prepared and long Sign Language monologues. This section presents a few challenging problems that lie in the way of such corpus collection.

3.1. Protocol

Contrarily to spoken languages, SLs have no written form which can be read out and captured without a trace of the preparation. Therefore, prepared discourse usually has either to be delivered by heart, or to work around the lack of written support. This means that all SL corpora containing prepared discourse fall in either of the categories below:

1. delivered by heart, and short enough to allow hesitation-free discourse in one take after enough rehearsal;
2. longer rehearsed productions, at the expense of discourse flow, i.e. containing some disfluency in the contents, e.g. backtracks for omissions, stalled and repeated items, eyes disengaging and recollecting thoughts;
3. well organised discourse sequences, at the expense of allowing visible prep notes during the shooting process, e.g. a knee pad, a whiteboard facing the informant, a screen under the camera ("teleprompter" technique).

Case (1) was fit for the average 30-second item of the "40 brèves" corpus, as it allowed fully controlled output, still ensuring a reasonable prep time. One could learn long discourse like actors literally learn hour-long plays, but it would take the rehearsal process to a level beyond what can reasonably be expected from a corpus informant. Short-term memory here does not allow fully controlled productions that exceed a minute in duration.

[3]It shut down with the company in 2015.

[4]Fingerspelling is a way of spelling a written name or word using an alphabet of handshapes.

The remaining techniques do allow for longer discourse, but have symmetric advantages and drawbacks. On the one hand, memorised discourse (2) inevitably results in disruptions in the discourse flow. On the other hand, allowing visual support at the time of capture (3) generates false articulations and dynamics because of the physical constraints added to head orientation, eye gaze direction, etc.

A map task in the DGS Korpus project was planned and actually even aborted after its pilot test for that reason (Hanke et al., 2010). A similar task was nonetheless conducted in DictaSign, using a screen under the camera so that the informant could see it at any time. The result is a strong effect throughout the task, frequently as obvious as postures held with arms stopped in a physically constrained position, back hunched, eyes apparently locked on the helper screen, squinting and scanning its contents, until the signing resumes (fig. 1). In such case, like in the DGS Korpus map task, data distortion is too strong and the data is no more useable to inform studies of regular SL gestures.

Figure 1: Data distortion in the DictaSign map task.

3.2. Elicitation

Elicitation of long discourse in Sign Language also becomes tricky as the material given to elicit the productions can create bias.

The "40 brèves" corpus was signed by professional translators, and we observe that all productions (100%) have changed—if not completely reversed—the order in which named entities were introduced in the discourse. Translation being precisely the task of not changing the meaning, it shows that a lot is imposed by the language on discourse construction, which is what we want to study, and not bias. Now source text given for translation is known to generally have an impact in the target language equivalents, especially if translated by people not professionally aware of the problem and trained to overcome it. So to allow for a wider range of informants in future tasks, including native signers uncomfortable with text all together, it is necessary to avoid text as elicitation material.

A common technique to elicit narration is to ask the informant to read a picture story (approx. 1 page long), and to retell it. Several corpora cited above incorporate such a

task. Interestingly, very little reordering takes place with picture stories in the resulting discourse. Unlike translation of texts for which sentence-by-sentence progression is very likely wrong[5], a picture-by-picture progression this time is often the result everybody expects. However, it is probably due to the chronological nature of the story itself, imposing its own sequence of events, hence may be limited to the story-telling genre.

4. Proposal

We explained that when collecting long prepared discourse, one generally chooses between discourse flow and undistorted articulations.
But our objective is to enable discourse-level observations, and SL cues meaning "that is why", "besides", etc. are very often conveyed by subtle head tilts or slight rhythmic breaks not necessarily accompanied by annotatable dictionary signs. We therefore can afford to sacrifice neither flow nor articulatory correctness. This section suggests a protocol to collect corpus data with the most of both.

4.1. Protocol

Memory works only up to a limit in discourse length. So allowing for prep notes seems unavoidable to secure long discourse flow. Now when notes are allowed, experimenters generally try to maximise the ratio between visibility of the notes to the informant and invisibility on the resulting film. So they strive to embed the blackboard or monitor in the camera setup as much as possible, like the ideal teleprompter is invisible to the television viewer.
However, we argue that note support undermines the production *whenever* they are visible to the informant while they are performing, hence must be avoided then. And sparing the later viewer (e.g. linguist) of the sight of notes in the result is no serious concern to them. Arguably even, in view of analysing the data, making the observer oblivious to the moments when informants might be looking at their notes can turn out a serious problem. Being able to tell which articulations are not the result of an external cue increases confidence in taking them for linguistic features, as opposed to possible corpus artefacts.
Moreover, we suggest that note support undermines the production *only when* they are visible to the informant while they are performing. Whether or not they look at notes outside of what is later analysed is not relevant.
In summary:

- only note-supported footage can reliably yield long yet fully fluent discourse;
- only non-supported signing can be analysed as fully acceptable language productions;
- memory allows for non-supported but short stretches of clean discourse;
- any break in the flow of discourse should be identifyable in the resulting data.

Therefore, our proposition to collect long and fluent elicited signed discourse is:

- submit the assignment in advance to the informants, together with a description of the following rules, for them to prepare their intended productions;
- allow any preparation notes, drawings, personal recordings, etc. in the studio at the time of collection;
- prevent any visual access to them during signing;
- allow any number of breaks at any moment of the production, provided they are made obvious (e.g. signing stops, pause is called, informant must turn around to further read notes).

The periods between points where the signing breaks and resumes are then marked and edited out. The resulting corpus data is left with fully unsupported discourse, of arbitrary length, whose rhetoric and logical sequence is reliably planned, only containing a number of breaking points.

4.2. Elicitation

As we said in §3.2., it is hard not to bias the discourse construction order when expecting longer eventful stories. We acknowledge that pictures are a good alternative to avoid signed input, only we wish also to avoid pre-constructed discourse which tends to be induced by picture stories.
We also acknowledge that translation is in principle a good way of eliciting any exact meaning, but aside from the very limited world of native professionals, it is difficult to find reliable informants for such task.
However, there is a middle ground which translators are often trained to move to as a first step from text, namely *deverbalising* (Seleskovitch and Lederer, 1985). To deverbalise a text is to draw an explanatory diagram, in which all entities are present, together with their relationships and all other relevant information. Discourse is then built to from this drawing, in other words without any influence of the sequence and lexical choices of the source.
In elaborating new elicitation material, we chose to try out two different strategies, respectively inspired by the two above. One involves a picture sequence corresponding to the chronological sequence of the events depicted, about RMS Titanic. However, to avoid productions too strictly focused on signing each picture in turn, we have included informational pictures (size of ship, number of life boats, etc.) which do not contain an event and whose contents informants could choose to include anywhere they saw fit. It is too long to be included here but it is available online for download[6].
The other strategy consists in a two-page deverbalising diagram about the Omar Raddad affair, a famous unsolved criminal case in France. It bears no inherent order or sequence, but contains a lot of information which had to be ordered. It is given in figure 2, and is available online[7].

[5] Working with news texts a lot, we observed that the headline information came first in a paragraph while pieces of context and satellite information were always appended. When translated to SL, this was systematically reversed to contextualise first and bring the major (focused) clause last, inside the context previously set up.

[6] perso.limsi.fr/filhol/research/files/elicitation-Titanic.pdf

[7] perso.limsi.fr/filhol/research/files/elicitation-OmarRaddad.pdf

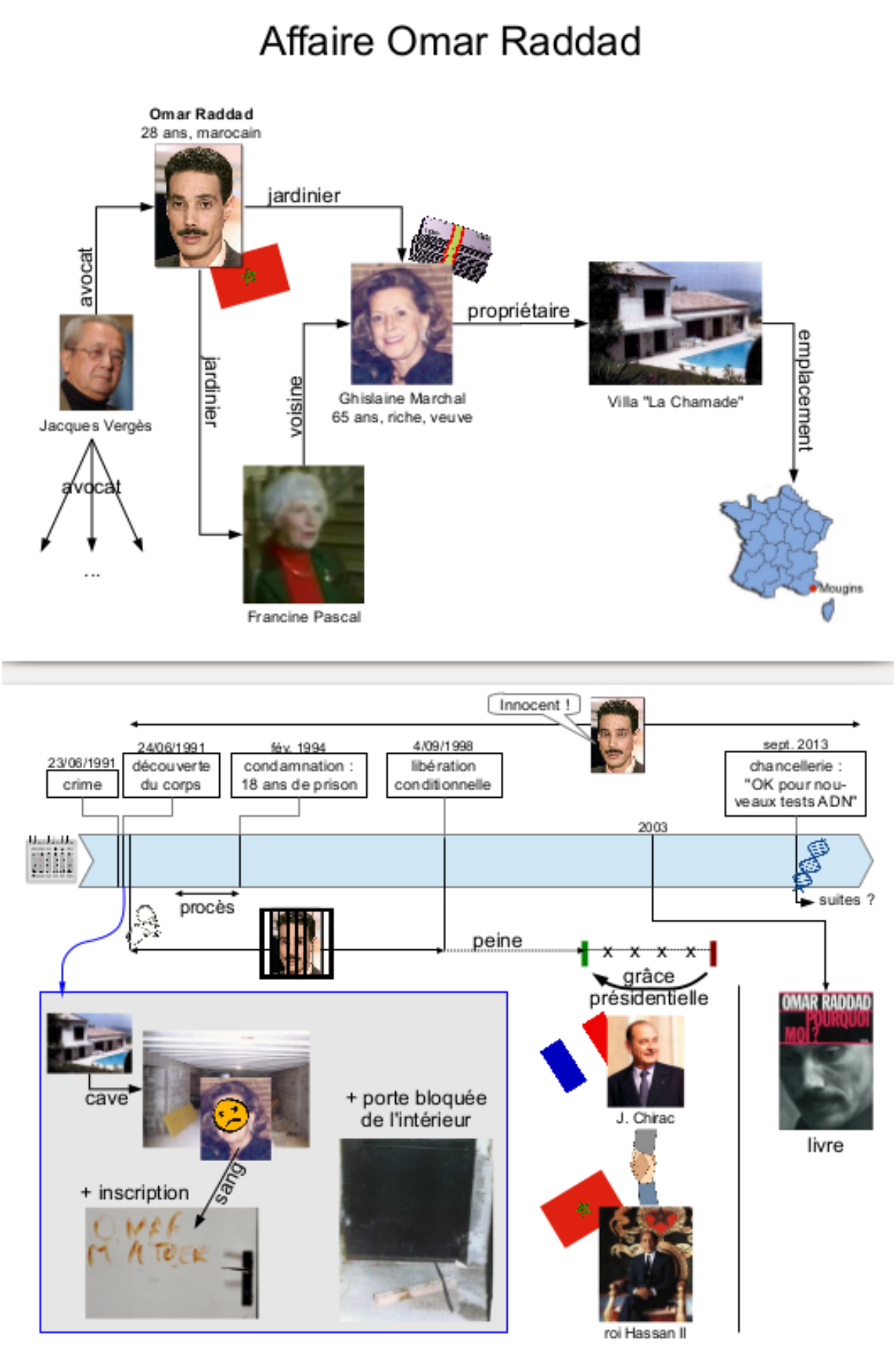

Figure 2: Two-page elicitation material for the “Omar Raddad” task

5. Capture and evaluation

We have conducted five sessions using the protocol and material above, each with a different informant. We included them as two tasks of the more general mocap[8] corpus collected last year (El-Fatah Benchiheub et al., 2016). In those sessions, a single face view was collected, including mocap markers on the face for the benefit of the other target studies (fig. 3). The data is still readable for traditional video observation.

Figure 3: Snapshot of the collected data.

This effort came as a first test before deploying it for more experiments, and possibly more lengthy elicitation yet, if evaluated positively. This section gives a few figures on the collected data, and evaluates the advantage of the protocol in terms of disfluencies per minute of resulting discourse.

Table 1 summarises the number of disfluencies found in a subset of the data collected using our protocol, per type and per session task. The "breaks" counted are the intentional (signalled) pauses taken by the informant. The disfluencies are counted and categorised in the following types: hesitations (H), filler/thinking gestures (F), reparation backtracks (B), interrputed/corrected signs (I) and superfluous repetitions (R).

Table 2 normalises the counts per minute of signing, giving both values including (*t.d.*: total disfluencies) and excluding (*n.b.d.*: non-break disfluencies) the called out breaks. The former gives a probable minimal rate of expected disfluencies in prepared discourse and will be compared to disfluency counts in non-paused discourse. The latter gives a lower rate, which quantifies the confidence in the data *between* the intentional breaks, subsequent studies now being able to call those out.

For comparison, we used some of the LS-COLIN data (non-chronological tasks), which we have full access to. Table 3 reports on the same counts in three sessions where informants explain the switch to the Euro currency in France, and their experience of it. The last column is the normalised number of disfluencies per minute of signed discourse.

[8]Motion capture.

Session	Duration	Breaks	H	F	B	I	R
Omar–S2	3 min 43 s	3	0	0	1	1	3
Omar–S3	3 min 15 s	4	4	0	1	4	2
Omar–S4	1 min 58 s	1	1	1	0	0	1
Titanic–S2	4 min 17 s	2	1	0	1	0	3

Average signing time without an intentional break: 56.6 s

Table 1: Disfluency and intentional break count per type in a collected data sample

Session	t.d./min	n.b.d./min
Omar–S2	2.15	1.35
Omar–S3	4.62	3.38
Omar–S4	2.03	1.52
Titanic–S2	1.63	1.17

Table 2: Normalised counts per minute of signing of the data analysed in table 1

Session	Duration	H	F	B	I	R	d./min
Euro–La	1 min 35 s	2	1	4	2	8	10.74
Euro–Kh	1 min	1	1	1	0	5	8
Euro–Ch	55 s	10	0	0	2	3	16.36

Table 3: Disfluency count per type and normalised counts per minute of signing in an LS-COLIN data sample

The disfluency rate here is over 11.7 per minute, as opposed to 2.6 when allowing intentional breaks, which is over four times less in density. When calling out the breaks, the rate is 1.855, i.e. 6.3 times less.

Admittedly, the elicitation material was different, thus the comparison bears some approximation. But apart from the material itself, the only major difference in the protocol used was the absence of an explicit permission to take breaks. We therefore believe the comparison to have relevance, especially as none of the observed ranges overlap.

This suggests that the proposed protocol significantly reduces disfluency in the resulting data, while still preserving it from distortion by external visual cues at the time of capture. Besides, the data pieces that are known not to deserve full trust in relevance are called out, which increases the observer's confidence in the remaining surface forms.

6. Conclusion

We have explained why it is difficult to collect long discourse that is both clean (undistorted) and prepared (constructed and flow intact) in Sign Language. With common studio elicitation techniques, captured discourse usually does away with one of those properties. We have proposed and tested an elicitation protocol which aims at ensuring most of both, and demonstrated some improvement in the produced data.

We have also produced elicitation material for two different tasks, based on two different strategies. One is derived from the translator's technique of deverbalising texts into diagram linking all elements of the source meaning; the

other remains closer to the more common technique of picture story elicitation.

Future work should include an evaluation of the respective impact of those strategies on the resulting data. We hypothesise that a trade-off exists here as well, this time between bias on discourse ordering (by picture stories, as noted in §4.2.) and bias on signing space usage (by deverbalised diagrams). Diagrams indeed tend to be mapped into signing space with little rearrangement by informants when they immediately make sense to them.

7. Bibliographical References

Crasborn, O. and Zwitserlood, I. (2008). The corpus NGT: an online corpus for professionals and laymen. In *3rd Workshop on the Representation and Processing of Sign Languages*, Marrakesh, Marocco.

Cuxac, C., Fusellier, I., Monteillard, N., Sallandre, M.-A., Jirou, G., Risler, A., Lejeune, F., Braffort, A., Choisier, A., Collet, C., Gherbi, R., Dalle, P., Jausions, G., and Lenseigne, B. (2002). LS-COLIN: Rapport de fin de recherche (final project report).

El-Fatah Benchiheub, M., Berret, B., and Braffort, A. (2016). Collecting and analysing a motion-capture corpus of french sign language. In *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*.

Filhol, M. and Tannier, X. (2014). Construction of a french–LSF corpus. In *Language resource and evaluation conference (LREC), Building and Using Comparable Corpora*.

Hanke, T., Hong, S.-E., König, S., Langer, G., Nishio, R., and Rathmann, C. (2010). Designing elicitation stimuli and tasks for the DGS corpus project. In *Theoretical Issues in Sign Language Research (TISLR conference)*, Purdue University, Indiana, USA.

Johnston, T. and Schembri, A., (2006). *Sustainable data from digital fieldwork*, chapter Issues in the creation of a digital archive of a signed language, pages 7–16. University of Sydney Press.

Johnston, T. (2008). *The Auslan Archive and Corpus. The Endangered Languages Archive*. University of London.

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

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation methods in the DGS (german sign language) corpus project. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta.

Schembri, A. (2008). British sign language corpus project: Open access archives and the observer's paradox. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, Marrakesh, Marocco.

Seleskovitch, D. and Lederer, M. (1985). Interpréter pour traduire. *L'Information Grammaticale*, 25:44–47.