Text–sign parallel corpus study to start designing an automatic translation system

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

Line Patris
Université Paris Diderot
F-75205 Paris cedex 13
magali.patris@etu.univ-paris-diderot.fr

Pierre Guitteny
"Signe"
Sign Language interpretation
Bordeaux, France
pierre.guitteny@gmail.com

ABSTRACT

This paper presents a new project whose goal is to design an automatic system to translate from French text to Sign Language, using a symbolic approach. After stating two essential properties of Sign Language that makes such a system different from text-to-text systems in terms of internal representational models, we present the 2006 Websourd AFP news corpus we chose for our design process. It is a parallel corpus consisting of journalistic texts in French and their video translations in Sign Language. Then we present our methodology, based on separate analyses of video description and text annotation first, and a comparison second. The idea is to annotate the entities in the texts thought to trigger some recurrent signed structures, and as a start we focused on three structures emerging from the video corpus observation: comparisons, oppositions and geographic localisations. Inspired by Guitteny's work on how to organise the signing space in an interpreting situation [7], they were chosen because they all strongly involve use of signing space, an essential notion in Sign Language with no equivalent in a written text. Using the highly-abstract model AZee [4] for representation of Sign Language rules, the ultimate goal is to build a set of translation mechanisms from annotated text to AZee operations, usable as input to a virtual signer animation system. Prospects are given to enrole theoretical frameworks capable of describing rhetorical/discourse structure representation.

Categories and Subject Descriptors

I.2.7 [**Artificial intelligence**]: Natural language processing—*Language models, Machine translation*

General Terms

Translation

Keywords

Automatic translation, Sign Language, corpus annotation

1. INTRODUCTION AND CONTEXT

Automatic translation has been a challenge for Natural Language Processing ever since the field existed, and it is known to be a most challenging scientific issue. Public translation systems like one can find online usually involve a choice in the source and target languages, but only written languages are made available. We now want to consider Sign Language (SL), in particular French Sign Language (LSF), as the target language. An example application would be for a SL user reading a textual webpage to select part of the text and run an applet program that would display an animated virtual signer producing the Sign Language for the selected text.

To translate from a language to another through an automatic process, some form of model is needed for each of the languages. Then an extra module is needed to bridge between the two. The ViSiCAST project about a decade ago did work on such a pipeline from text to sign [8]. It used HPSG, a unification grammar based on syntactic relations in text. Sign Language being far from the linearity of a textual input, we propose to use an approach to it and translation to it that fit its specificities.

This paper presents our start on a new translation project, dedicated to Sign Language as target. We want the system built bottom-up from data to enforce fully-acceptable output with no text-to-text bias. The next section presents two SL-specific properties, which will guide us on our way to it. Then, we introduce the corpus we based our preliminary work on and from which we will derive our first ideas and conclusions. After commenting on the choice for symbolic output of the translation module, the paper describes our methodology for the study of the corpus. We conclude with short- and long-term objectives for the future of the project.

2. SIGN LANGUAGE SPECIFICITIES

The task we address here is that of a translation system between languages using different channels: the source language is linear text; the target is gestural thus multi-linear. This leads to many differences of properties between the two, and we present two of them here.

2.1 Use of signing space

In sign linguistics, it is universally observed that sign languages make a relevant and consistent use of space, assigning locations of the signing space to parts of the signed productions. Lexical units (signs) can be relocated in space for particular reasons, and full clauses can also be signed leaning towards one side or the other.

In terms of implementation, the "lexical agreement" problem has been addressed more than once [8, 9], using syntactic relations and unifications systems. All of these are based on the assumption that a sign production is a sequence of lexical signs in the first place, whose order is syntactically constrained. Therefore, relocation is possible but limited to the cases where the reason for it is syntactic, e.g. to direct a verb.

The point in this paper is to look at spatialisation in the general sense, including the majority of cases where it is not analoguous to lexical agreement.

2.2 Discourse-grain observations or you lose

Another property of Sign Language—which is arguably a more or less direct consequence of 2.1—is its unique preferred order for the clauses of a discourse. In Sign Language, it is more acceptable [7]:

- to sign any element of context first (location, time...) and sign anything that takes place in that context afterwards;
- to sign the fixed and immutable things first in a setup (ground, houses...) and sign any animated object moving in it afterwards;
- to sign anything that identifies a topic or a target first, and sign the "hot news" or point at the target last.

For example, "Bob is playing with his sister in the garden" will typically be translated with the garden first (place) and the verb last (action and point of the sentence). This is unless the actual information is that the children are in the garden, as when phrased "It is in the garden that Bob and his sister are playing"; then the sign [garden] will come in last with a marked form.

Now this SL property, illustrated with lexical ordering in a sentence here, holds even for discourse-level ordering of sentences/clauses:

- everything participating in the setup is always signed before the action;
- chronological order is preserved when narrating a sequence of events—no lexical before/after mix;
- cause canonically preceeds consequence;
- etc.

While written language authors change orders and use meaningful prepositions to direct the link between two clauses, it is close to illegal or very misleading in Sign Language to do the same. We will see that this is consistent throughout our whole corpus study hereafter.

The reason we bring up this issue here is to emphasise on how much SL discourse structure differs from that of text, and to address the consequent fact that translating from one to the other cannot be carried out in a sequential fashion following the flow of text. Word-to-word translation is generally known to be of very poor quality; it is likewise for sentence-by-sentence translation of a text into Sign Language. In practice for instance, Sign Language interpreters would always rather be given summaries of the talks they interpret, and the overall quality of a resulting interpreted talk does vary according to how much is known of the talk beforehand [7].

Therefore, we claim that an automatic translation system should try and approach a text on that global level to carry the whole meaning and logic of a text into sign, knowing that clause order may change.

3. CORPUS-DRIVEN METHODOLOGY

To design a symbolic translation system from text, we need to parse the input into a symbolic representation of it. But as we have said, we want to place our study on the discourse level first rather than on the sentence level. Also, our approach is corpus-driven, but not based on machine learning techniques. Indeed, they usually need a linear signal and a tremendous amount of data to be trained, two premises which available SL corpora do not yet fulfill. This section presents and justifies the parallel corpus on which we base the rest of the work we report here.

3.1 The corpus

The corpus we use for this study is a parallel corpus of 1,000 French news item summaries, translated into LSF by professionals [12]. For each item we have a text comprising a 1-sentence title and a paragraph-long news content, together with their translated equivalents in a sign video (title + content). It is important to note, in relation with section 2.2, that the signed version is never a sentence-by-sentence translation, and that the whole text is always digested into a fully-acceptable SL production.

For our purpose, there are four advantages to this corpus which we explain hereafter.

Corpus genre.

By design, the corpus holds a series of texts with comparable length and style, which makes the data homogeneous, and its informational purpose is consistent. However, the vocabulary it not controlled or limited to any specific field. These two properties make a good tradeoff between convenience of the data to work with and potential for scalability.

Clean data.

Journalistic information is always written with the purpose of concision and clarity. None of the news content is bloat or misleading, which is clearly a plus. Besides, it is to note that unlike unverified web content, there are virtually no syntactic or spelling errors, which avoids relying on too much robustness from any text parsing system.

News content.

Of course, news reports almost systematically mention several people or place names, as well as events and clear relationships between them. This type of situation is known to be prone to spatial relationships in the signing space.

Local business.

Finally, we also appreciated that much of the recent years' Natural Language Processing research in our lab had produced systems trained on the huge AFP news agency feed. We could work with local tools kindly made available to us, highly trained for our genre of input texts.

3.2 AZee as output

During the past few years, LIMSI has been developing abstract sign description models capable of specifying various

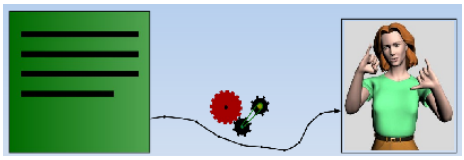

Figure 1: Sign Language animation from formal description input

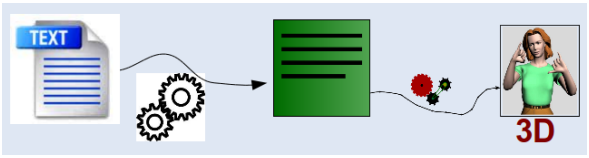

Figure 2: AZee input to a virtual signer, and output from the text translation system

SL features, including:

- the *Zebedee* model for lexical signs including their geometric variability, in particular the dependency of their "location" on signing context [3];
- the *Azalee* model for linguistic structures and the way they involve and synchronise the different articulators of the body [4].

AZee is intended to be a general description language to be used as input to virtual signers in the future, combining Azalee and Zebedee, thereby capturing all of these features. In figure 1, AZee is the format of the input represented on the left-hand side, and the gears represent the animation system needed to produce the signed output on the right-hand side.

Now if we consider this part of the work available, a good question in view of our full symbolic translation system is whether this input to the avatar animator can be used as output of the translator. In other words, could AZee be the interface between those two processes of the pipeline? See fig. 2.

Factoring out the invariants of a linguistic structure and parameterizing the variables into a dedicated AZee operator builds a production rule which we call "azop". An azop is named with a short explanatory description of the operation performed, and reused for any instance form. For example and as has already been reported in the DictaSign project [5], an "enumeration" operator can be written and applied to a list of items to enumerate, which implements a synchronised head movement as invariant and can be used with any list of things to enumerate.

AZee allows to build production rules on very different levels. A rule can be created to:

- add a facial expression to a single lexical item in order to add some form of adjectival meaning to it, e.g. a rule "adj-big" on a lexical sign that will typically generate a cheek puff on the sign [bear] to sign "a big bear";
- constrain the order of two (or more) full clauses if a structure is found to govern that order, e.g. a rule "setup" applying to a situational context and an event taking place in that context, that will make an avatar sign them in that order as it is preferred in SL syntax;
- even, on a much higher level, add discourse markers to punctuate full outline sections, e.g. a rule to generate automatic transition sentences like "now let's talk about", with the adequate pauses, eye blinks, etc.

AZee rules can represent phonetic details, lexical precedence, semantic and rhetorical discourse operations... There is no separation in these levels; they can all be dealt with using the same formalism. One goal of AZee is actually to build a "semantic grammar", to govern surface productions from the intended meaning, which discards syntax (in the linear ordering sense) from the top position of structuring elements. Also, we have seen that a coarse-grain level of discourse study was the preferred working level when dealing with translation to Sign Language. We are led to hypothesise that AZee is a good candidate to serve in our symbolic translation system. It takes the production rule system closer to the semantics of the discourse as a whole, which is what we want in an ideal output for such a system.

4. METHODOLOGY

Our methodology is based on a separate observation of text and video. This section describes what was done on each side of the corpus, and brings up a comparison task afterwards. Working from the videos first guarantees that we always work on the texts knowing what we are looking for and why, regarding the signed production, we are looking for it. Separating the steps and not studying the two languages side-by-side guarantees that there is no word-to-word or sentence-to-sentence bias in our study, i.e. that we do concentrate on a linguistic system alone regardless of the task, and that we keep regarding meaning as the ultimate element to translate.

4.1 Video analysis

The analysis of the video part of the corpus, before any kind of comparison with the text, is made in collaboration with a professional Sign Language interpreter. Its process is somewhat empirical as there are no tools able to serve as counterpart to the ones used on text, like syntactic parsers or text scripts able to locate and tag a given sequence or pattern of lexical units. However, it is bound to stabilise with time and we intend to formalise it as we go.

The first step we took is ask the interpreter to write a free-hand description regarding information order in the video and everything regarding the signing space. This study, in relation to section 2.2, takes place on the discourse level as the interpreter describes the signing based on a segmentation in coarse-grain "chunks", which can be:

- a situational context, in which the following signed chunks will take place;
- a time/place setup for an event;
- the description of an event, person or place or any entity or group which will be referred to afterwards;
- the main point of the discourse, usually observed last.

The list is non-exhaustive and built on the fly. The chunking taking place here relies more on the intuition of where a new purpose of a type above starts or ends, but SL being an oral language, and similarly to any vocal-oral corpus, the notion of syntactic sentence becomes blurred. Then we draw

graphs to represent links between the chunks, using every chunk as potential node, preferably written top to bottom to keep track of signing order. When they are assigned a specific location they are tagged with the linguistic reason for that spatialisation, and when two chunks are semantically bound, we draw edges between the nodes and tag the edge with identification of the link such as "cause", "opposition", "precedence", "context of". Once again, this is not yet claimed to be a formal framework, but an exploratory means of using a Sign Language professional's intuition to search for regularities and indeed quite a few emerged:

- every pair of chunks linked by an edge tagged "opposition" has two different locations, most the time followed by a verb "agreeing" with one or both locations (see example below);
- probably a particular case of the above, the same applies to comparisons between two entitites;
- every sequence of signs involving several geographical places locates one with reference to the previous—usually up to two or three and typically country, then area/region, and/or finally town;
- every video ends with the clause carrying the actual news, as expected...

Figure 3 shows snapshots of the clear opposition in the news item reading:

> With two cowboys, the beautiful country of the American West and a love story, the film *Brokeback Mountain* combines all the ingredients of the typical western film. However, they are arranged in a way John Wayne would surely not have appreciated.

In 3.a, we see a snapshot of the first half, where everything describing the film is signed to the signer's left. In 3.b, the signer is referring to John Wayne and we see that his body has turned to his right. Finally, in the last two seconds of the video, the signer has turned back to face the camera and signs "we can be sure" / [not like] / flat hand demonstrative to the left.

The point of this was to know where to start with annotating the texts, i.e. parsing and tagging the texts knowing what to look for. It is the purpose of the next section.

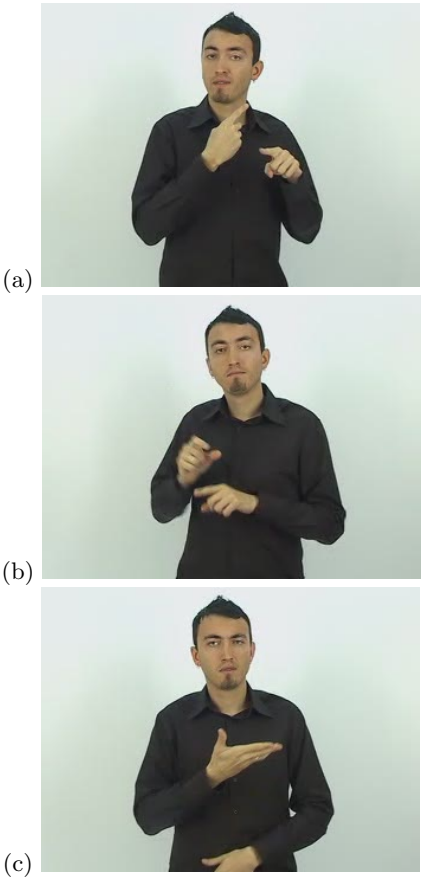

Figure 3: Example of opposition structure observed in the corpus

4.2 Text annotation

Three interesting phenomena were selected in the study of the videos described above, now the next step to take is to examine the text corpus and annotate the text structures related to the semantics of the chosen signed structures. In this section this time, the whole job will be text-only. It specifies the way we annotate the French text corpus, regarding the three following items: opposition, comparison and structures we call "geographic localisation". Also, we describe the tools we used to carry out the annotation task, part of it automatically.

4.2.1 Annotation specification

Opposition is a self-explanatory concept, but can be very subjective thus quite hard to define an annotation guideline for, and it can occur between segments of various lengths,

from word to long sentence. This would enable us to analyse various situations but of course complicates the annotation process. To tackle this, we give an objective definition for a subset of opposition constructions, based on syntax: two entities are opposed when they are syntacticly conjoined through one of a set list of lexical opposition conjunctions.

Comparison is also a relation between two distinct entities, but there is a subtle difference in the way they are respectively treated: both opposed entities are at the same level, while in a comparison one is taken for the norm against which the other is measured. Observations also show that the segments involved in a comparison are rather shorter than those opposed, a simple word or syntagm on average. Comparison in French is structured as follows: "*a* est plus/moins/aussi *adj* que *b*" (in English: *a* is more/less/as *adj* than/as *b*). We chose to focus once more on this lexically marked (hence objective) pattern. To locate occurrences of this pattern in the texts, we used a semantic tagger called Wmatch, whose behaviour is described in the "tools" section below.

Geographic localisation is the last phenomenon we chose to annotate in the texts. It is interesting to include it in our study for two reasons, first because it seems always to trigger a use of signing space in a very precise way, and second because it is frequent in journalistic texts. Indeed, it consists in selecting a smaller portion of a given location with the use of cardinal directions, as we see in "au sud de Gaza" (south of Gaza), or "le nord de l'Italie" (the northern part of Italy). It can be extended with the name of the geographical entity being located, which is usually named first, then located with reference to the former. These patterns can also be tagged using Wmatch.

Here is an example of our XML-style annotation for geographic localisation in a text (the tags' names are still subject to modifications):

```
<_loc>
  <_pt_cardinal> sud </_pt_cardinal>
  <_prep> de </_prep>
  <_det> la </_det>
  <_loc>
    <_province>
      <Gaza> bande de Gaza </Gaza>
    </_province>
  </_loc>
</_loc>
```

4.2.2 *Tools*

In the LIMSI–CNRS environment, several software tools are used daily for parsing and editing text. We present three of them in this section, two of which we use in our study, while the third provides interesting extension prospects for our annotation processes.

The latter is named XIP, which stands for Xerox© Incremental Parser, a syntactic parser able to extract dependencies, named entities and semantic roles [1]. We had first decided to use the dependencies to annotate oppositions, because the conjunctions were well tagged and considered as connectors, entering a binary dependency with the second verb of the sentence. The semantic parser, Wmatch described below, was not so good at finding them, but this was remedied by a new set of rules added to its grammar for our purposes, so we were able to put XIP aside with no loss of efficiency and a gain in the homogeneity of our set of XML

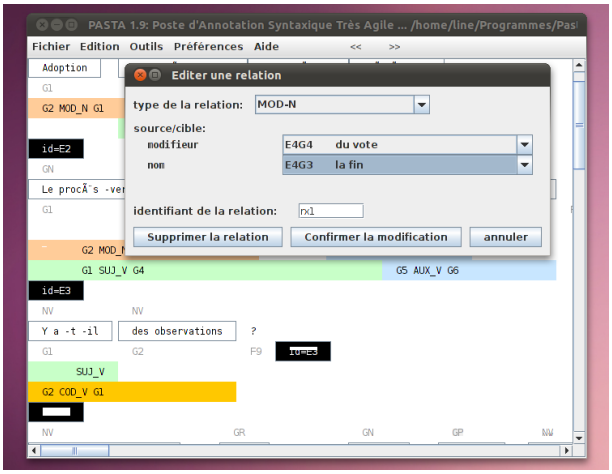

Figure 4: Annotating a relation with the PASTA program

tags. Still, XIP is not definitively removed from our panel of tools, because its output information on syntactic node dependencies will be useful in a later step of our project.

Wmatch [6] is a semantic tagger, and as such functions only at a local level. It is therefore unable to annotate far-reaching relations as XIP does, but local focus makes for greater precision at a local level, which is what we needed in the first steps of our project. It detects perfectly the patterns of comparison, annotating the adverb and the adjective under a distinctive tag, as well as the cardinal directions and locations associated to them.

PASTA, for Poste d'Annotations Syntaxiques Très Agile (very agile syntactic annotation post) is a software for annotating XML data with syntactic relations, originally designed for the Passage project [11]. In Passage, the types were used for syntagm type labels such as NPs, VPs, etc. and the relations for linking syntactic relations, for example a verb and its subject. A screenshot is given in figure 4, showing a dialog box where the user is prompted to fill in two slots of a relation. We chose to use this tool because it enables us to annotate both types and relations in a given text. Moreover, PASTA does not discriminate between a word or a group of words as source and/or target of a relation. With some modifications to its list of types and relations, PASTA enables us to annotate as groups the entities being opposed or compared (we did not specify their nature any more for the moment, but it may be done later as a refinement in our annotations). The relations highlight the connections between the two entities being compared or opposed and their connector, the cardinal direction and the place located with it, or the name of a geographic entity and its precise location. For that matter, our progress is as follows: automated annotations with semantic tagging are completed—and in the next part we will describe how we obtained them—; we are currently carrying out the manual pass of the annotations.

4.2.3 *Methodology and tagging conventions*

Once we had narrowed our project down to the three phenomenons described in the previous section, and isolated their characteristics, the task was a simple matter of determining the input and output formats, and programming the

necessary modules to obtain the desired effects. We made our XML tagging conventions quite simple, and mostly followed Wmatch's tagging style, for the sake of consistancy.

We first tagged the texts with Wmatch to obtain the first version of our XML files, which was done using the native grammar of Wmatch [10]. The next step was to write a grammar extension to include new rules for opposition conjunctions, and adjust the comparison conjunctions to our writing conventions. We then applied this new grammar to the already-tagged files to produce our working version. In this version, we have kept all the tags supplied by Wmatch, as they may be useful when we need to be more precise in our analysis, and our tags highlighting the opposition and comparison conjunctions. We will then use PASTA manually to annotate the relations between the different components of our three structures. At the end of this task, we will fully have annotated the XML files with the desired information.

4.3 Comparison

Once we have annotated the text we can compare it to the videos, looking at both following cases:

- a structure is annotated in the text and a spatialisation of the corresponding entities takes place in the video; these are cases to validate the systematic link between the annotated lexical structures and a use of space in SL;
- a structure is annotated but the equivalent sign production falls out of the pattern above; these cases will be interesting to comment on, perhaps to refine some of the categories (e.g. more constraint on the type of comparison: comparing to one vs. comparing to many, etc.).

For example, we already see a very clear rule for geographic localisations. Let us look at the example given in section 4.2.1 again, with the semantically tagged structure for "south of Gaza". There are three important tags in this hierarchical structure: the two `_loc` tags—the first holding the whole structure and the second nested in the first—and the cardinal direction tag `_pt_cardinal`. The geographic localisation structure is the whole group: the first encapsulating `_loc`. When translated into Sign Language, every structure using this template consistently raises the same signed structure, i.e. in order:

1. a sign is performed for the second '`_loc`'-tagged item;
2. a zone is activated in space with a circular movement of a hand on a vertical plane and a quick look of eye gaze;
3. a subpart of that zone is designated (e.g. pointing sign) according to the `_pt_cardinal` term.

Now for each XML tag structure that we will obtain once the whole corpus is annotated with relations through PASTA, we hope to find such translation rules, and start validating our corpus-driven approach to translation.

5. OBJECTIVES AND CONCLUSION

In short, this paper has presented a methodology to derive a first set of symbolic translation rules from text to sign, specifically looking at the spatialisation issue. Using a parallel corpus of text-to-sign translation, a first study was made of the videos (disregarding the texts) to find relevant reasons for sign and sentence relocation. Three first clear categories emerged: comparisons, oppositions, and geographic localisations. Scripts were run and tools used to tag the texts where lexically marked instances of these structures could be found, this time disregarding the videos. Then by looking at the two separate annotations, we will state whether systematic rules can be derived to translate any of those three structures.

A short term objective with this study is to find systematic rules governing use of space in LSF, when translating comparisons, oppositions and geographical localisations, in the case of textual constructs that are possible to pick up automatically. By the date of the conference, we should be able to produce the first findings on this, as the text annotation will be finished and compared to the video descriptions.

In a longer run, if we first succeed in formalising invariants on these sample structures, we will continue with more structures. There are other reasons for spatialising a sign sequence, and there are other features to look at and capture in production rules. Clause and statement order, for which we give a prospect below, is interesting to look at too. It is hoped that with this methodology, a full translation architecture with AZee support can be specified and implemented as the interface between the text and a Sign Language animation module.

To tackle clause/sentence order in Sign Language, we need to look at the rhetorical function of each counterpart in text, as a lot of reordering seems to take place to avoid linking lexical tokens like text equivalents "therefore", "because", etc. We intend to initiate collaborations with experts on rhetorical structure or formal discourse representation theories like RST or SDRT [2], which formalises the way in which parts of a discourse link to one another. While it can still not be in reach to annotate this automatically, it will be interesting to work with a manual annotation of the discourse parts with this type of theory and see what emerges from a comparison of those to the sign chunks observed in the videos.

6. ACKNOWLEDGEMENT

We would like to thank Sophie Rosset for helping with the semantic tagging of our text corpus and use of Wmatch, and Patrick Paroubek for the time he spent adapting the PASTA program to our needs.

7. REFERENCES

[1] S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. Robustness beyond shallowness: Incremental deep parsing. *Natural Language Engineering*, 8:121–144, 2002.

[2] N. Asher and A. Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

[3] M. Filhol. Internal report on zebedee. Technical Report 2009-08, LIMSI-CNRS, 2009.

[4] M. Filhol. A combination of two synchronisation methods to formalise sign language animation. In *Proceedings of the 9th international Gesture Workshop*, 2011.

[5] M. Filhol, A. Braffot, and S. Matthes. Dictasign deliverable d4.2: Sentence descriptions for bsl, dgs, gsl, lsf, 2010. DictaSign project, EU-FP7.

[6] O. Galibert. *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris Sud, 2009.

[7] P. Guitteny. Langue des signes et schémas. *Traitement Automatique de Langue (TAL)*, 48, 2007.

[8] T. Hanke et al. Visicast deliverable d5-1: interface definitions, 2002. ViSiCAST project report.

[9] L. Kervajan, E. Guimier De Neef, and J. Véronis. French sign language processing: verb agreement. In *Gesture in Human-Computer Interaction and simulation, LNCS/LNAI*, volume 3881. Springer, 2006.

[10] S. Rosset, O. Galibert, G. Bernard, E. Bilinski, and G. Adda. The limsi participation to the qast track. In *Working Notes of CLEF workshop*, 2008.

[11] A. Vilnat, P. Paroubek, E. Villemonte de la Clergerie, G. Francopoulo, and M.-L. Guénot. Passage syntactic representation: a minimal common ground for evaluation. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2478–2485, Valletta, Malta, 05 2010. European Language Resources Association (ELRA).

[12] Websourd. L'actualité en bref et en lsf de l'année, 2006. DVD.