Using sign language corpora as bilingual corpora for data mining: Contrastive linguistics and computer-assisted annotation

Meurant, Laurence; Cleve, Anthony; Crasborn, Onno

Published in:
Proceedings of the 7th workshop on the Representation and Processing of Sign Languages:Corpus Mining

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Using Sign Language Corpora as Bilingual Corpora for Data Mining: Contrastive Linguistics and Computer-Assisted Annotation

Laurence Meurant[1], Anthony Cleve[2], Onno Crasborn[3]
[1]F.R.S. - FNRS and University of Namur, [2]University of Namur, [3]Radboud University Nijmegen
[1,2]61, rue de Bruxelles, B-5000 Namur, Belgium; [3]PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
laurence.meurant@unamur.be, anthony.cleve@unamur.be, o.crasborn@let.ru.nl

Abstract

More and more sign languages nowadays are now documented by large scale digital corpora. But exploiting sign language (SL) corpus data remains subject to the time consuming and expensive manual task of annotating. In this paper, we present an ongoing research that aims at testing a new approach to better mine SL data. It relies on the methodology of corpus-based contrastive linguistics, exploiting SL corpora as bilingual corpora. We present and illustrate the main improvements we foresee in developing such an approach: downstream, for the benefit of the linguistic description and the bilingual (signed - spoken) competence of teachers, learners and the users; and upstream, in order to enable the automatisation of the annotation process of sign language data. We also describe the methodology we are using to develop a concordancer able to turn SL corpora into searchable translation corpora, and to derive from it a tool support to annotation.

Keywords: Sign Languages, parallel corpora, annotation automatization

1. Introduction

As more and more sign languages nowadays, the French Belgian Sign Language (LSFB) is now documented by a large scale digital corpus (Meurant, 2015): the Corpus LSFB. This dataset includes around 150 hours of multi-camera recorded data, from which 12 hours are so far annotated with ID-Glosses (Johnston, 2010) (104,000 tokens, from which 98,200 fully lexicalized signs), and 2.5 hours translated into written French (2,400 sentences) and is supplemented by the metadata about the participants and the tasks. An online lexical database contains all the sign types glossed up to now, and serves as a dynamic external controlled vocabulary for the annotation process in ELAN. These data are made available online via a user-friendly web site. The French counterpart of the Corpus LSFB is now being collected: in the same setting and following the same protocol, pairs of French speaking informants are currently videorecorded. The collected data will be transcribed and translated into LSFB. When this work will be completed, we will for the first time benefit of a bidirectional translation corpus between a sign language (SL) and a spoken language (SpL). The Corpus NGT (Crasborn et al., 2008) has been an inspiring model for the Corpus LSFB. It includes NGT video data (72 hours), gloss annotations (150,000 tokens, 3,300 types), sentence-level translations into written Dutch (15 hours, 15,000 sentences), and the lexical database NGT Signbank (including translation equivalents and a detailed phonological description).

This kind of resource is not only essential to the linguistic description of sign languages, but it is also a potential wealth of information for pedagogic purposes, for the field of translation and interpretation studies and for the field of contrastive linguistics between signed and spoken languages. Exploiting corpus data remains subject to the time-consuming and expensive manual task of annotating, i.e. from the ID-Glossing to the analytic annotations. This slow process is unavoidable at this stage and crucial in enlarging the available data set that is needed to automate the annotation process in the near future.

In this paper, we present ongoing research that aims at testing a new approach to exploit SL data. This approach relies on the methodology of corpus-based contrastive linguistics. It exploits the fact that many SL corpora (including the LSFB and the NGT ones) do not only consist of video recordings of SL, but also glosses and translations into spoken language (SpL). In other words, our approach considers SL corpora as bilingual corpora. We present and illustrate the main improvements we foresee in developing the use of sign language corpora within a corpus-based contrastive methodology: downstream, for the benefit of the linguistic description and the bilingual (signed - spoken) competence of teachers, learners and the users; and upstream, in order to enable the automatisation of the annotation process of sign language data.

We first (Section 2.) present the major types of multilingual corpora used for the purpose of contrastive linguistics. Then we explain why SL data can be considered as translation corpora and we show how valuable the combination of translation corpora and comparable corpora would be in the fields of SL linguistics and of SL-SpL contrastive linguistics. We provide an overview of the possible uses of such bilingual data, not only to the benefit of linguists, but also of interpreters, translators, teachers and learners. In Section 3., we show the modelling of the different data sources we are using for the development of a concordancer between SL and SpL data, and we specifically detail the way they interact with one another. Section 4. describes the methodology we are using in order to develop the concordancer, from the challenging alignment of written texts and video recorded productions at the level of the word-sign, to the extraction of semantic equivalents in context, and the way these development are expected to automatically assist the annotation process.

2. Corpus-based contrastive analysis and Sign language data mining

2.1. Multilingual corpora

Beyond the domain of SL linguistics, the computer revolution also impacted the domain of contrastive linguistics in general by having enabled the development of multilingual corpora. Multilingual corpora, combined with alignment and search tools, are today acknowledged for their theoretical as well as practical importance in cross-linguistic studies and applications: they provide a rich basis of language correspondences in context that are able to provide new insights into the languages that are being compared (Altenberg, B. and Granger, 2002; Johansson, 2007). Multilingual corpora are the basis of all multilingual concordancers such as TransSearch (Bourdaillet et al., 2010) or Linguee (Linguee, 2015). Following the terminology of Altenberg, B. and Granger (2002), we will distinguish between translation corpora and comparable corpora, although both sometimes fall under the heading of "parallel corpora" used in a generic sense .

Translation corpora consist of original texts in one language and their translations into one or several other languages. They are unidirectional when the translation goes only in one direction, from the original language A to the target language B. If the translation goes in both directions, that is if each language is both source and target language, they are said bidirectional. Some translation corpora are aligned, which means that each unit of the original text (it can be a paragraph, a sentence, a phrase, or even a word, in the case of written texts) is linked to its corresponding unit in the other language. Such aligned translation corpora are also called "parallel corpora"[1]. The Hansard Corpus is a well known example of parallel bidirectional translation corpus. It has been the first one to be digitized and made available to linguists. It consists of parallel texts in English and Canadian French, drawn from official records of the proceedings of the Canadian Parliament.

Comparable corpora are made of texts in original language only, i.e. non translated ones, that share the same type, subject matter and communicative function. The gathered texts may be restricted to a specific domain (e.g. newspaper articles about football in English and French) or on the contrary they may represent a wide range of text types (e.g. balancing general news with economical, legal, medical, and political texts).

Each kind of multilingual data has its advantages and disadvantages. When using a translation corpus, one can rely on the semantic similarity of the texts in both languages: the objective, the discursive function, the register as well as the audience is typically the same in both version of the texts. But translated texts may always be suspected to reflect the transfers of features from the source language to the target language, or "translationese" (Gellerstam, 1996), and individual variations specific to the translators. On the contrary, the texts contained in a comparable corpus reflect the natural use of language, but it is sometimes difficult to know whether the compared texts are really comparable, for example in terms of register or discursive function. Therefore, the combination of both types appears to be particularly relevant since it eliminates or mitigates the disadvantages and strengthens the advantages of each type.

[1]Henceforth, we will use 'parallel corpora' with this meaning.

2.2. Sign language corpora as translation corpora

Due to the visual-gestural nature of SLs, most modern SL machine-readable corpora like the Corpus LSFB and the Corpus NGT are bilingual ones, since the videotaped data are accompanied by the written glosses of the signs and by the translation of the videos in written language. But as far as we know, this bilingual property of SL corpora has not been exploited yet for the development of contrastive linguistics. However, we see contrastive corpus linguistics between a signed and a spoken language as an effective solution to the current difficulty to detect interesting data amongst the sign language corpora.

Most SL corpora can at least be seen as unidirectional translation corpora, which provide a good basis for comparing how a specific meaning (retrieved from the SpL translation, in our cases in French or Dutch) is rendered in SL (LSFB or NGT respectively). If combined with alignment at the level of the sign and word, it would be an efficient means to extract aligned bilingual examples of words and signs in context. For example, a request on the word *même* ('same') within the French translations of the Corpus LSFB would provide the various signs used by the signers and that were translated by *même* (AUSSI, COMMUN, EGAL, MEME-AVANCER, MEME.MAFIA, MEME.REPETITION, MEME.Y, STABLE, but also MOI-MEME, PERSONNE.MOI, VOIR.MOI which have a reflexive meaning ('self') translated by *même*[2]. These signs would be presented in their context of use, that means within the video clip where they appear, and aligned with the corresponding contexts of occurrence in French. And conversely, a request on the sign AUSSI ('also, same') will provide the various words and word constructions used to translate the various tokens of this sign into French (such as *comme* 'as', *disons* 'let's say', *un genre de* 'sort of', *aussi* 'also', *et puis* 'and then'), according to their context of use. This type of information can be harvested from corpora to enrich the current lexical databases of the Corpus LSFB and the Corpus NGT with a classification of the meanings of the signs in context and their frequencies, which in turn will be used to assist the annotation process (see Section 4.).

By comparing sequences of signs or words using a concordancer, it will also be possible to search for the translation equivalents of non-lexical elements, as for example the equivalents of the passive forms of French, or the French and Dutch equivalents of the spatial left vs. right oppositions in LSFB or NGT, the ways LSFB or NGT expresses what is translated by prepositions into French or Dutch, or the ways partly-lexicalized signs (Johnston and Schembri, 2010) of LSFB or NGT are translated into French and Dutch, respectively. In particular, research on discourse can

[2]The signs corresponding to these glosses written in capitals can be seen on the lexicon part of the LSFB corpus website (`http://http://www.corpus-lsfb.be/lexique.php`)

greatly benefit from this methodology. For example, requests on French or Dutch discourse markers will provide examples in LSFB and in NGT that will illustrate how diverse the SL expressions of the equivalents of these markers are: Do SLs use discourse markers as equivalents, or other lexical and/or non-lexical resources, articulated manually and/or non-manually?
In their present state, both the LSFB and the NGT corpora (blocks A and B in figure 1) are unidirectional (relations 1 and 2 in figure 1): the SL original productions are translated into written French and Dutch respectively. We are currently testing the feasibility of building the counterpart of the LSFB corpus (block C in the figure), which means videorecording spoken French data and translating them into LSFB. In this way, we will be able to count on a bidirectional translation corpus between a SL and a SpL.

2.3. Towards comparable corpora

In building the French counterpart of the LSFB corpus, our data gain various additional dimensions, and especially the possibility to compare original LSFB productions and original French ones (relations 4 in figure 1). The French data are elicited in the same conditions as were the LSFB data, and following the same protocol. The informants are invited by pairs in the LSFB-Lab studio. A French-speaking moderator is leading them through the same tasks as the one used for the LSFB corpus.
The content of the tasks were minimally adapted to fit to the hearing and Belgian French culture of the informants, but the dialogic setting as well as the discourse genre of each task have been preserved, which make the French and the LSFB productions closely comparable.
Together, the LSFB corpus and its French counterpart (B and C in figure 1) provides a rich variety of possible comparisons (referred by the numbers of the arrows in the figure):

1. Comparison of original discourses in one language and their translation in the other one (relations 2 and 3);
2. Comparison of original discourses in both languages (bidirectional relation 4);
3. Comparison of original and translated texts in the same language (relations 5 and 6).

These three types of corpus-based comparisons have an heuristic power in the sense they offer the opportunity to discover features of the languages in contrast that could not be expected without the automatic comparison of large amounts of parallel data (Altenberg, B. and Granger, 2002; Gilquin, 2000). It is the reason why we are testing the feasibility of building such combination of translation and comparable corpora and its efficiency for the issue of corpus mining.
When it will be possible to link these bilingual corpora to data from LSFB and French learners, the *Constrastive Interlanguage Analysis* method (Granger, 1996) could be used to better understand the specific difficulties of LSFB signers learning French and of French speakers learning LSFB.

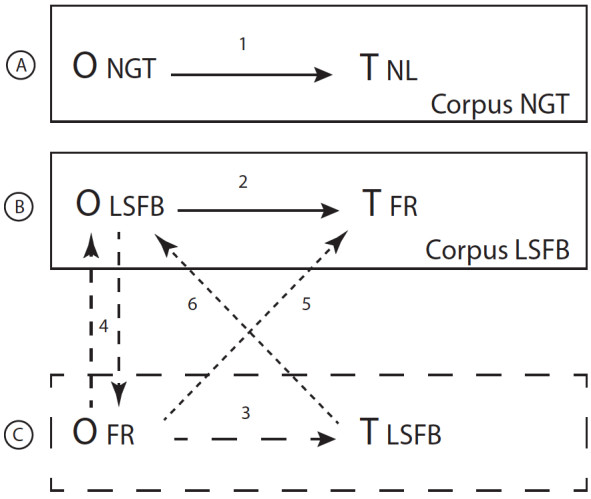

Figure 1: Corpora at our disposal (continuous line) and under construction (dotted line). O = original texts, T = translated texts. Both blocks A and B constitute translation corpora. Together, the original texts of B and C constitute comparable corpora.

2.4. Beyond the linguists' needs

Once these bilingual data gathered (the translation corpora or even the comparable ones) and the concordancer developed, linguists will take advantage of novel and effective means to detect interesting data within sign language corpora. But the resources that we are building for corpus-based contrastive analysis for the purpose of SL research also will benefit other kinds of users, ranging from interpreters, translation and interpreting trainers, and teachers for the deaf, to indeed all signing learners of French/Dutch and French/Dutch speaking learners of LSFB/NGT. A searchable database of aligned bilingual examples of language in use will constitute a useful resource for expanding one's knowledge of a second language and increasing one's level of bilingualism. It can be used to assist a wide range of tasks, among which the comprehension of LSFB/NGT or French/Dutch texts, the production of LSFB/NGT or French/Dutch texts, as well as the translation between LSFB/NGT and French/Dutch (in both directions).
When it comes especially to deaf pupils who learn a spoken language, these bilingual data can be seen as an efficient tool to support their learning and to foster their autonomy in the use of the spoken language, just as *TransSearch* or *Linguee* are supporting the speakers of one language who learn another one, at any level. For example, a deaf learner of French or Dutch may discover the variety of meaning of signs and above all learn to distinguish the various meanings with accuracy thanks to the signed equivalents at her/his disposal for each written example. She/he may also be helped in their use of some idiomatic features like gender of names, prepositions, *avoir/être* auxiliaries in French, etc.
Figure 2 shows a mock-up user interface of the tool that will be derived from the aligned bilingual data (in its LSFB-

French version). The terms between asterisks refer to the entities' names used in Figure 3.

3. Modelling the data resources

The combination of the Corpus LSFB and the Corpus NGT provides a relatively (i.e. for the present time) significant size of exploitable data which will be involved in the machine learning process underlying the development and the exploitation of our parallel concordancer.

Figure 3 provides a simplified "helicopter-view" of the various data artifacts that are involved in the creation of our multilingual and multimodal corpus-based concordancer. This model represents the main concepts involved in the data and tools, as well as their characteristics and relationships.

The figure shows the parallelism between the components available from the LSFB corpus (at the top), and the ones provided by the NGT corpus (at the bottom). The new data set under construction (in dotted line in Figure 1 and in grey in Figure 3) can be seen as a mirror of the existing Corpus LSFB: spoken French data (we foresee 40 hours of video), their transcription and translation into LSFB, as well as the annotation of the translations.

Each corpus consists of a set of videos (LSFB_VIDEO and NGT_VIDEO) where two signers achieved a task. Each video is identified by a unique ID, corresponding to its Unique Resource Identifier (uri), and is characterised by the duration of the video (Duration), and a brief description of the task (Task_Description).

Each corpus also includes a large set of signs (LSFB_SIGN and NGT_SIGN). For the LSFB corpus, this set of signs corresponds to the Lex-LSFB lexical database; for the NGT corpus it corresponds to the NGT Signbank. In both corpora each sign is characterised by a unique ID-gloss (ID_Gloss), and is linked to a set of keywords (FR_KEYWORD or NGT_KEYWORD) that represent (some of) the different possible meanings of the sign[3]. Note that the NGT Signbank also includes extra information about the signs, such as phonological descriptions, that are not depicted in Figure 3.

The occurrence of a given sign in a video is represented through an entity type SIGN_ANNOTATION. An annotation indicates the exact time period during which the sign appears in the video, in the form of a time interval (Begin and End). Note that when the same sign S occurs N times in the very same video V, there are N annotations linking S and V in the corpus, each with a distinct time interval. The annotation also records which of the two signers is the author of the sign, via attribute Turn.

As mentioned above, the corpus also provides, for a subset of the videos, the full French/Dutch translation (FR_TRANSLATION and NL_TRANSLATION) of the task. Each translation is made up of a set of translation fragments (FR_TRANSLATION_FRAGMENT and NL_TRANSLATION_FRAGMENT), that is a French/Dutch text fragment (Text) translating what is expressed in the respective sign language by one of the two signers (Turn) during time interval [Begin, End] of the video.

External tools and related data resources are also available. The CoBRA (Corpus Based Reading Assistant) tool (Deville et al., 2013) is based on bilingual corpora (Dutch-French and English-French) aligned at the level of the sentence, and allows the teachers to create labelled texts in Dutch (NL) or in English (EN) and French-speaking learners to be assisted in their reading by clicking on any word in order to know its meaning in its particular context of occurrence. CoBRA is based on a searchable concordancer, called the "Dico Corpus" tool, and on two bilingual dictionaries (FR-NL and FR-EN) called "DiCoBRA" that are (1) produced from a contrastive approach of the existing dictionaries of each language and (2) completed by the contrastive data provided by "Dico Corpus".

The CoBRA resources (CoBRA Corpus) currently include a global text corpus of over 30,000,000 words among which circa 15,000,000 French words, about 10,000,000 concordances (i.e. aligned bilingual examples), an English-French glossary of about 19,000 entries, and a Dutch-French glossary of about 20,000 entries. CoBRA's dictionary (DiCoBRA) includes circa 87,000 lemmas and 300,000 inflected forms of French.

4. Automated support for annotation

On the basis of the available data described above, we are currently developing a concordancer in order to exploit the LSFB and the NGT corpora as aligned (at the level of the sign and word) and searchable translation corpora (LSFB-French and NGT-Dutch). While doing this, we also aim at providing support to the annotation process of the lexical part of the signed data that have not been annotated to date. The methodology used is organised in three steps, the first one being in progress: building the alignment tool, extracting semantic equivalents from the annotated and translated data, and eventually developing the tool support for the annotation process.

4.1. Alignment

In order for the translation corpora to be exploited, they need to be aligned. This means that each unit from one language must be linked to its corresponding unit in the other language. Translation corpora of written texts can be aligned at the level of the paragraph, at the level of the sentence, or even at the level of the phrase or the word. The automatic alignment sentence by sentence is the most common. The matching between a sentence in the source language and a sentence in the target language is based on statistics exploiting information about typographical features (capitals and punctuation marks), the length of the sentence, and cognate words (Altenberg, B. and Granger (2002), p. 10).

In the case of SL data, the alignment cannot rely on any typographical feature. And, as has been extensively shown, the identification of sentences in SL remains a difficult task (Crasborn, O. (2007), Fenlon et al. (2007), Ormel and Crasborn (2012), Börstell et al. (2014), to name a few). Therefore, we decided to avoid investing in segmenting the data

[3]Within the NGT data, those meanings in context are retrieved from a specific tear from the annotations files where a word-level translation equivalent is created for every sign.

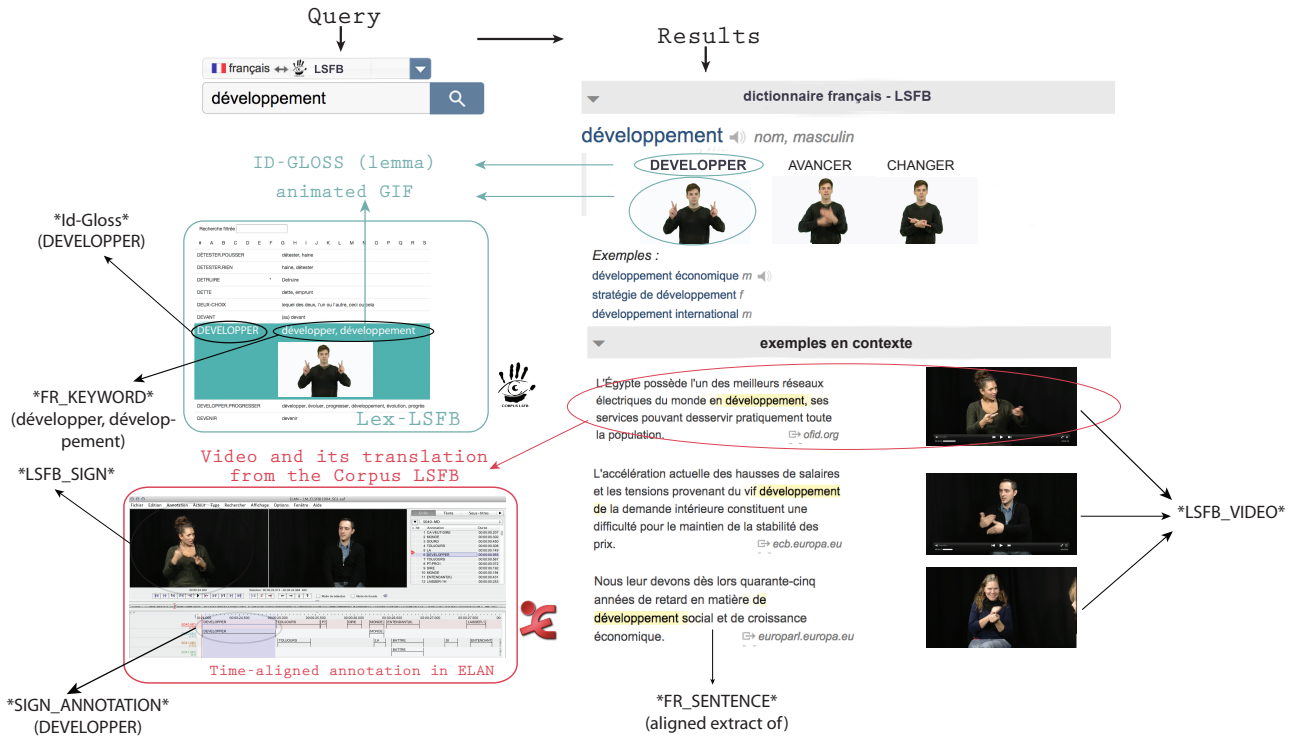

Figure 2: Model of the possible user interface of a bilingual tool (LSFB-French) derived from the parallel data. This figure is based on fictive examples and inspired by the Linguee user interface. The terms between asterisks refer to the entity names used in Figure 3

into sentences, and to establish the automatic alignment between the SL data (LSFB and NGT) and the written translation (French and Dutch respectively) based on other cues. We are currently working with a set of four kinds of information.

First of all, we can rely on the existing alignment created at the time of the translation process. Indeed, the peculiarity of our translation corpora consists in the fact that in both the LSFB and the NGT corpora, the translations have been encoded in ELAN, and thus time-aligned on the video data in the same way as the glosses are. In the LSFB corpus, the translators themselves were not asked to align their text to the video: the minimal unit they had to take into account was the turn. Afterwards, the alignment in ELAN was made by combining the segmentation of the translator into French sentences or paragraphs, the thematic coherence of the discourse, and finally the pragmatic display constraint of not making the translation segments too long to read in the website of the corpus. As for the NGT corpus, the translators directly entered their Dutch text in ELAN, aligning it at the level of the sentence-like unit in NGT. In any case, both corpora already provide a first alignment between the SL data and the written translations: in the LSFB corpus at the level of a paragraph-like group of French sentences, and in the NGT corpus at the level of the sentence. The issue lies in narrowing the scope of this existing alignment. Three other elements are exploited for this purpose.

Second, anchor signs and words are identified within the existing segments. A sign-word pair is considered as anchor when, in the available data, the sign and the word are strongly related or, in other words, have a high level of translation correspondence. Examples of such signs-word pairs are expected to be found amongst the numbers, the manually spelled words, or the colours signs, to name a few.

Third, the identification of the anchors pairs is supported by the semantic information provided manually during the annotation process. In the LSFB corpus, each entry of the lexical database has been provided with a list of possible meanings of the sign in French. In the NGT corpus, the meaning of each sign in context is specified for each sign within a dedicated tier of the ELAN file. This semantic content can be considered as starting bilingual lexicons for each pair of languages. In addition, this information constitute a first indicator of whether a sign is a good candidate to function as an anchor.

Fourth, we extracted for each gloss of the lexical databases a list of the preceding and following context of the sign. The result is a list of collocations for each entry, i.e. a list of the common sign combinations harvested from the data. Doing the same for each lemma of the translations, and linking the lists of each pair of contrasted languages (LSFB-French or NGT-Dutch), we expect to identify potential translation blocks, and thus refine the localisation of the semantic equivalents from the one language to the other, and finally improve the automatic alignment of the bilingual data.

The efficiency of the combination of these four resources for the alignment of the data will be tested pretty soon.

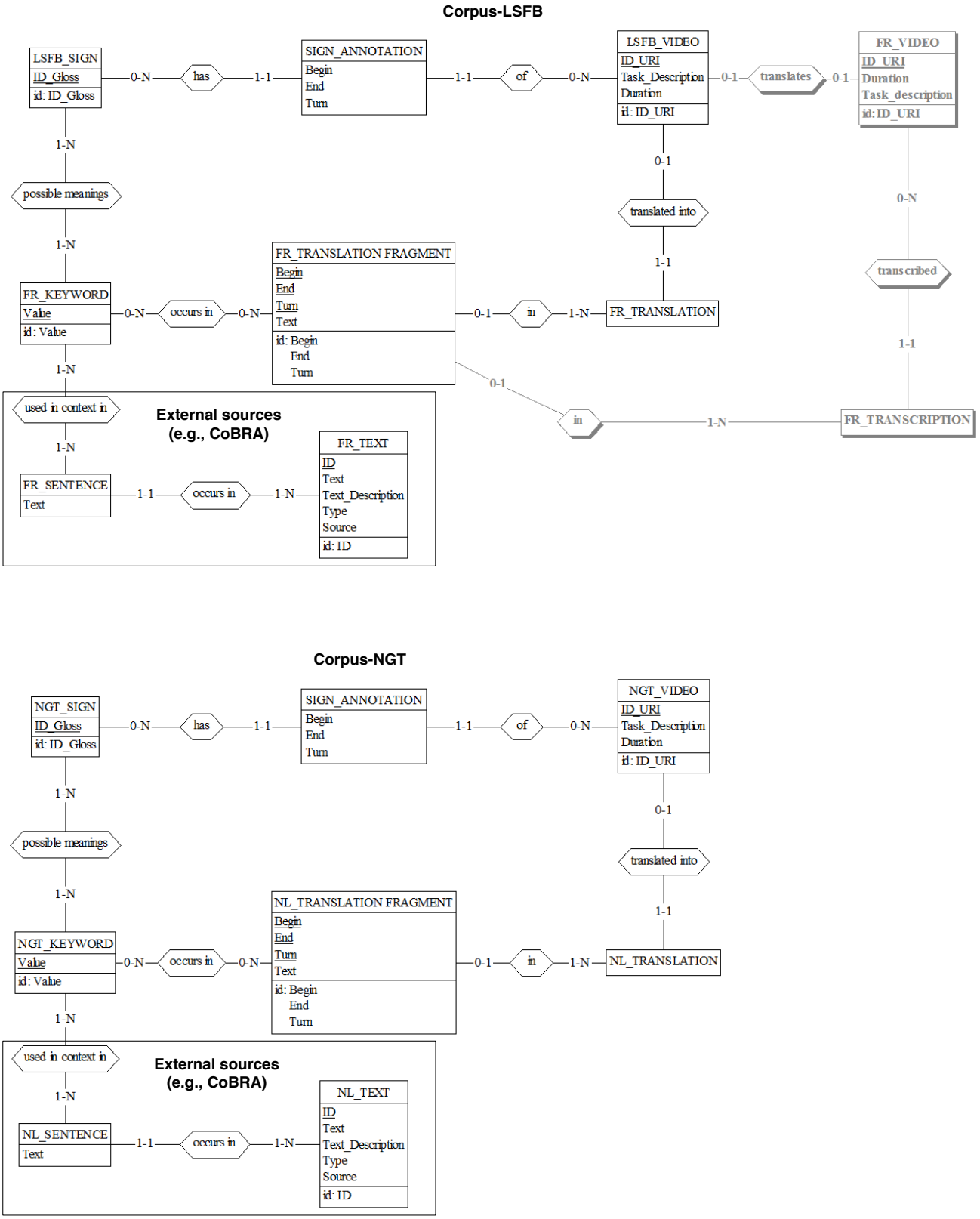

Figure 3: Entity-Relationship model of the data at our disposal. Black elements represent already available data. Light grey elements represent work in progress.

4.2. Extraction of semantic equivalents

Thanks to the alignment set up for the available data (104,000 tokens of LSFB and 2,400 sentences of French translation; 150,000 tokens of NGT and 14,000 sentences of Dutch translation), the second step will be to extract for each sign, starting with the more frequent ones, a series of parallel examples of use from the videotaped data and their written translation (see Section 2.2.). Among those examples of semantic equivalents, we will manually distinguish when the different meanings of a single sign in context occur. This task will be done with the help of the translations, but also with the help of the external data provided by the CoBRA (Corpus Based Reading Assistant) tool (Deville et al., 2013) (see Section 3.). CoBRA will be a source of additional examples of the French and Dutch words in use, and even more importantly, the DiCoBRA dictionaries that includes semantic information about French and Dutch words[4].

Harvesting this semantic information on the meanings of signs in context will supplement the starting data currently available in the lexical database of each corpus. In turn, the semantic equivalents and the information about the polysemy of the signs will be re-injected for the automatic alignment learning, together with the resources presented in Section 4.1. And, last but not least, the outcome of this second step is aimed to be exploited in order to provide an assistance the the annotators of SL data while annotating.

4.3. Towards a tool support to annotation

The task of the annotator should be assisted in different ways by the outputs of the preceding steps of the development of the concordancer. Having learned from the available aligned data, the program will be able to invite the annotator, when she/he chooses a gloss, to select a meaning amongst suggested options. Suggestions will also be improved by the collocations tables established as presented in Section 4.1. As the annotation progresses (within the phrase, then the clause or even the thematic chunk), taking into account the collocations of the annotations and of the selected meanings, the suggestions are expected to be more and more accurate. Eventually, we plan to investigate the efficiency of this "meaning in context" database and the associated list of anchor signs-words pairs and collocations for the purpose of suggesting annotations based on a previously translated text of the sign language data.

5. Perspectives

Now that more and more SL corpora are created, linguists face the challenge of mining interesting data amongst the large amount of the collected video files and the patiently accumulated annotations. This paper suggests directions towards the use of translated SL corpora as parallel corpora, and indicates how they could be exploited as a way to speed up the annotation process, but also as an insightful probe to get new insights into SLs and on the comparison between SLs and spoken languages in their written form. We propose to combine statistical and machine learning approaches to the manual work of annotators and translators. Our motivation for this bilingual approach to SL corpus linguistics eventually has the benefit of supporting the development of deaf learners and deaf users of written language, and will stimulate SL-SpL translation and interpreting studies.

[4]This semantic information has been produced from a contrastive approach of the existing dictionaries of each language and completed by the contrastive data provided by the concordancer the authors called 'Dico Corpus'. In other words, DICoBRA corresponds, for French and Dutch words, to the outcome we aim for LSFB and NGT signs: a corpus-based dictionary of signs in context build on parallel data.

6. Acknowledgements

Meurant and Cleve are supported by the University of Namur, within the framework of its research incentive program. Crasborn is supported by NWO grant 360.70.500 'Form-Meaning Units'.

7. Bibliographical References

Altenberg, B. et al. (Eds.). (2002). *Lexis in contrast: corpus-based approaches*. John Benjamins Publishing.

Börstell, C., Mesch, J., and Wallin, L. (2014). Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Crasborn, O. et al. (Eds.), *Beyond the manual channel. 6th Workshop on the Representation and Processing of Sign Languages*, pp. 7–10, Reykjavik. ELRA.

Bourdaillet, J., Huet, S., Langlais, P., and Lapalme, G. (2010). TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24:241–271.

Crasborn, O. (Ed.). (2007). *Identifying sentences in signed languages*, volume 10-2 of *Special issue of Sign Language and Linguistics*. John Benjamins, Amsterdam; Philadelphia.

Deville, G., Dumortier, L., Meurisse, J.-R., and Miceli, M. (2013). Ressources lexicales: Contenu, construction, utilisation, évaluation. In Gala, N. et al. (Eds.), *Ressources lexicales pour l´aide à l´apprentissage des langues*, volume 30, pp. 291–312. John Benjamins.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. In Crasborn, O. (Ed.), *Special issue of Sign Language & Linguistics*, volume 10-2, pp. 177–200. John Benjamins, Amsterdam; Philadelphia.

Gellerstam, M. (1996). Translations as a source for cross-linguistic studies. *Lund studies in English*, 88:53–62.

Gilquin, G. (2000). The integrated contrastive model: Spicing up your data. *Languages in Contrast*, 3:95–123.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Aijmer, K. et al. (Eds.), *Languages in Contrast. Text-based cross-linguistic studies*, volume 88 of *Lund Studies in English*, pp. 37–51. Lund University Press.

Johansson, S. (2007). Seeing through multilingual corpora. *Language and Computers*, 62:51–71.

Johnston, T. and Schembri, A. (2010). Variation, lexicalization and grammaticalization in signed languages. *Langage et société*, 131:19–35.

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15:106–131.

Linguee. (2015). `http://www.linguee.com`. [accessed 25-10-2015].

Ormel, E. and Crasborn, O. (2012). Prosodic correlates of sentences in signed languages: A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2):109–145.

8. Language Resource References

Crasborn, O. and Zwitserlood, I. and Ros, J. (2008). *The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands*. Centre for Language Studies, Radboud University Nijmegen, ISLRN 175-346-174-413-3.

Meurant, L. (2015). *Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). Laboratoire de langue des signes de Belgique francophone (LSFB-Lab).* FRS-F.N.R.S et Université de Namur.