Modelling a parallel corpus of French and French Belgian Sign Language (LSFB)

Meurant, Laurence; Gobert, Maxime; Cleve, Anthony

Published in:
Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016

Publication date:
2016

Document Version
Peer reviewed version

Towards a corpus-based online tool for French - Sign Language (LSFB) needs

The overarching objective underlying this proposal is to develop an online tool, based on a parallel corpus of French Belgian Sign Language and written French data, and aimed to assist a various set of tasks related to the comparison of LSFB and French. These tasks include (1) the comprehension of LSFB or French texts, (2) the production of LSFB or French texts, (3) the translation between LSFB and French (in both directions) and (4) the comparative description of these languages. To the best of our knowledge, this kind of tool has never been created before.

The first step of investigation aims at creating a (unidirectional) French-LSFB (then, written-video) concordancer associated with a colocation calculator and to test the efficiency of this concordancer for the extraction of a dictionary of meanings in context. In this proposal, we will present the modelling of the different data sources at our disposal and specifically the way they interact with one another.

Linguistic resources and Sign languages

Sign languages (SLs) are among the less-resourced languages of the world, due to the combined impact of various factors including their status of minority languages, the lack of written form of these visual-gestural languages, and their only recent official acceptance and recognition in the society. Research in SL linguistics is generally considered to have begun with the works of Tervoort (1953) and Stokoe (1960). The foundation of the Sign Language Linguistics Society in 2004 symbolized that research on SLs had become a worldwide effort. The digital revolution has had an important impact on the knowledge of Sign Languages since it opened (in the early 2000s only) the possibility to develop Corpus linguistics on sign languages by collecting, archiving, annotating and documenting large scale videotaped data (Johnston, 2010). Nowadays, annotating SL data remains a manual and time-consuming task. But this slow process is unavoidable at this stage in order to enlarge the available data needed to automate the process in a short future. We hypothesize that building our bilingual tool will contribute to this movement towards annotation automation.

In 2012 started in the University of Namur the “Corpus LSFB” project. The project will be published (online and open access) at the end of 2015. If this resource is essential to the linguistic description of LSFB, it is 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.

From Sign language corpora to Sign-spoken languages tools

Beyond the domain of SL linguistics, the computer revolution also impacted the domain of contrastive linguistics in general by having allowed 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 languages correspondences in context that can serve as testbeds for linguistic theories and hypotheses, but they are also essential for applications in the fields of lexicography, natural language processing, automatic or machine-assisted translation and language teaching (Altenberg et al. 2002, Johansson, 2007). Multilingual corpora are the basis of all multilingual concordancer as TransSearch (Bourdaillet et al. 2010) or Linguee (Linguee 2015).

Due to the visual-gestural nature of SLs, most modern SL machine-readable corpora, as the Corpus LSFB is, are multimodal corpora, 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 property of SL corpora has not been exploited yet for the development of bilingual tools. On the one hand, sign language engineering is mostly devoted to automatic or assisted translation tools (e.g. the ‘SignSpeak’ project[1]; Filhol et al., 2014; Dreuw et al., 2010; Morrissey et al., 2005), or for SL recognition (e.g. the ‘Dicta Sign’ project[2]; Dreuw et al., 2008). On the other hand, as SL corpora are recent and their number is still small, corpus-based SL dictionaries are scarce (the German Sign Language (DGS) dictionary in preparation[3] is an exception).

The innovation of this research lies in making a SL corpus, i.e. the Corpus LSFB, an **aligned and searchable translation corpus** (namely LSFB productions translated by human translators into French), **providing aligned bilingual examples of words and signs in context**, at the service of language teaching, translation, and contrastive linguistics. An overview of the possible user interface of the tool is shown in Figure 1 (the terms between asterisks refer to the entities’ names used in Figure 2).

---

[1] http://www.signspeak.eu/

[2] http://www.dictasign.eu/

[3] http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/dictionary.html

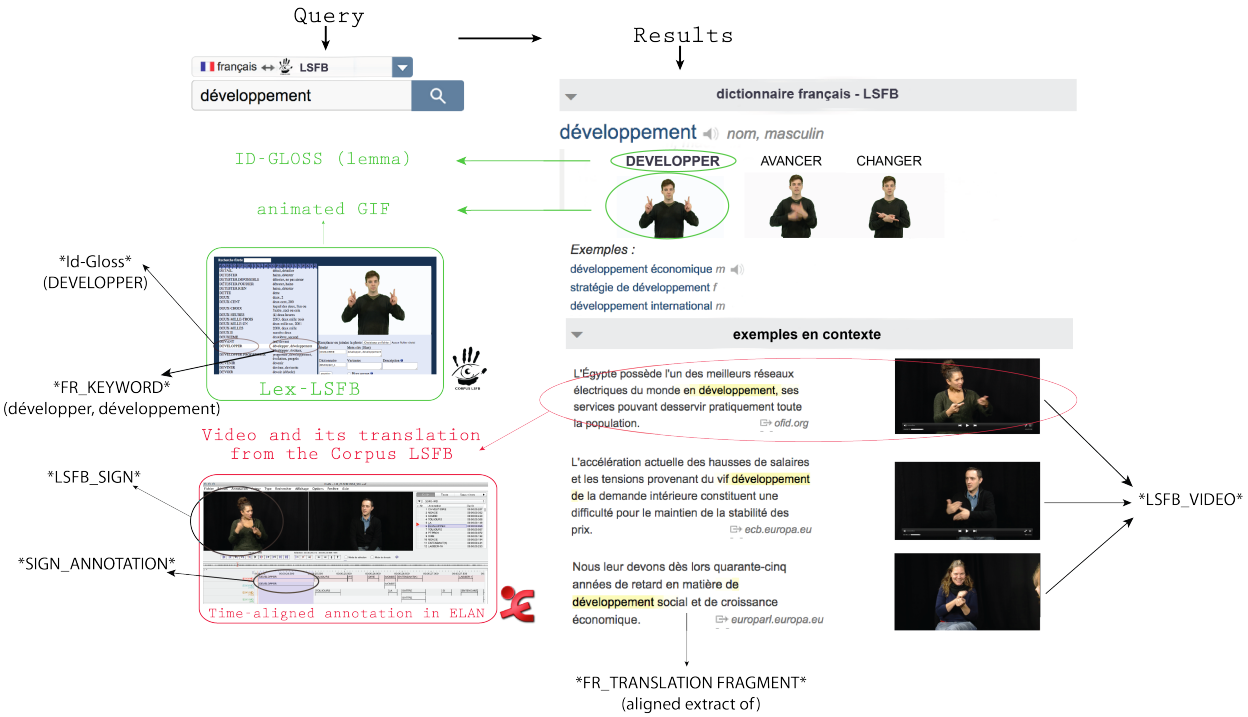

Figure 1 : Model of the possible user interface based on fictive examples (inspired by the Linguee user interface). The terms between asterisks refer to the entities' names used in Figure 2

Available resources

Data and tools at our disposal are:

- LSFB data: 125 hours of HD, 50 f/sec. videos from "the Corpus LSFB", containing semi-directed spontaneous productions of 50 pairs of signers from Brussels and all regions of Wallonia. The productions are elicited by a systematic list of 18 tasks guided by a deaf moderator and covering a variety of genres (narratives, conversations, explanations, argumentations and descriptions) and topics.
- LSFB annotations: 10 out of these 125 hours of videos are glossed, which means that each sign is given an ID-Gloss (Johnston, 2010), namely a written label of the lemma corresponding to the sign token (e.g. the ID-Gloss "PENSER" ('think') for all the possible forms of the verb). The annotation is made in ELAN[4], which allows creating and searching annotations aligned with video sources. The annotation files are associated and time-aligned with the corresponding videos.
- Translations: 5 out of the 10 hours of annotated data have a free translation in French.
- Lex-LSFB: all the ID-Glosses entered in the annotation files (currently 2500 entries) are collected within an online, constantly evolving lexical database: Lex-LSFB. Each entry of Lex-LSFB includes the ID-Gloss of the sign, one or several possible translations of the sign in French, an animated GIF file showing the sign in isolation, and information about the variants of the sign. Lex-LSFB and ELAN are connected: each entry of Lex-LSFB is linked to the various occurrences of the sign in the videos.

External tools and related data resources are also available:

- CoBRA (Corpus Based Reading Assistant) is an online and interactive tool developed at University of Namur (Deville et al., 2013), based on bilingual corpora (Dutch-French and English-French) aligned at the level of the sentence, that 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. The translation of each term is illustrated by a series of bilingual citations extracted from bilingual corpora. 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".
- CoBRA Corpus: the CoBRA resources 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.

[4] http://tla.mpi.nl/tools/tla-tools/elan/

aligned bilingual examples), an English-French glossary of about 19.000 entries, and a Dutch-French glossary of about 20.000 entries.

- DiCoBRA: CoBRA's dictionary includes circa 87.000 lemma and 300.000 inflected forms of French.

Modelling Corpus-LSFB data resources

The first step of this project consisted in formally modelling the various data artifacts involved in the Corpus-LSFB and its environment. Figure 2 provides a simplified "helicopter-view" of this data model, by means of an *Entity-Relationship* (ER) diagram. This model represents the main concepts involved in the corpus, as well as their characteristics and relationships.

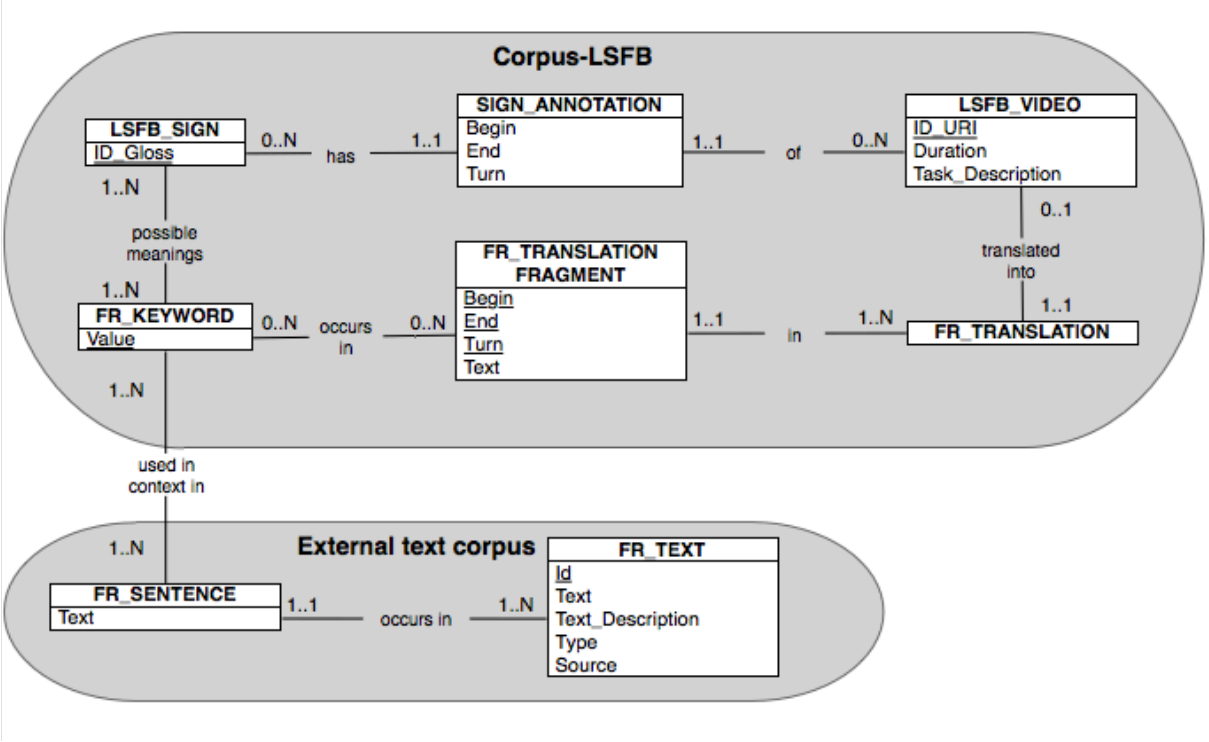

Figure 2 : Simplified Entity-Relationship diagram of the data model

The **Corpus-LSFB** consists of a set of videos (**LSFB_VIDEO**) where two signers achieved a **task** in LSFB. 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**).

The corpus also includes a large set of LSFB signs (LSFB), characterised by a unique id gloss (**ID_Gloss**). Each LFSB sign in the corpus is linked to a set of french keywords (**FR_KEYWORD**) that represent the different possible meanings of the LSFB sign.

The occurrence of a given LSFB sign in a video is represented through 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.** By convention, the value of attribute **Turn** is either 'A' (signer A) or 'B' (signer B).

The corpus also provides, for a subset of the LSFB videos, the full French translation (**FR_TRANSLATION**) of the task. Each translation is made up of a set of French translation fragments (**FR_TRANSLATION_FRAGMENT**), that is a French text fragment (**Text**) translating what is expressed in LSFB by one of the two persons (**Turn**) during time interval [**Begin**, **End**] of the video**.**

An **external text corpus** gracefully complements the Corpus-LSFB. This corpus consists of a large set of French texts, available through the CoBRA toolsuite. Those texts are in turn composed of French sentences (**FR_SENTENCE**), where contextual occurrences of each **FR_KEYWORD** (or one of its inflected forms) may possibly be found.

Populating the Corpus-LSFB data model

Once the various Corpus-LSFB data artifacts have been modelled in a conceptual way, an important and challenging task remains to be done: populating the data mode in order to eventually build a queryable Corpus-LSFB database. At the time of writing this abstract, this task has *almost* been achieved in the sense that most data artifacts are currently recorded and/or referenced in the ELAN tool (**LSFB_VIDEO**, **SIGN_ANNOTATION, FR_TRANSLATION** and **FR_TRANSLATION_FRAGMENT**) and in the Lex-LSFB database (**LSFB_SIGN** and **FR_KEYWORD**). Both environments are connected to each other since the ELAN annotations make use of the same **ID_Gloss** values of the Lex-LSFB database. The external resources are mainly available through the CoBRA toolsuite, the database of which allows one to recover the different contexts of occurrence (**FR_SENTENCE**) of a French keyword (**FR_KEYWORD**), or of one of its inflected form, in a large aligned corpus of French texts (**FR_TEXT**).

Exploiting the Corpus-LSFB populated data model

Once a complete and queryable Corpus-LSFB database will be available, we anticipate several usage scenarios of this database in the short term, as a first stage of the overall bilingual tool that we want to develop in the longer term.

First, we will need to develop a tool allowing us to further align (at the level of the French translation fragment) both sides of the Corpus LSFB, namely the annotated videos of LSFB productions and their French textual translations. The challenge of this task will be to automatically relate each sign annotation (i.e., an ID_Gloss and a [begin,end] time interval) to the corresponding *part* of the French translation fragment where the *meaning* of this sign is given *in context*. The [Begin, End] time interval of the French translation fragments will allow the alignment tool to identify the *right* fragment (the one that *includes* the [Begin, End] time interval of the sign). However, since the very same translation fragment may be linked to *several* (possibly numerous !) successive LSFB signs, the tool will then need to further slice the translation fragment in smaller fragments, each relating to *one* “clause-like” unit in LSFB.

With the help of the alignment tool described above, we will derive a finer-grained Corpus-LSFB database, that we will then use as a basis to build advanced interactive tools and conduct systematic (contrastive) linguistic studies. The first interactive tool we want to build is a searchable concordancer that exploits the aligned, fine-grained Corpus-LSFB database as well as the external corpus in order to find the LSFB concordances related to a given input term in French, as shown in Figure 1.

From a scientific point of view, the very same database can also be exploited for conducting linguistic studies and enrich our knowledge of the contextual usages and meanings of a LSFB sign, but also to compare the ways (lexicon, paraphrase, depicting structures, etc.) French and LSFB express the same ideas.

References

Altenberg, B. & Granger, S., eds. *Lexis in contrast: corpus-based approaches*. Vol. 7. John Benjamins Publishing, 2002.

Bourdaillet, J., Huet, S., Langlais, P., & Lapalme, G. *TransSearch: from a bilingual concordancer to a translation finder*. *Machine Translation*, *24*(3-4), 241-271, 2010.

Deville, G., Dumortier, L., Meurisse, J-R. & Miceli, M. Ressources lexicales pour l’aide à l’apprentissage des langues In Gala, N. & Zock, M. (eds). *Ressources Lexicales: Contenu, construction, utilisation, évaluation*. John Benjamins, Vol 30 , p. 291-312, 2013.

Dreuw, P., Ney, H., Pérez, G. M., Crasborn, O., Piater, J. H., Moya, J. M., & Wheatley, M. The SignSpeak Project-Bridging the Gap Between Signers and Speakers. In *LREC,* 2010.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., & Ney, H. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In *LREC*, 2008.

M. Filhol & Tannier, X. *Construction of a French-LSF corpus*, Building and Using Comparable Corpora. In *LREC*, 2014.

Johnston, T. From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1), 106-131, 2010.

Johansson, Stig. Seeing through multilingual corpora. *Language and Computers* 62.1 (2007): 51-71.

Linguee, http://www.linguee.com/, last visit 25/10/15.

Meurant, L. *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, 2015.

Morrissey, S., & Way, A. An example-based approach to translating sign language, In *Workshop Example-Based Machine Translation (MT X–05)*, pages 109–116, Phuket, Thailand, September, 2005.

Stokoe, William C. Sign Language Stucture: An Outline of the Visual Communication Systems of the American Deaf. *Studies in Linguistics: Occasional papers* 8, 1960.

Tervoort, B. T. M. (1953). *Structurele analyse van visueel taalgebruik binnen een groep dove kinderen*.