Modelling a Parallel Corpus of French and French Belgian Sign Language (LSFB)

Laurence Meurant, Maxime Gobert, Anthony Cleve
LSFB-Lab, PReCISE research center, PReCISE research center
University of Namur - rue de Bruxelles, 61 - 5000 Namur - Belgium
laurence.meurant@unamur.be, maxime.gobert@unamur.be, anthony.cleve@unamur.be

Abstract

The overarching objective underlying this research is to develop an online tool, based on a parallel corpus of French Belgian Sign Language (LSFB) and written Belgian French. This tool is aimed to assist various set of tasks related to the comparison of LSFB and French, to the benefit of general users as well as teachers in bilingual schools, translators and interpreters, as well as linguists. 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 contrastive analysis of these languages. The first step of investigation aims at creating an unidirectional French-LSFB concordancer, able to align a one- or multiple-word expression from the French translated text with its corresponding expressions in the videotaped LSFB productions. We aim at testing the efficiency of this concordancer for the extraction of a dictionary of meanings in context. In this paper, we will present the modelling of the different data sources at our disposal and specifically the way they interact with one another.

Keywords: Sign language, written language, parallel corpus

1. 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 as 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 SLs 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. The basic annotation task consists in identifying in the signing flow each sign type, or lemma, and associating it to a written gloss (an ID-gloss) so that each lemma is labelled with the same and unique gloss throughout the data. This slow process is unavoidable at this stage because the amount of data needs to be enlarged in order to automate the process in a short future. We hypothesize that building our parallel corpus and its concordancer will contribute to this movement towards automated annotation. In 2015, the first online, large scale and searchable corpus of LSFB was published (Meurant, 2015)[1]. This resource is not only essential to the linguistic description of LSFB, but also a potential wealth of information for pedagogic purposes, for the field of translation and interpreting studies and for the field of contrastive linguistics between signed and spoken languages.

2. From Sign Language Corpora to Parallel Corpora

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 language 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 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).

Due to the visual-gestural nature of SLs, most of the modern SL machine-readable corpora, as the Corpus LSFB is, are multimodal corpora: 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[2]; Filhol and Tannier (2014); Dreuw et al. (2010); Morrissey and Way (2005)), or for SL recognition (e.g. the "Dicta Sign" project[3]; 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[4] is an exception.

[1] http://www.corpus-lsfb.be

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

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

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

The innovation of this research lies in making a SL corpus, i.e. the Corpus LSFB, an aligned and searchable translation corpus made of LSFB productions and their human translations into written French. These aligned data will be exploited to provide examples of words and signs in context, at the service of language teaching, translation, and contrastive linguistics.

Figure 1 provides a fictive user interface of the tool we expect to derive from such a searchable translation corpus. The building of this interface is planned as the final step of our research. In order to reach it, the ongoing work focuses on the development of the concordancer.

Since most modern SL corpora are made of the same data components, namely glossed videos, translations into written language and lexical database, the concordancer we are developping is expected to be transferable to other signed - written languages pairs: e.g. Sign Language of the Netherlands (NGT) and written Dutch via the Corpus NGT (Crasborn and Zwitserlood, 2008); Flemish Sign Language (VGT) and written Dutch via the VGT Corpus (Van Herreweghe et al., 2015), or French Sign Language (LSF) and written French via the CREAGEST data (Balvet et al., 2010)[5].

3. Available Resources

The data and tools at our disposal come from the components of the Corpus LSFB (Meurant, 2015):

- LSFB data: 150 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. A presentation of the content of the corpus is available online[6].
- LSFB annotations: 12 out of these 150 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 sign). To date, 104,000 tokens are glossed. The annotation is made in ELAN[7], an annotation tool designed for the creation of complex annotations aligned to video and audio streams. The annotation files (.eaf format) are associated and time-aligned with the corresponding videos. Within the web interface of the Corpus LSFB, the annotations may be shown, when they are available, by clicking on the appropriate symbol above the video viewers.
- Translations: 3 out of the 12 hours of annotated data have been translated into French (2,400 sentences). The translations are target oriented: the text is produced in the most natural French possible, reflecting the influence of LSFB in the French lexicon or syntactic structure as little as possible. The oral features that characterize the LSFB semi-spontaneous conversations have been translated into the French text as magazines do when transcribing an interview. Within the Corpus LSFB website, the translations may be shown at the same time the video is playing; a specific button serves this purpose.
- Lex-LSFB: all the ID-glosses entered in the annotation files (currently 2,500 entries or types) are collected within an online, constantly evolving lexical database: the Lex-LSFB. Each entry of the Lex-LSFB includes the ID-gloss of the sign, one or several possible translations of the sign into French, an animated GIF file showing the sign in isolation, and information about the variants of the sign. This lexical database is visible on the Corpus LSFB website[8]. The Lex-LSFB and the annotation files are connected: each entry of the Lex-LSFB is linked to the various occurrences of the sign in the videos.

Content available from the Corpus LSFB	
Videos	150h x 4 synchronized camera shots
	HD quality, 50f/sec.
	Semi-spontaneous dialogs
Annotations	12h, i.e. 104,000 tokens
(ID-glosses)	ELAN files (.eaf)
	Separate annotations for right and left hand
	Each token is linked to its entry in the lexical database
Lexical database	2,500 entries/types For each entry: ID-gloss,possible phonetical variants, possible translation(s) in French, animated GIF file showing the sign
Translations	3h, i.e. 2,400 sentences
(French text)	

Table 1: Summary of the data available to date from the online Corpus LSFB

These existing data and tools will be complemented by external tools and related data resources:

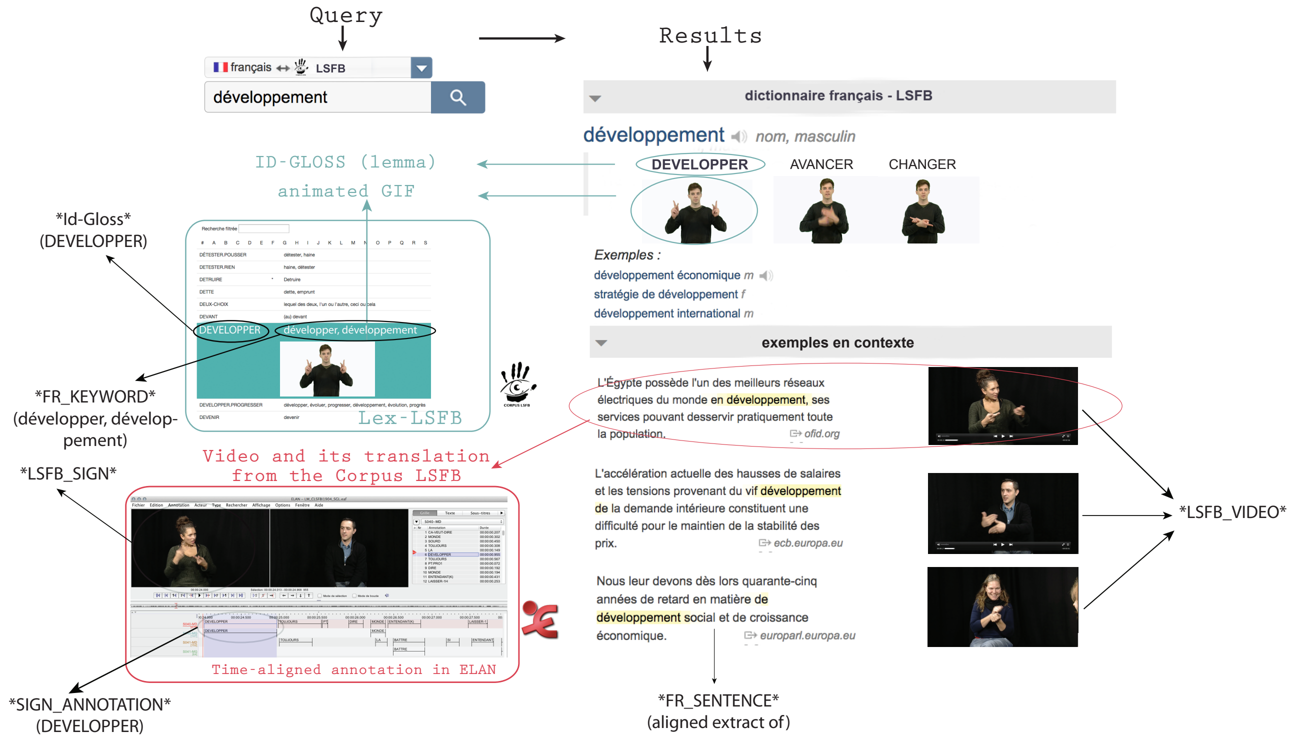
- CoBRA (Corpus Based Reading Assistant) is an online and interactive tool developed at University of Namur (Deville et al., 2013), based on bilingual corpora (Dutch-French and English-French) aligned at the level of the sentence. It allows the teachers to create labelled texts in Dutch (NL) or in English (EN) and French-speaking learners to be assisted in their reading by clicking on any word in order to know its mean-

[5] LSFB is historically related to LSF: both come from Old French Sign Language. Despite the lack of comparative studies between LSF and LSFB, it can be asserted on the experience of signers and interpreters that both languages differ mainly in terms of their lexicon and share the same morpho-styntactic features. LSFB is also closely related to VGT. In a recent past, people usually referred to one Belgian Sign Language, with regional varieties both in the North and the South of the country. But because of the Belgian federalization process in the 1970s, the contacts between the deaf communities from the North and from the South became less and less frequent, and both sign languages developped separately.

[6] http://www.corpus-lsfb.be/content.php

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

[8] http://www.corpus-lsfb.be/lexique.php



L'Égypte possède l'un des meilleurs réseaux électriques du monde en développement, ses services pouvant desservir pratiquement toute la population. ofid.org

L'accélération actuelle des hausses de salaires et les tensions provenant du vif développement de la demande intérieure constituent une difficulté pour le maintien de la stabilité des prix. ecb.europa.eu

Nous leur devons dès lors quarante-cinq années de retard en matière de développement social et de croissance économique. europarl.europa.eu

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 entity names used in Figure 2.

ing 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. 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.

4. Modelling the Data Resources

In order to build the concordancer, we first need to model the data at our disposal, exploiting the various data artifacts involved in the Corpus LSFB as well as additional external data sources. 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 data, as well as their characteristics and relationships.

In an ER diagram, each box represents an entity type (i.e. a concept of the application domain), that represents a collection of entities of the same nature. Each entity type may have a certain number of attributes, representing properties that can be associated with the entities of the collection. If a subset of attributes is underlined in the diagram, it means that those attributes constitute a unique identifier of the entity type, i.e. there cannot be two distinct entities in the collection with the same respective values for those attributes. For example, there cannot be two LSFB_SIGN with the same value of attribute ID_Gloss. In other words, each LSFB sign has a unique ID_Gloss. The links between the entity types are called relationship types. They represent the set of possible relationships that may hold between the instances of each entity type. Each relationship type R linking two entity types E1 and E2 has two roles, one played by E1 and one played by E2. Each role played by an entity type E has a minimum cardinality and a maximum cardinality, specifying the minimum (resp. maximum) number of relationships of type R that can link a given instance of E to an instance of the other entity type playing a role in R. For example, the role played by entity type LSFB_SIGN in relationship type is of cardinality 0..N (i.e. the minimum cardinality is 0 and the maximum cardinality is N). This means that each instance of LSFB_SIGN has between 0 and N corresponding instances of SIGN_ANNOTATION. The cardinality of the role played by SIGN_ANNOTATION is 1..1, meaning that an instance of SIGN_ANNOTATION is associated to 1 and only 1 LSFB_SIGN.

Within the terms of an ER diagram, the data described in Section 3. can be described as follows. The Corpus-LSFB consists of a set of videos (LSFB_VIDEO) where two signers achieved a task in LSFB. Each video is iden-

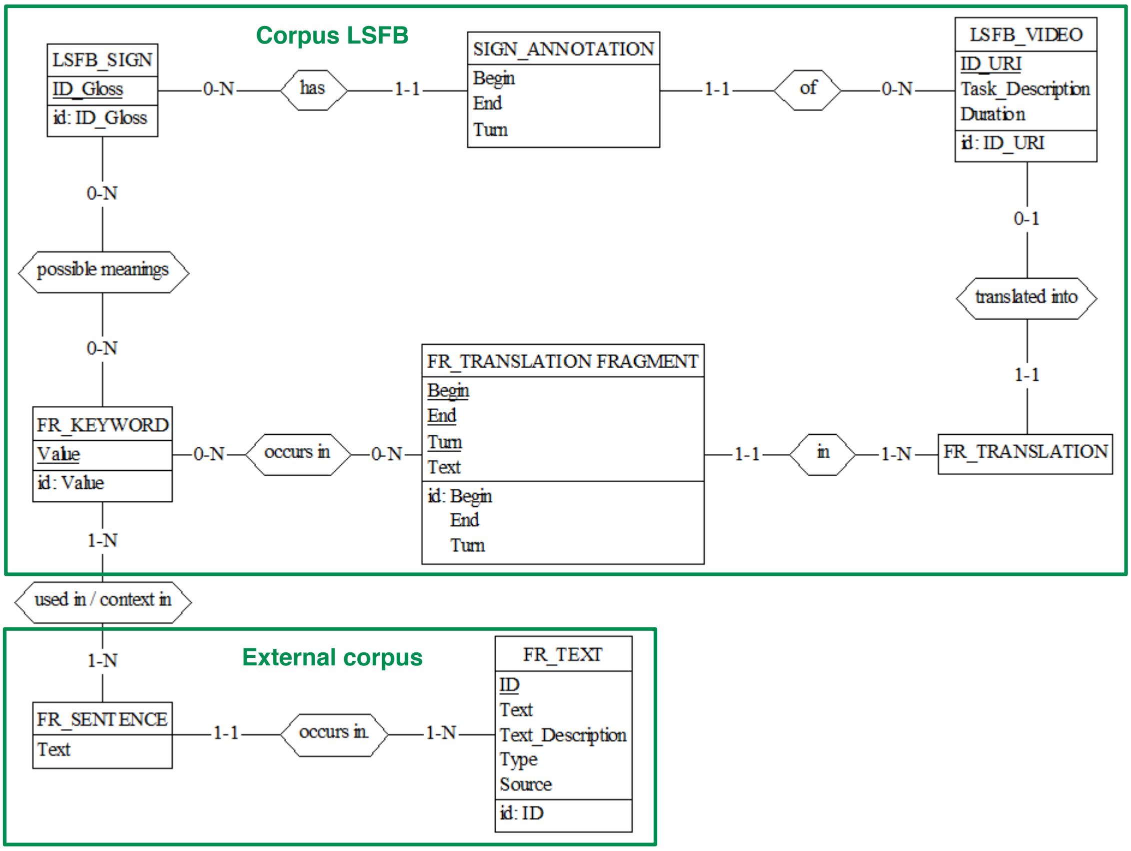


Figure 2: Simplified Entity-Relationship diagram of the data model.

tified 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 LSFB 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 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. By convention, the value of attribute Turn is either ’A’ (signer A) or ’B’ (signer B).

As mentioned above, 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.

5. Populating the Corpus LSFB Data Model

At the time of writing this paper, most data artifacts are 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 to recover the different contexts of occurrence (FR_SENTENCE) of a French keyword (FR_KEYWORD), or of one of its inflected forms, in a large aligned corpus of French texts (FR_TEXT).

6. Exploiting the Corpus LSFB Populated Data Model

Since the complete and queryable Corpus LSFB database is available, our ongoing work consists in aligning (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 is 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 and 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 in development, we expect to feed and to make more precise the current lexical database. For each entry of the Lex-LSFB, additional meanings will be provided by the various translations associated with the various tokens of the sign in context. We will then use this finer-grained database derived from the parallel data in order to improve the annotation process, by giving suggestions to the annotator. We then plan to build an advanced interactive tool such as the one fictively depicted in Figure 1, and thanks to it, 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 to enrich our knowledge of the contextual usages and meanings of a LSFB sign, but also to compare the ways (lexicon, paraphrase, depicting structures, etc.) in which French and LSFB express the same ideas.

7. Acknowledgements

This research is funded by the University of Namur, within the framework of its research incentive program.

8. Bibliographical References

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

Balvet, A., Courtin, C., Boutet, D., Cuxac, C., Fusellier-Souza, I., Garcia, B., L’Huillier, M. T., and Sallandre, M. A. (2010). The creagest project: A digitized and annotated corpus for french sign language (lsf) and natural gestural languages. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 469–475.

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. and Zwitserlood, I. (2008). The corpus NGT: an online corpus for professionals and laymen. In O. Crasborn, et al., editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pages 44–49. ELDA.

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

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., and Ney, H. (2008). Benchmark databases for video-based automatic sign language recognition. In *Proceedings of LREC 2008*.

Dreuw, P., Ney, H., Pérez, G. M., Crasborn, O., Piater, J. H., Moya, J. M., and Wheatley, M. (2010). The SignSpeak project – bridging the gap between signers and speakers. In *Proceedings of LREC 2010*.

Filhol, M. and Tannier, X. (2014). Construction of a french-lsf corpus, building and using comparable corpora. In *Proceedings of LREC 2014*.

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

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].

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

Stokoe, W. C. (1960). Sign language stucture: An outline of the visual communication systems of the american deaf. *Studies in Linguistics: Occasional papers 8*.

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

Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., and Verstraete, S. (2015). Het corpus VGT. Een digitaal open access corpus van video’s and annotaties van Vlaamse Gebarentaal. www.corpusvgt.be. ontwikkeld aan de Universiteit Gent ism KU Leuven.

9. 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.