Toward Categorization of Sign Language Corpora

Jérémie Segouat
LIMSI-CNRS / Orsay, France
WebSourd / Toulouse, France
jeremie.segouat@limsi.fr

Annelies Braffort
LIMSI-CNRS / Orsay, France
annelies.braffort@limsi.fr

Abstract

This paper addresses the notion of parallel, noisy parallel and comparable corpora in the sign language research field. As it is quite a new field, the categorization of sign language corpora is not well established, and does not rely on a straightforward basis. Nevertheless, several kinds of corpora are now available and could raise interesting issues, provided that adapted tools and techniques are developed.

1 Introduction

Sign Language (SL) is a visual-gestural language, using the whole upper body articulators (chest, arms, hands, head, face, and gaze) in a simultaneous way. Signs (in some way, equivalent to words in vocal languages) are articulated in the signing space located in front of the signer. This is a natural language, with its own linguistic structures and specificities, used by deaf people to communicate in everyday life. It can be considered that there is one SL for each country, as for vocal languages. One particularity is that there is no written form of SL (Garcia, 2006): corpora take the form of videos, thus specific design and analysis methods have to be used. Therefore, NLP and corpus linguistics definitions may have to be adapted to this research field.

1.1 Brief History of Sign Language Corpora

Research in SL has begun with the creation of notation systems. These systems aim to describe in a written form how SL could be performed. Bébian (1825), a French teacher, wrote a book where he proposed a description of the French Sign Language (LSF) using drawings. This description took into account facial expressions and manual gestures. A major study was conducted by Stokoe (1960) on American SL. The aim was also to describe SL, but this time only focused on manual gestures. These studies were based upon live analyses: no video corpus was created. The researchers had to watch how signers were performing SL, and then write down or draw what they were observing.

In the 1980s, Cuxac (1996) created one of the first video SL corpora for linguistic studies. From the 1990s until now, video SL corpora have been created both to be used in linguistic studies, as listed by Brugman (2003), and for gathering lexicons to create dictionaries[1]. A few years ago, some video SL corpora were designed to serve as the basis for NLP and Image Processing (Neidle, 2000).

1.2 Definitions

Fung (2004) distinguishes four kinds of corpora: parallel (“a sentence-aligned corpus containing bilingual translations of the same document”), noisy parallel (“contain non-aligned sentences that are nevertheless mostly bilingual translations of the same document”), comparable (“contain non-sentence-aligned, non-translated bilingual documents that are topic-aligned”), and very-non-parallel (“contains far more disparate, very-non-parallel bilingual documents that could either be on the same topic (in-topic) or not (off-topic)”). If these definitions are still under discussion in the NLP community, there is no such discussion in the community which studies SLs. Would it be possible to apply such definitions to Sign Languages corpora?

Many corpora are mere dictionaries[2], i.e. they only contain isolated signs and no utterances, just signs, but could be considered as very basic parallel SL corpora. As far as we know, there exists very few noisy parallel SL corpora (see section 2.2), and very few comparable SL corpora (Bungeroth 2008, ECHO project[3]).

[1] http://www.spreadthesign.com/country/gb/
[2] http://www.limsi.fr/Scientifique/iles/Theme5/corpus
[3] http://www.let.ru.nl/sign-lang/echo/

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 64–67,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

Because not enough data can be found on the way these corpora have been built and the way they are used, it seems difficult to discuss whether Fung's definitions apply to them. Thus, we present in this paper the corpora we have built (section 2) and explain why they could be considered as parallel, noisy parallel or comparable. Section 3 discusses the use of NLP processes for SL corpora analysis, and section 4 presents prospects on existing or possible SL corpora.

2 LIMSI's Sign Language Corpora

2.1 Parallel Corpora

We are currently building a French Sign Language (LSF)-French dictionary (Segouat 2008) that will be available on the Web. We will provide not only French and LSF translations, but also linguistic descriptions of signs, and a functionality to search for signs from their visual aspects or their linguistic descriptions. This is a mere parallel corpus that will be using to analyze the variety of LSF in France (according to where people live, where they have grown, where they learned LSF, etc.).

We have recently built a corpus related to the railway information domain (Segouat, 2009). The starting point is written French sentences that exactly correspond to the vocal announcements made in railways stations. The goal is to provide information in LSF as it is provided vocally: by coarticulating pieces of utterances. Written French sentences were translated into LSF and filmed, in order to study coarticulation in LSF. We use this corpus to analyze how signs are modified according to their context.

We participate in the DictaSign European project (Efthimiou, 2009) that aims at gathering parallel SL corpora from four countries (Greece, England, Germany, and France). One of its purposes is to study translations between different sign languages (SLs) of these four countries. The welcome page of the website[4] includes presentations of the project in the four different SLs that are each direct translations of the corresponding written texts. As it is a starting project, this corpus has not yet been studied nor considered from a comparability point of view.

2.2 Noisy Parallel Corpora

We have taken part in the creation of the LS-COLIN corpus (Cuxac, 2001). The aim of this project was to design a corpus that could be used by linguists and computer scientists. The methodology was the following: each deaf signer (i.e. a person who performs SL) was explained the protocol. The person had to perform several kinds of stories, on several given themes or elicited by using pictures. For the picture based story, the deaf signer was shown six pictures that draw a line for the story, and then expressed the story in LSF. This corpus could be considered as a noisy parallel one, because the LSF version is a translation of the pictures with addition of details. The linguists have created a noisy parallel version of some parts of LS-COLIN, by providing a transcription with glosses (sign to word translation, without taking into consideration the grammatical structure involved: thus there is a lack of information). All the annotations were made in French text, and were used to analyze the grammatical structure of LSF.

We have participated to the WebSi project (Martin, 2009), which aims at evaluating whether common representations could be designed for gestures performed by speaking and signing persons, allowing bilingual applications to be developed. The first step was a study dedicated to the comparison of deictic gestures, both with multimodal-French and LSF utterances. The corpus consists of answers, by a deaf and a hearing person, to eleven questions eliciting responses with deictic gestures of various kinds. A French/LSF interpreter formulated the questions so that both subjects were in the closest possible interaction conditions. The observed productions were indeed very different. In the deaf person's answers, a more complex structure was observed in deictics, because the deictic function is incorporated into the lexical signs, forming what is called indicating signs. However, common global aspects were observed in both types of productions, which are all constituted by pointing using gaze and manual gestures organized with a given temporal structure.

2.3 Comparable corpora

In the LS-COLIN corpus, each deaf signer had to perform a story on several given themes, for example September 11 tragic events. This can be considered as a synchronous comparable corpus because each signer expressed his own version of the same event. The picture-based stories may also be considered as comparable corpora, because deaf signers were asked to perform the story twice: at the beginning and at the end of the recording. Thus it is the same topic, and the two versions are not translations of one another; but

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

we are not certain that it can be considered as “non-sentence-aligned” because they both follow picture order. Computer scientists have used LS-COLIN from a comparability point of view, to analyze the visual modality in LSF: they studied torso (Segouat, 2006) and facial (Chételat-Pelé, 2008) movements. These studies were made on same-topic stories performed by different deaf signers. While these studies did consider the comparability of the corpus, they were not focused on that aspect. Thanks to these studies, we may observe differences in sign performances among deaf signers, from crossed linguistics and computer science perspectives.

3 Computations on Sign Language Corpora

The computations in use for written data cannot be used directly for video SL corpora. Nowadays though, a way to study SL corpora is to annotate them. Annotations are mainly in written form, thus one might think of applying existing NLP methods to the resulting “texts”. But would the conclusions be relevant enough? A bias is that annotations do not exactly represent SL utterances. Annotations can be made with glosses or complete translations but these written data cannot describe in an efficient way typical SL properties such as simultaneity, spatial organization, non-manual features, etc.

In our opinion, it would thus be difficult to apply the computations used on written comparable corpora (Fung, 2004; Morin, 2006; Deléger, 2008) or on parallel corpora to comparable or parallel SL corpora.

Some studies currently focus on graphical annotations, or use image processing to analyze video SL corpora (Bungeroth, 2008). It is a first step towards an analysis without any written text processing. Suitable tools to deal with this kind of annotations still have to be set up.

4 Promising Sign Language Corpora

4.1 Existing Corpora

The Dicta-Sign project already provides a quadrilingual corpus: the website contains four versions of the same presentation in four different sign languages. An analysis of this corpus would be interesting, because all SL videos were made from the English text. The British SL, and also the other texts in French, Greek, and German were obtained from the English written source. Then the corresponding SL videos in LSF, Greek SL, and German SL were translated from the texts in written French, Greek, and German. This corpus is therefore parallel, although probably noisy because of the double written-to-written then written-to-SL translation process. Comparing these videos would allow us to notice changes in the translations between SLs, using knowledge from the written-text translation field of research.

The corpus dealing with information in French railway stations is a bilingual parallel corpus. Other corpora are going to be designed and used in projects related to bus stations, airports, etc. Therefore we will have interesting parallel (French-LSF) and comparable (same topic) about transportation systems, to study.

4.2 Other Possible Corpora

The WebSourd Company’s website[5] provides everyday news translations in LSF, displaying both the text that has been translated and the video in LSF. Each year, all videos are archived on a DVD. WebSourd is, as far as we know, the only company that provides everyday information in LSF. Collecting other sources for the same types of information would yield an interesting synchronous comparable corpus.

In SL we distinguish “translation” from “interpretation”. Both could be performed either by hearing persons from vocal languages to SLs, and vice and versa, or by deaf persons from SLs to SLs. A translation is done with significant time taken for preparing the work. It looks more like a “written” form of language, thus such translations can create parallel corpora. Interpretation is done live, and often without any preparation of what is going to be interpreted. It is more like “oral” expression, with discourse corrections, repetitions, etc., thus it is likely to produce noisy corpora. SL interpretation corpora are available (e.g. every live interpretation on TV), but as far as we know they haven’t yet been analyzed, although such study looks interesting.

There are in France[6] and in Great Britain[7] two TV programs presented in SL and made accessible with oral and written translations. These constitute a huge amount of parallel corpora (vocal language-sign language translations) that have not yet been used in any research field.

[5] http://www.websourd.org

[6] http://www.france5.fr/oeil-et-la-main/index-fr.php?page=accueil

[7] http://www.bbc.co.uk/blogs/seehear/

5 Conclusion

Until now very few parallel or comparable sign language corpora of SL have been built, and the few which exist were not studied from these points of view. Studying these parallel and comparable SL corpora for linguistics, computer science analysis, and for translation is therefore a new, yet to investigate area. What we should consider now is to set up a methodology to create those corpora with the aim to study them as what they are: parallel orcomparable. Moreover, we have to develop new tools, and adapt existing ones, that will fit this goal.

Reference

Roch-A. Bébian. 1825. *Mimographie, ou essai d'écriture mimique, propre à régulariser le langage des sourds-muets*. Paris. L. Colas eds.

Annelies Braffort, Christian Cuxac, Annick Choisier, Christophe Collet, Patrice Dalle, Ivani Fusellier, Rachid Gherbi, Guillemette Jausions, Gwenaelle Jirou, Fanch Lejeune, Boris Lenseigne, Nathalie Monteillard, Annie Risler, Marie-Anne Sallandre. 2001. *Projet LS-COLIN. Quel outil de notation pour quelle analyse de la LS ?* Colloque Recherches sur les langues des signes. Toulouse UTM eds. 71-86.

Hennie Brugman, Daan Broeder, and Gunter Senft. 2003. *Documentation of Languages and Archiving of Language Data at the Max Planck Insitute for Psycholinguistics in Nijmegen*. Ringvorlesung Bedrohte Sprachen. Bielefeld University, Germany.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way and Lynette van Zijl. 2008. *The ATIS Sign Language Corpus*. 6th International Conference on Language Resources and Evaluation. Marrakech. Morocco.

Émilie Chételat-Pelé, Annelies Braffort. 2008. *Sign Language Corpus Annotation: Toward a New Methodology*. 6th International Conference on Language Resources and Evaluation. Marrakech. Morocco.

Christian Cuxac. 1996. *Fonctions et Structures de l'iconicité dans les langues des signes; analyse d'un idiolecte parisien de la Langues des Signes Française*. Doctoral Thesis, Paris V University, France.

Louise Deléger and Pierre Zweigenbaum. 2008. *Paraphrase acquisition from comparable medical corpora of specialized and lay texts*. AMIA. Annual Fall Symposium. Washington, DC. 146-150.

Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. *Sign Language Recognition, Generation and Modelling: A Research Effort with Applications in Deaf Communication*. 13th Internation Conference on Human-Computer Interaction. San Diego, CA. USA.

Pascale Fung, Percy Cheung. 2004. *Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM*. 12th Conference on Empirical Methods in Natural Language Processing. Barcelona. Spain. 57-63.

Brigitte Garcia, 2006. *The methodological, linguistic and semiological bases for the elaboration of a written form of LSF (French Sign Language)*. 5th International Conference on Language Resources and Evaluation. Genoa. Italy.

Jean-Claude Martin, Jean-Paul Sansonnet, Annelies Braffort, and Cyril Verrecchia. 2009. *Informing the Design of Deictic Behaviors of a Web Agent with Spoken and Sign Language Video Data*. 8th International Gesture Workshop. Bielefeld, Germany.

Emmanuel Morin and Béatrice Daille. 2006. *Comparabilité de corpus et fouille terminologique multilingue*. Traitement Automatique des Langues. Vol 47. 113-136.

Carol Neidle. 2000. *SignStream(TM): A Database Tool for Research on Visual-Gestural Language*. American Sign Language Linguistic Research Project, Report No. 10. Boston University. USA.

Marie-Anne Sallandre. 2006. *Iconicity and Space in French Sign Language*. Space in languages: linguistic systems and cognitive categories. Collection Typological Studies in Language 66. John Benjamins. 239-255.

Jérémie Segouat, Annelies Braffort, and Émilie Martin. 2006. *Sign Language corpus analysis: Synchronisation of linguistic annotation and numerical data*. 5th International Conference on Language Resources and Evaluation - LREC, Genova, Italia.

Jérémie Segouat, Annelies Braffort, Laurence Bolot, Annick Choisier, Michael Filhol, and Cyril Verrecchia. 2008. *Building 3D French Sign Language lexicon*. 6th International Conference on Language Resources and Evaluation – LREC. Marrakech, Morocco.

Jérémie Segouat. 2009. *A Study of Sign Language Coarticulation*. Accessibility and Computing. SIGACCESS Newsletter. Issue 93. 31-38.

William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Washington DC. Gallaudet College Press.