Issues underlying a common Sign Language Corpora annotation scheme

Antonio Balvet

UMR 8163 STL (Université Lille 3 & CNRS),
Université Lille Nord de France F-59653 Villeneuve d'Ascq,
antonio.balvet@univ-lille3.fr

Abstract

Corpus-based Sign Language linguistics has emerged as a new linguistic domain, and as a consequence large-scale and controlled video data repositories are under construction for different Sign Languages. Nevertheless, as pointed by (Johnston, 2008) no unified annotation scheme is yet available, which compromises any chance of comparing or reusing corpora across research teams. Another related issue is the comparability of descriptions and formalizations between SL linguistics and mainstream linguistics. In this paper, we address the issue of the definition of a common annotation scheme for Sign Language corpora annotation, distribution, exchange and comparison. In section 2. we discuss the challenge of building inter-operable corpora for corpus-based linguistics. We also examine existing annotation schemes or strategies proposed for SL linguistics. In section 3. we propose a small set of annotation tiers, based on Frame-Semantics, as a common annotation scheme. We also propose to add text-level as well as utterance-level metadata to this common annotation scheme, in order to broaden the range of future uses of SL corpora.

1. Introduction

Mainstream corpus-based linguistics for oral and written languages is a flourishing research domain now that the capabilities of computers and linguistic software meet the demand of corpus-based and corpus-driven approaches both for linguistic research and applied domains of linguistics (second-language learning, lexicography, machine translation).

Sign Languages, on the other hand, are visuo-gestural and multi-segmental languages. Moreover, they have no stabilized written form, as of today, which hinders their computational processing. To make things even worse, Deafs over the world have generally been forbidden to use their natural language up until very recently[1], which has yielded great linguistic diversity. As a consequence, every aspect of their description, from the identification of basic units to the description of SL syntax or semantics, is a challenge to linguists, and even more so for computational or corpus linguists.

Sign Language linguistics can therefore be considered as a new and very challenging linguistic domain. Since most SL linguists are not native speakers of the language they are engaged in describing, at least some resort to actual language usage is necessary, even in the most formal approaches to SL linguistics. As a consequence, large-scale and controlled video data repositories are under construction for different Sign Languages: Auslan (Australian Sign Language), BSL[2] (British SL), DGS[3] (German SL), LSF[4] (French SL), and SSL[5] (Swedish SL) to name but a few. The constitution of such controlled corpora is essential to the preservation and (formalized) description of Sign Languages in their diversity. Nevertheless, as pointed by (Johnston, 2008) no unified annotation scheme is yet available, which compromises any chance of comparing or reusing corpora across research teams.

Another related issue is the comparability of descriptions and formalizations between SL linguistics and mainstream linguistics: given a set of SL corpora and their associated annotations, would a mainstream linguist be able to compare the syntax (or semantics, or any other traditional domain) of a given SL and the syntax of an oral language? Probably not, as most SL annotation schemes do not offer transcriptions (in their usual sense), and the glosses they provide are generally Sign-to-words intermediate associations rather than true morpheme-based interlinear glosses, as can be found in comparative linguistics and linguistic typology[6].

In this paper, we address the issue of the definition of a common annotation scheme for Sign Language corpora annotation, distribution, exchange and comparison, focusing on some of the necessary features of such an annotation scheme, both from SL in general and from a computational (NLP or corpus-linguistics) perspective. In section 2. we discuss the challenge of building inter-operable corpora, for SL as well as mainstream corpus-based linguistics. We also examine an existing annotation scheme proposed for the Auslan project. In section 3. we propose a tentative common annotation schemes based on Frame-Semantics. We also propose to add text-level as well as utterance-level metadata to this common annotation scheme in order to broaden the range of future uses of SL corpora, with a computational perspective (corpus-linguistics and Natural Language Processing) in mind.

2. The challenge of corpus distribution, exchange and comparison

As stated above, SL linguistics has reached a crucial point: large-scale, controlled corpora are being devised all over

[1] In the case of LSF, young Deafs were forbidden to sign during classes, up until 1991.

[2] http://www.bslcorpusproject.org/.

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

[4] http://www.creagest.cnrs.fr/.

[5] http://www.ling.su.se/pub/jsp/polopoly.jsp?d=12405&a=57659

[6] See "The Leipzig Glossing Rules" for interlinear glosses examples: http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

the world, both for preservation and description objectives, mirroring a general trend in linguistics, which has caused large-scale, representative audio and text corpora to come to existence. Nevertheless, due to different practices, backgrounds (generative grammar vs. cognitive linguistics), funding opportunities, experimental setup (elicitation vs. free interaction, monologues versus dialogs), initial application (teaching SL vs. SL research), and also computer equipment and skills, no unified annotation scheme is yet available for all these projects, as pointed by (Johnston, 2008), which jeopardizes any chance of comparing or reusing corpora across research teams. This situation is not a privilege of SL research, though: it could be said that whenever two electronic corpora for any given (oral) language exist, they only seldom share the same tagsets, linguistic material, purposes, general methodology or even size. For example, if we consider two well-known English corpora such as the Penn Treebank (Marcus et al., 1993) and the Susanne corpus (Sampson, 1994), they vary wildly in coverage: over 1 million words for the Penn Treebank, versus around 130,000 for the Susanne corpus. They also vary wildly in their initial objectives: a large-scale "quasi-industrial" syntactically annotated corpus project for the Penn Treebank, versus a small-scale consistency-oriented project for the Susanne corpus. Of course, no common annotation scheme or even metadata exist for these corpora, which entails that each end-user should learn each corpus's peculiarities. English is a well-established language, with a normalized written form and which has benefitted from a long grammatical history and an enduring research effort throughout the years. Therefore, corpus and computational linguists would be able to provide conversion tools whenever the need for inter-operability between the Penn Treebank and the Susanne corpus should arise. This is not the case for SLs in general: due to their multi-segmental and visuo-gestural modality, SLs have no normalized written form, which dramatically hinders their computational processing. At best, automatic recognition of only isolated parameters can be achieved, even with state-of-the art algorithms and pattern-recognition methods. Nevertheless, SL linguistics can benefit from the experience accumulated in mainstream corpus-linguistics. In our view, one way of guaranteeing SL corpora inter-operability are metadata.

2.1. Metadata: documenting and structuring corpora

Metadata can be considered as structured data on the data. In the framework of corpus linguistics, metadata generally serve two main purposes: The first one is the overall documentation of the source of the data, which generally entails identifying the speaker and his/her background (age, sex, education etc.), the interviewer or field linguist responsible for the data collection, the particular experimental setting used (types of cameras, exact reference, type of compression, type of recording medium: tapes, disks, flashdrives, use of lights, disposition of speakers, stimulus etc.), and other experimental variables. The second one is the structuring of each recorded corpus using in situ metadata so as to identify relevant discourse-level or utterance-level units (beginning and ending of a story, utterance or proposition boundaries, phonological/morphological/syntactic boundaries). For the purpose of corpus building, type 1 metadata are not necessarily included in the annotations associated with a given recording, while the latter generally are. Moreover, type 2 metadata are bordering on annotations, as the proper and consensual identification of many discourse or utterance-level units is a rather complex task. For example, even for written languages like English or French, the proper identification of such basic linguistic units as sentence boundaries or words is generally not an altogether easy task as inter- and even intraindividual variance are generally observed. In the domain of mainstream corpus-linguistics, the Text Encoding Initiative (TEI)[7] offers guidelines and tools for the declaration of metadata (what to document) and the proper structuring of both overall metadata (type 1 above) and *in situ* metadata (type 2). In this framework, both discourse-level (text units) and utterance-level (sentences, words) units are identified, generally in order to support further annotations (e.g. lemmatization, part-of-speech tagging, syntactic parsing, semantic tagging).

How are *in situ* metadata crucial for SL corpora? Because they provide the only proper (controlled) way, once a corpus is completely structured, to build sub-corpora out of the original corpus and the *in situ* metadata. In future uses of the SL corpora being devised to this date, we might want to consider cases where a researcher would need to study "the introduction of actants in stories told by left-handed Deaf children with a cochlear implant, from ages 5 to 7". This would only be possible if such *in situ* metadata were included in the annotated files. To our knowledge, no SL annotation scheme allows for just such *in situ* labelling and subsequent potential selection of discourse as well as utterance-level units. Therefore, in our proposal for a common SL annotation scheme, we include metadata of the type discussed above: beginning and end of stories, utterances, propositions and possibly signed units.

2.2. A discussion of the Auslan annotation scheme/strategy

The Auslan project is a large corpus archive for Australian Sign Language: annotations are expected to take at least 10 years before they reach a stage compatible with extensive corpus-based research. To our knowledge, it is one of the only SL corpus annotation projects for which an annotation strategy has been explicitly devised and published, even though the same general approach can be found in other SL corpora projects, such as NGT. In the Auslan project, one of the solutions adopted by (Johnston, 2008) for consistent annotation relies on the concept of lemmatization, applied to Sign Language annotation: "the classification or identification of related forms under a single label or lemma (the equivalent of headwords or headsigns in a dictionary)". Johnston describes the annotation protocol used for lexical signs in the framework of Auslan, where local interpretations of signs are normalized and constrained, in order to keep the set of lexical signs as small as possible: "[w]ithout lemmatization a collection of recordings [...] with various related annotation files [...] will not be able to be used as a true linguistic corpus as the counting, sorting, tagging.

[7] See TEI and TEI-Lite recommendations http://www.tei-c.org/Guidelines/Customization/Lite/

etc. of types and tokens is rendered virtually impossible.". This lemmatization process entails a high level of normalization and regularization, which in itself is not unusual in the course of corpus annotation. One of the key features of modern SL corpora, and more broadly of linguistic corpora in general, is their association with an annotation tool (Elan, Anvil, Transcriber, Praat, NiteXML...), which makes it possible to align annotations with the time indexes of the annotated media files (audio, video). Modern corpora are therefore associated with several time-aligned annotation layers, generally referred to as "tiers". One of the most important feature of these annotation tiers is that they are not intended to preserve information (encode the original information in a different format), but rather to interpret and abstract over the original signal, in order to be integrated in a formalized description, and hopefully a model (a grammar) of the described language. Therefore, every time a linguistic corpus is built, annotation issues arise, requiring linguists to arrive at a compromise between faithfulness to the original data and consistency. As Johnston points out: "[w]ithout consistency (...) it will be impossible to use the corpus productively and much of the time spent on annotation will be effectively wasted because the corpus will cease to be, or never become, machine readable in any meaningful sense."

3. Proposals for a common annotation scheme for lexical and non lexical signs

Lemmatization, or lexical sign normalization, appears as a necessary annotation strategy in the perspective of large and controlled SL corpora annotation. But, as (Johnston, 2008) points out: "[l]emmatisation can only apply to lexical signs. However, many signed meaning units found in natural signed language texts are not lexical signs." For Johnston "[lexical signs are] essentially, equivalent to the commonsense notion of *word*" whereas "the term *non-lexical sign* is reserved for a form that has little or no conventionalized or language-specific meaning value beyond that of its components in a given context." Johnston proposes annotation conventions for such non lexical signs, of which the sub-category "depicting signs" seems to encompass what (Cuxac, 1996), and more specifically the Creagest team (Balvet et al., 2010), label **Highly Iconic Structures**. In the perspective of Cuxac's semiological model of sign creation and development, these non lexical structures are a central linguistic device, both for natural human gesturality and Sign Languages. As Johnston's citation above illustrates, this position is not shared by the vast majority of Sign Language linguists, who generally assume these structures to be peripheral at best, or even outside the range of language altogether (Garcia, 2010) and (Boutet et al., 2010).

Are lemmatas enough to ensure the linguistic exploitation and reusability of SL corpora among the SL linguistics community? Moreover, are lemmatas, in association with fine-grained postural and gestural descriptions, enough for ensuring comparability between SL and oral languages corpora? Could a mainstream linguist use SL annotations to compare structures among SLs and oral languages? Probably not, especially if one aims at describing not only lexical signs, but also Highly Iconic Structures which have been shown to represent over 40% of the semantic units in LSF (and other LSs) stories and discourse[8]. Such structures are a major challenge for the formalized description of SLs: no oral language lemmatas are always available for each Transfer Structure, as they generally represent whole discourse units (propositions).

For all these reasons, we advocate in favor of Frame-Semantics primitives (Fillmore, 1977) and a Framenet-supported (Collin et al., 2008) annotation scheme for SL corpora. Frames are defined as "[having] many properties of stereotyped scenarios – situations in which speakers expect certain events to occur and states to obtain. In general, frames encode a certain amount of "real-world knowledge" in schematized form." (Lowe et al., 1997). A typical example is the "commercial transaction" Frame, in which four Frame elements are generally required: two animated actants, an amount of money and an object. The result of the process associated with this Frame is the change of ownership of the object, in exchange for money. This stereotyped scenario can be associated with a relatively large set of lexical units in different languages (*buy*, *acheter*, *kaufen*, *comprar* etc.). Moreover, even though Frames are probably not universal concepts by essence, in our view they are likely to be learned and understood across different cultures and languages. And, as they represent basic stereotyped scenarios, they could be used to label complex Highly Iconic Structures, for which no direct mapping to a given oral language lemma can be found. Therefore, we feel that Frames are probably a useful tool for a common SL annotation scheme, not necessarily for glossing individual signed units, but at least as *in situ* metadata.

Therefore, we propose the following annotation tiers as a minimal common annotation scheme:

- text-level and utterance-level segments: START and END of stories, utterances, propositions;
- oral language glosses (e.g. English, French);
- Frame instance and core elements labels: Experiencer, Instrument, Goal, etc. based on the existing Framenet lexicon;
- lexical unit sets associated with Frame instances, as lemmatas for both lexical signs and Highly Iconic Structures.

To our knowledge, these annotations are not standard procedure in SL linguistics, except for glosses. Of course, they are not exclusive of finer-grained descriptions of phonological, morphological, syntactic or rhetorical constructs. But we believe this annotation strategy could overcome the limitations of resorting to lemmatas following Johnston's annotation strategy. Moreover, including such Frame instances and core element labels could provide a common inter-operable indexing strategy, allowing researchers to extract comparable SL corpora segments based on their Frame instance labels, regardless of the particular sign languages or of the structures supporting the Frame instance (lexical sign, HIS).

[8] See (Sallandre, 2003) and (Cuxac and Sallandre, 2007).

In the figures below we give an example of the annotation of the sign GIVE[9] and a Transfer Structure[10] as instances of a GIVING Frame. In the LSF non lexical sign structure, signer Christelle signs "and then she gives her chicks a nice worm" using a complex Transfer Structure combining Situational Transfer (TREE) with Personal Transfer (signer = mother bird) and a clever adaptation of sign GIVE in order to resemble a beak configuration[11]. This example is a clear instance of a whole proposition denoting a GIVING Frame, which cannot easily be mapped into lemma "GIVE". It illustrates the necessity and usefulness of identifying such Frame instances, whether they are expressed with lexical signs or other more complex structures.

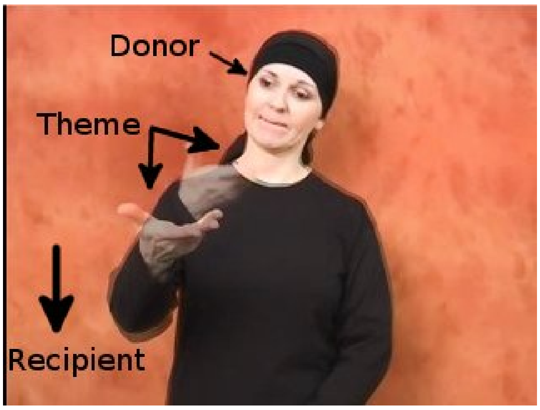

Figure 1: BSL Standard GIVE

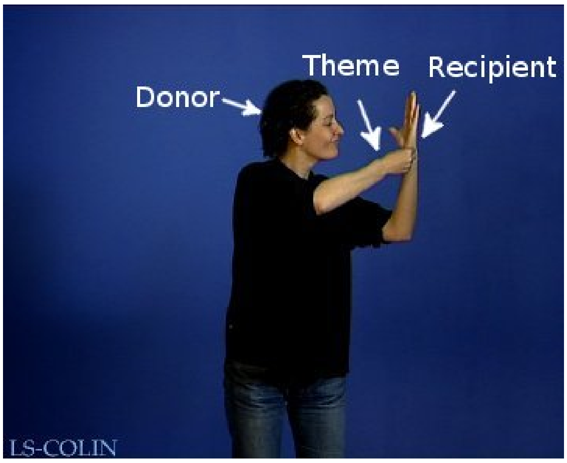

Figure 2: LSF GIVE Transfer Structure

4. Conclusion and perspectives

In this paper, we have outlined a tentative common annotation strategy for SL corpora inspired by Frame-semantics, for the annotation of Frame instances, rather than just lemmatas, regardless of the particular SL or sign structure used. We believe this strategy could provide inter-operable SL corpora, which is crucial for their distribution, exchange and comparison. We include text-level and utterance-level metadata to our proposal, in order to broaden the future uses of the corpora being devised by allowing to derive narrower sub-corpora out of more generic ones.

5. References

A. Balvet, C. Courtin, D. Boutet, C. Cuxac, I. Fusellier-Souza, B. Garcia, M-T. L'Huillier, and M-A. Sallandre. 2010. The Creagest project: a digitized and annotated corpus for French Sign Language (LSF) and natural gestural languages. In *LREC (Language Resources and Evaluation Conference) 2010 Proceedings*, Malta.

D. Boutet, M-A. Sallandre, and I. Fusellier-Souza. 2010. Gestualité humaine et langues de signes : entre continuum et variations. In B. Garcia and M. Derycke, editors, *Langage et Société*, number 131, pages 55–74. Maison des sciences de l'homme.

C.F. Collin, C.J. Fillmore, and J.B. Lowe. 2008. The berkeley framenet project. In *Proceedings of the COLING-ACL*, pages 23–63.

C. Cuxac and M-A. Sallandre. 2007. Iconicity and arbitrariness in French Sign Language: Highly Iconic Structures, degenerated iconicity and diagrammatic iconicity. In E. Pizzuto, P. Pietrandrea, and R. Simone, editors, *Verbal and Signed Languages - Comparing structures, constructs and methodologies*, pages 14–33. Mouton De Gruyter.

C. Cuxac. 1996. *Fonctions et structures de l'iconicité. Analyse descriptive d'un idioloecte parisien de la Langue des Signes Française*. Ph.D. thesis, Université Paris 5.

C.J. Fillmore. 1977. The need for a frame semantics in linguistics. *Statistical Methods in Linguistics*, (12):5–29.

B. Garcia. 2010. *Sourds, surdité, langue(s) des signes et épistémologie des sciences du langage. Problématiques de la scripturisation et modélisation des bas niveaux en Langue des Signes Française (LSF).* Habilitation thesis, Université Paris 8–Saint-Denis.

T. Johnston. 2008. Corpus linguistics and signed languages: no lemmata, no corpus. In *LREC (Language Resources and Evaluation Conference) 2008, Proceedings of the workshop on the representation and processing of Sign Languages*, pages 82–87.

J.B. Lowe, C.F. Baker, and C.J. Fillmore. 1997. A frame-semantic approach to semantic annotation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C. SIGLEX.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, june.

M-A. Sallandre. 2003. *Les unités du discours en Langue des Signes Française (LSF). Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*. Ph.D. thesis, Université Paris 8–Saint-Denis.

G. Sampson. 1994. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press.

[9] BSL, source: Spread The Sign web page, http://www.spreadthesign.com.

[10] LSF, source: LS-Colin corpus, see (Sallandre, 2003) for more details on the LS-Colin corpus.

[11] See (Sallandre, 2003) for detailed transcriptions of HIS structures.