Issues in creating annotation standards for sign language description

Adam Schembri[1], Onno Crasborn[2]
[1]Deafness Cognition and Language Research Centre, University College London
49 Gordon Square, London, WC1H 0PD, United Kingdom
[2] Centre for Language Studies, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: a.schembri@ucl.ac.uk; o.crasborn@let.ru.nl

Abstract

In this paper, we discuss the need for a standardised system of annotation for sign language corpora. Although several tools exist for the annotation of video data (such as ELAN or iLex), and some existing projects have annotation guidelines (e.g., Crasborn et al., 2007; Johnston, 2010), a widely adopted standard is currently unavailable. First, we discuss the purpose of a set of unified annotation standards for sign languages: such standards would provide a shared set of conventions for the easy exchange of data across different sign language corpus projects and may increase consistency within corpora. Next, we discuss the properties that would define a good set of shared annotation conventions (Beckman et al., 2009). We examine some of the proposed annotation standards for spoken language description, such as the ToBI conventions for prosody and the Leipzig Glossing Rules for morpho-syntax. Lastly, we discuss the relationship between theory and description. Dryer (2006) pointed out that linguists often contrast ‘theoretical linguistics’ with ‘descriptive’ work. But if one accepts the argument that there is indeed no ‘atheoretical description’, then sign language linguists need to agree on a shared theory for basic sign language description, and how this translates into annotation practices.

1. Introduction

In this paper, we discuss the need for standardised annotation conventions for the creation of signed language corpora. The paper has come about partly in response to an increasing interest in annotation standards amongst spoken language linguists, as manifested in the report by the annotation standards working group at the 2009 Cyberling workshop (Beckman et al., 2009), as well as among some sign language researchers (e.g., Hermann, 2008; Johnston, 2010). Annotation is used here to refer to written material that is added to, and time-aligned with, the primary sign language digital video data, and represents a description and/or an analysis of the data. Several multimedia annotation tools are currently available (e.g., ELAN, iLex, Anvil, Transana, and SignStream), and are increasingly becoming adopted in the sign language linguistics community. Despite the fact that sign language researchers form a relatively small community of practice and that some projects have made their annotation guidelines available (e.g., Neidle, 2002; Crasborn et al., 2007; Zwitserlood et al., 2008), widely accepted conventions for sign language transcription and annotation are lacking. In the absence of any agreed set of standards, the conventions adopted by the ECHO project[1] have become the basis for some researchers’ annotation guidelines (e.g., Johnston & Schembri, 2006; Herrmann, 2008; Leeson & Nolan 2008), but we feel that the time for wider discussion and dissemination of an agreed set of standards has come.

Note that we are not proposing the widespread adoption of any sign language writing or notation system, nor for a movement away from the increasing use of primary video data in the field: we are focusing here on the use of annotation as means of tagging the primary data and allowing us to create machine-readable corpora.

2. Sign language annotation

Ide and Romary (2004) suggested that there are two fundamental types of annotation activity: (1) *segmentation* and (2) *linguistic annotation*. The first activity consists of identification of the observable elements in the primary data (e.g., signs) using glosses, and should involve some kind of *tokenisation* or *lemmatisation* of the data (Johnston, 2010). The second activity might be further subdivided into at least two subtypes: *syntagmatic* and *paradigmatic* annotation (Beckman et al., 2009). Syntagmatic annotation involves a description of the relationship between the elements identified in the segmentation process (e.g., a noun phrase), while paradigmatic annotation involves the identification of segments as members of particular linguistic classes (e.g., nouns or verbs). Sign language glossing techniques used in the literature often attempt to combine all of these aspects into a single string (e.g., glosses representing signs combined with class labels, such as ‘CL’ for classifier, and superscript lines showing the scope of non-manual markers, such as ‘neg’ for a headshake over a verb phrase).

3. Why do we need sign language

[1] http://www.let.ru.nl/sign-lang/echo

annotation standards?

Annotation of sign language video data serves a number of different functions in corpus sign linguistics, reflecting a researcher's interest in the specific phonetic, phonological, lexical, morphological, syntactic and/or discourse organisation of the data. Often, annotation guidelines are created to serve very specific purposes. In the current British Sign Language (BSL) Corpus Project, for example, a study investigating the linguistic and social factors influencing variation in signs produced with the 1 handshape (the index finger extended from the fist) uses dedicated single character codes for each of the relevant factors, such as the handshape in the preceding sign, or the gender of the signer (Schembri et al., 2009). Annotation conventions will thus always be complemented by project-specific annotations, and are by no means intended to replace these.

The issue of annotation standards becomes more important as opportunities for researchers to share data grow. As Johnston and Schembri pointed out (in press), very few sign language corpora in the modern sense of the term 'linguistic corpus' currently exist (i.e., a representative collection of language samples in a machine-readable form that can be used to study the type and frequency of linguistic units, see McEnery & Wilson, 2001). But many corpus projects are now underway, and this provides the field with a window of opportunity to address the issue of annotation standards. We should begin focussing on the issue of standardised conventions now to ensure that future data exchange between these various projects will be possible, and to provide a basis for future projects. Beckman et al. (2009) suggested that an annotation standard will only succeed if it is associated with a commitment by a community of users to adhere to such conventions. As more and more sign language researchers begin to work on similar issues in corpus sign linguistics, meet regularly in specific workshops and share resources through the Sign Language Linguistics Society[2] and the Sign Linguistics Corpora Network[3], there are now structures in place that can support the development, codification and transmission of annotation standards.

Aside from being able to exchange data between corpora, annotation standards might also encourage consistency within corpora. Good standards will be based on experiences from multiple researchers and research areas and are more likely to have well-developed manuals for annotators or other training methods like dedicated workshops.

4. What are the characteristics of best practice annotation standards?

Beckman et al. (2009) proposed a number of properties as features of 'best practice' annotation standards. First, standards have to be *consistent* and *reliable*. If we look at the history of sign language representation practices in the sign language literature, there are have been few attempts to evaluate the reliability of our means of representing sign language data (such as glossing). This is because there have been few opportunities for sharing primary data, and thus issues around the reliability of particular practices have been avoided. Thus, in order to ensure consistency and reliability for any proposed set of standards, there may be a need to conduct studies into the intra-annotator and inter-annotator reliability rates of any such system, and structures in place that will allow revisions of the standards to be disseminated. Independent validation of a whole corpus is impossible if there is not explicit agreement on the annotation standards that should apply and if these standards are not described in detail.

Second, standards should be *useable*. Any proposed set of conventions must be accompanied by extensive documentation (e.g., reference and training manuals) and perhaps specially-designed annotation software, be relatively easy to teach, should allow the data that has been annotated to be searched used already available query tools, and should comply with the technical demands of a specific annotation tool (e.g. on the text encoding standard to follow).

Third, annotation conventions should be *resilient*. Often there may be uncertainty about how best to annotate some aspect of the primary data, so the standards need clear mechanisms for marking uncertainty about ambiguous cases.

Fourth, standards should be *accountable*. The amount of information contained in the annotations, for example, should stay within the limits of confidentiality agreed to by corpus participants.

Fifth, annotation conventions need *interoperability*: the standards need to be useable within different annotation software packages. They must be clearly related to existing descriptions of the specific linguistic phenomena in the literature, and users should be able to translate the annotation conventions into the terminology used by their own particular theoretical framework.

Lastly, the standards need *extensibility* and *adaptability*. The annotations should be able to be extended to describe new linguistic phenomena in undocumented sign language varieties. There are also need to be practices related to versioning the conventions, so that metadata about which version of the standards are used in particular corpora are available, together with mechanisms for translating across corpora that have been annotated at different stages during the evolution of the conventions.

5. Case studies of spoken language annotation standards

Beckman et al. (2009) review many of the existing standards for annotation for spoken languages. Two examples that illustrate different aspects of the issues involved in the creation of standardised annotation conventions include the Leipzig Glossing Rules and the ToBI Framework.

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

[3] http://www.ru.nl/slcn

5.1 Leipzig Glossing Rules[4]

The Leipzig Glossing Rules[5] are a de facto standard for glossing morphosyntactic phenomena proposed by linguists at the Max Planck Institute for Evolutionary Anthropology and the University of Leipzig. The conventions have emerged out of the typological literature, building on work by Lehmann (1983) and Croft (2003). The rules includes recommendations for best practice with interlinear glosses, such as a requirement for word-by-word alignment of glosses with words, with segmentable morpheme glosses separated by hyphens and fused morphemes represented by glosses separated by periods. Infixes are shown using angled brackets in the gloss, and reduplication shown by a tilde. The rules also include a lexicon of abbreviation conventions for various morphosyntactic categories. These include labels such as 'AGR' for agreement markers, 'OBL' for oblique arguments and 'VOC' for vocative constructions. The rules reflect common usage in the typological literature (and indeed some of the practices and labels will be familiar from published sign language research), with only a few innovations proposed.

Documentation consists of a website, with the rules downloadable as a PDF document. Feedback is welcome, with possible revised versions of the rules promised for the future (the current version dates from February 2008), but currently there is little information available about the consistency and reliability of their use. Beckman et al. (2009) suggest that the creation of some software that allowed users to check their annotations for internal consistency would be useful.

5.2 ToBI

Unlike the Leipzig Glossing Rules, the ToBI (Tone and Break Indices) conventions were originally language-specific, intending to work as a set of annotation standards for the description of the prosody and intonation of American English. This has since been extended to other varieties of English and to a number of other spoken languages. Although these different systems share some basic design principles, they are language-specific, as each set of annotation conventions 'must be guided by an inventory of its prosodic and intonation patterns' (Pierrehumbert, 2000: 26)[6].

Nevertheless, the standards to provide a basis for comparing prosodic systems across languages using shared terminology. A ToBI annotation for American English includes six obligatory parts (Beckman et al., 2005): (1) an audio recording, (2) a record of the fundamental frequency contour, (3) an autosegmental transcription of the intonation contour, (4) an representation of each lexical item, (5) a numeric index from 0 to 4 of the perceived degree of juncture after each lexical item, and (6) markers for disfluencies, commentaries and other miscellaneous annotations. Symbols include L and H for low and high tones, with % representing boundaries, and ? for uncertainty about the annotation. The system for English represents a consensus model of intonation and prosody, drawing on common elements in the 80 years of inter-disciplinary basic and applied research into English prosody. ToBI has had considerable development, testing and a history of use since the early 1990s. Documentation includes websites[7] and published articles, and there have been a number of workshops held at international conferences.

6 Theory and sign language description: Implications for sign language annotation standards

An issue that has been clearly stated in the work on prosodic systems (Beckman et al., 2005) and morphosyntax (Dryer, 2006) is that a theory-neutral annotation system is impossible. Beckman et al. (2005) pointed out that even the most widely-accepted annotation standard, the International Phonetic Alphabet (IPA), is based on two strong theoretical claims: that utterances in any spoken language can be divided into basic vowel and consonant segments (rather than taking syllables as the basic smallest unit, for example), and that each spoken language has a limited inventory of speech sounds that are not radically different from the languages on which the IPA was initially based. Dryer (2006) pointed out that linguists often characterise certain work as 'atheoretical', with some researchers, for example, contrasting 'theoretical linguistics' with 'descriptive' work on particular languages or in cross-linguistic typology. But if one accepts the argument that there is indeed no 'atheoretical description', then sign language linguists need to agree on what sort of shared theory we need for basic sign language description, and how this translates into annotation practices. This will be a challenge, particularly in sign language morphology, where, for example, there is a lack of consensus in the field about whether or not signed languages have verb agreement (e.g., Liddell, 2000; Meier, 2002) and verbal classifier systems (e.g., Schembri, 2003; Zwitserlood, 200x).

7 Towards annotation conventions

We can see the beginnings of standardised annotation conventions for sign language corpora in the ECHO project (Crasborn et al., 2007). The ECHO guidelines were the outcome of a pilot project on the creation of open access sign language corpora on the internet, in which researchers from three universities in different countries and with different research interests aimed to establish a set of basic annotation layers that would be of use for various research endeavors in the future. This led to annotation guidelines and a set of short annotated narratives and poetry (Crasborn et al., 2007). The

[4] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

[5] http://www.eva.mpg.de/lingua/resources/glossing-rules.php

[6] See for example http://todi.let.kun.nl/ToDI/home.htm for the Transcription Of Dutch Intonation (TODI).

[7] See for example http://www.ling.ohio-state.edu/~tobi/.

annotation layers (tiers in ELAN) included glossing separately for the left and right hand, some phonetic annotations appended to the glosses, and a selection of articulatorily independent non-manual properties such as broad categories for eye blinks and head movements (Nonhebel et al., 2004a). Separate conventions were created for the annotation of mouth actions (Nonhebel et al., 2004b). These proposals have influenced the work on the Auslan, ISL, NGT and BSL corpus annotation guidelines, as well as those used in more specific cross-linguistic projects (e.g., Herrmann, 2008; Zwitserlood et al., 2008). The ECHO project, and subsequent work by Zwitserlood et al. (2008), for example, proposed that terminology for segmentation and linguistic annotation has to be very general, and these suggestions will serve as a basis for future work. It is not sufficient, however, for single individuals or research groups to propose standardised conventions, as any annotation standards must develop out of some consensus view about what aspects of sign language linguistic theory and description are important.

8. Practical implications for sign language annotation standards

It is clear that the creation of standards will require a substantial effort on the part of the corpus sign linguistics community. The field lacks the long tradition and widespread shared terminology that forms the basis of the Leipzig Glossing Rules for morphosyntax, and has not experienced the widespread movement towards the creation of consensus-based conventions that we see in the ToBI standards. Despites this, current infrastructure in the field would lend itself to the creation and dissemination of any such proposed standards for sign language annotation. Metadata standards for sign language corpus work already exist (Crasborn & Hanke, 2003), for example, and to appear to be gaining acceptance amongst sign language researchers.[8]

There clearly appears to be the need for dedicated funding beyond the current Sign Linguistics Corpora Network to support a project focused on the creation of annotation standards, and the preparation of necessary documentation that can be distributed to potential users. Any annotation-related project would also possibly require studies into intra-annotator and inter-annotator reliability, as well as the creation of computational tools that can increase the reliability of annotators' work. Moreover, the large-scale validation of whole corpora will be dependent on well-documented annotation conventions, and the validation process would be of a higher standard if the annotation can indeed rely on shared standards. Moreover, any such project needs to put into place some kind of institutional framework for the ongoing maintenance of the conventions, to provide training, and to support ongoing revisions of the conventions and of the accompanying documentation.

Finally, it would be a good idea to explore to which extent the standardisation efforts currently encouraged by the pan-European CLARIN project[9] could be employed. This especially holds for the standard data categories that define widely agreed-upon linguistic terms in the ISOcat[10] concept registries. These might contribute to conventions for sign language annotation, while at the same time maintaining strong links with the spoken language research domain.

9. Acknowledgements

This work was supported in part by the Economic and Social Research Council of Great Britain (RES-062-23-082), by the Netherlands Organisation for Scientific Research (NWO 236-89-002), and by the European Research Council (ERC Starting Research Grant awarded to Onno Crasborn).

10. References

Dryer, M.S. (2006). Descriptive theories, explanatory theories, and basic linguistic theory. In F. Ameka, A. Dench & N. Evans (Eds.), *Catching language: Issues in grammar writing*. Berlin: Mouton de Gruyter, pp pp. 207-234.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005) The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic yypology. The phonology of intonation and phrasing*. Oxford: Oxford University Press, pp. 9-54.

Beckman, M., Robinson, S., Churng, S., Corbett, G., Fillmore, C., & Wright, R. (2009). *Annotation standards.* Retrieved on 24 March 2010 from the Cyberling Wiki: http://cyberling.elanguage.net/page/Group+1%3A+Annotation+Standards.

Crasborn, O. & Hanke, T. (2003). *Metadata for sign language corpora*. Online document, http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf

Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, 12(4), pp. 537-564.

Croft, W. (2003). *Typology and universals*. 2nd ed. Cambridge: Cambridge University Press.

Hermann, A. (2008). Sign language corpora and the problems with ELAN and the ECHO annotation conventions. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 68-73.

Ide, N. & Romary, L. (2004). International standard for a linguistic annotation framework. *Journal of Naturaql*

[8] This early standard on sign metadata has recently been re-evaluated at a workshop of the Sign Linguistics Corpora Network, see http://www.ru.nl/slcn.

[9] http://www.clarin.eu

[10] http://www.isocat.org

Language Engineering, 10:3-4, 211-225.

Johnston, T. (2010a). *Guidelines for the annotation of the video data in the Auslan corpus.* Unpublished manuscript, Macquarie University.

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

Johnston, T. & Schembri. A. (in press). Corpus analysis of sign languages. In C. A. Chapelle, (Ed.), *Encyclopedia of Applied Linguistics.* Wiley-Blackwell.

Leeson, L. & Nolan, B. (2008). Digital deployment of the Signs of Ireland Corpus in E-learning. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 112-122.

Lehmann, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica* 16: 199-224.

Liddell, S.K. (2000). Indicating verbs and pronouns: Pointing away from agreement. In K.D. Emmorey & H. Lane (Eds.), *The Signs of Language revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 303-320.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics,* 2nd ed. Edinburgh: Edinburgh University Press.

Meier, R.P. (2002). The acquisition of verb agreement: pointing out arguments for the linguistic status of agreement in sign languages. In G. Morgan & B. Woll (Eds.), *Directions in sign language acquisition.* Amsterdam, John Benjamins, pp. 115-142.

Neidle, C. (2002). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. file://localhost/Online report, http/::www.bu.edu:asllrp:asllrpr11.pdf.

Nonhebel, A., Crasborn, O. & van der Kooij, E. (2004a) Sign language transcription conventions for the ECHO project. Version 9, 20 January 2004. Radboud University Nijmegen. http://www.let.ru.nl/sign-lang/ECHO/docs/ECHO_transcr_conv.pdf.

Nonhebel, A., Crasborn, O. & van der Kooij, E. (2004b). Sign language transcription conventions for the ECHO project. BSL and NGT mouth annotations. Radboud University Nijmegen. http://www.let.ru.nl/sign-lang/ECHO/docs/ECHO_transcr_mouth.pdf

Pierrehumbert, J. (2000) Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and experiment: Studies presented to Gosta Bruce.* Dordrecht, Netherlands: Kluwer, pp. 11-26.

Schembri, A., Fenlon, J., & Rentelis, R. (2009). Sociolinguistic variation in the 1 handshape in British Sign Language. Paper presented at *NWAV 38: The 38th New Ways of Analyzing Variation Conference*, University of Ottawa.

Schembri, Adam. (2003). Rethinking 'classifiers' in signed languages. In K.D. Emmorey (Ed.), *Perspectives on classifier constructions in sign languages*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 3-34.

Zwitserlood, I. (2003). *Classifying hand configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands)*. Utrecht, The Netherlands: LOT.

Zwitserlood, I., A. Özyürek & P. Perniss (2008) Annotation of Sign and Gesture Cross-linguistically. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 185-190.