# ISOcat data categories for signed language resources

Onno Crasborn[1] and Menzo Windhouwer[2]

[1] Radboud University Nijmegen, Centre for Language Studies
[2] Max Planck Institute for Psycholinguistics, Nijmegen
o.crasborn@let.ru.nl, menzo.windhouwer@mpi.nl

As the creation of signed language resources is gaining speed world-wide, the need for standards in this field becomes more acute. This paper describes the role that ISOcat may play in this process, and makes some initial proposals for the Thematic Domain 'sign language' that was introduced in 2010.

## 1 Background and context

Sign languages are unwritten languages world-wide. Recording language use implies making video recordings. Time-consuming manual annotations are necessary to make such recordings machine-readable. Automatic processing by video recognition techniques is not yet available to linguists at this stage, although the developments in this area are promising (e.g. Cooper et al., to appear, Dreuw et al. 2010). Either way, linguists need to agree on how to classify linguistic events in video recordings, whether they are manually labelled or detected by computer algorithms (Schembri & Crasborn 2010). While there has been some discussion and a concrete proposal on sign-specific metadata categories (Crasborn & Hanke 2003), no such concrete proposal has been made for annotation of signing at any level. The meetings of the Sign Linguistics Corpora Network (SLCN; Crasborn 2010) in 2009 and 2010 have established that the time is right for proposals in this area, as the first large sign language corpora with their annotation conventions are being published online and a large number of corpus creation projects are currently underway (e.g. Crasborn et al. 2007, Johnston 2009).

## 2 ISOcat

The ISOcat Data Category Registry (DCR) is the ISO 12620:2009 compliant registry of the ISO technical committee for "Terminology and other language and content resources" (ISO TC 37) (Kemps-Snijders et al. 2008a). The DCR plays a role in various flexible standards developed by this committee. For example, the Lexical Markup Framework (LMF; ISO 24613:2008), a standard for lexical resources, provides a meta model which can be adorned with ISOcat data categories to create an actually usable project or application specific model. But in fact any linguistic

resource (schema) can be linked to data categories, and can thus make the semantics of the elementary descriptors explicit.

The DCR data model (Kemps-Snijders et al. 2008b) distinguishes two major types of data categories: complex and simple categories. Complex data categories have a conceptual domain for which the values can be described in the registry as simple data categories. Each data category has an elaborate specification. The administrative section contains, among others, a technical persistent identifier, a human readable mnemonic, and various types of notes. The descriptive section then contains alternative names, definitions, examples, explanations, et cetera, in various working languages. An English name and definition is mandatory for every data category. Finally there is a linguistic section that contains the conceptual domain of a complex data category, which can be described specifically for various domains and object languages.

Every linguist can register with ISOcat, at http://www.isocat.org/, and use, create and share data categories. The idea is that this grassroots approach will feed a standardised core, a coherent set of data categories reviewed and maintained by a group of international experts. Multiple of these expert groups have been established by ISO TC 37, e.g., one for metadata and one for morphosyntax, and more can be added upon need of the linguistic community. Recently, a thematic domain group for sign languages has been introduced.

Linguists can use the ISOcat web interface, but also a tool-oriented interface is provided. This interface has been used to create a close integration with the ELAN annotation tool, which is used by many linguists to create sign language resources. In this tool, reference to ISOcat data categories can be made in the definition of controlled vocabularies. This makes ISOcat an attractive platform for the development of standards for sign language annotation.

## 3 Sign language-specific terminology

Many data categories that are being developed for spoken languages will also apply for sign languages. However, we see a need for sign-specific standardised terminology in three different areas. First of all, the metadata categories that were proposed in Crasborn & Hanke (2003) were additions to the IMDI metadata set, including actor properties such as ‘hearing status’. These have been further discussed at an international workshop in November 2009, where little need for amendments was established. They form excellent candidates for ISOcat metadata categories, whether in the thematic domain ‘sign language’ or ‘metadata’.

Secondly, there are some general terms relating to sign language and deaf communication that would be good to standardise, so that they can be profitably used in descriptions of sign language resources. They include the term ‘sign language’ itself, making explicit that this is not any type of visual communication, but specifically refers to the natural languages used in deaf communities. Other terms with a similar extent include ‘fingerspelling’, ‘sign-supported speech’, ‘CODA’ (child of deaf adults), and ‘auxiliary sign language’.

The aspect of annotation that has received most discussion in the recent literature is glossing, as there are specific problems inherent in the representation of a signed language in the written code of a spoken language (Johnston 1991, 2010, Ormel et al. 2010). This does not lend itself easily to the formation of ISOcat data categories, as there will be few lexical items that are shared across languages. In many other areas of linguistic description, however, there will be many concepts where enough consensus in the linguistic literature has arisen. This does not necessitate agreement on the correctness of a specific theory: as long as concepts within a theory are sufficiently clear, they can be described as a data category with a specific label and a persistent identifier. For example, the theory of 'grand iconicity' by Cuxac (2000) is not accepted by all linguists as a correct analysis of the grammar of sign languages, yet specific concepts from this theory may be useful for the annotation of sign language data. By using a standard reference to a data category such as 'personal transfer' (which others would characterise as 'role-taking' or 'constructed action'), this can be used in the annotation of videos in a way that is transparent to other researchers. Thus, it will facilitate the automated analysis of the corpus in question.

## 4 How to move forward

It is high time that linguists and language technologists come together within the Thematic Domain Group 'sign language', to make concrete proposals about ISOcat data categories that can then be proposed to the wider community for discussion. The nature of the ISOcat workflow implies that categories that have been created can be used right away, so that discussion need not only take place on the proposals on paper but can also be tested in the development of concrete corpora. For this to become a success, it will be crucial that corpora are actually published and shared. In that way, searches using ISOcat data categories as criteria can pay off directly, and the benefit of standardisation will not just be a long-term ideal.

## References

1. Cooper, H., Holt, B., Bowden, R.: Sign Language Recognition. In: Looking at People: Automatic visual analysis of humans, Part D. Accepted, (to appear)
2. Crasborn, O.: The Sign Linguistics Corpora Network: towards standards for signed language resources.. In: N. Calzolari, K. Choukri, B.Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 457-460Paris: ELRA (2010)
3. Crasborn, O. Hanke, T.: Additions to the IMDI metadata set for sign language corpora. http://www.let.ru.nl/sign-lang/echo/docs/SignMetadata_May2003.doc (2003)
4. Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Bergman, B., Woll, B.: Sharing sign language data online. Experiences from the ECHO corpus. Int. J. Corpus Linguistics 12(4), 535--562 (2007)

5. Dreuw, P., Forster, J., Ney, H.: Tracking Benchmark Databases for Video-Based Sign Language Recognition. Proceedings of ECCV International Workshop on Sign, Gesture, and Activity (SGA), Crete, Greece (2010)
6. International Organization for Standardization: Language resource management — Lexical markup framework (LMF). ISO 24613 (2008)
7. International Organization for Standardization: Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources. ISO 12620 (2009)
8. Johnston, T.: Transcription and glossing of sign language texts: examples from Auslan (Australian Sign Language). International Journal of Sign Linguistics, 2, 3-28 (1991)
9. Johnston, T.: Guidelines for annotation of the video data in the Auslan Corpus. Online ms., Department of Linguistics, Macquarie University, Sydney, Australia (2009)
10. Johnston, T.: From archive to corpus: transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics 15(1): 104-129 (2010)
11. Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.E.: ISOcat: Corralling Data Categories in the Wild. In: European Language Resources Association (ELRA) (ed), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)
12. Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.E.: A Revised Data Model for the ISO Data Category Registry. In: Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE'08), Copenhagen, Denmark (2008)
13. Ormel, E., Crasborn, O., van der Kooij, E., van Dijken, L., Nauta, E., Forster, J., Stein, D.: Glossing a multi-purpose sign language corpus. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, 22-23 May 2010, Malta. Pp. 186-191 (2010)
14. Schembri, A., Crasborn, O.: Issues in creating annotation standards for sign language description. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, 22-23 May 2010, Malta.