Workshop Proceedings

6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel

Language Resources and Evaluation Conference (LREC)
Reykjavik, Iceland, 31 May 2014

Editors and Workshop Organizers

Onno Crasborn	Radboud University, Nijmegen NL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Julie Hochgesang	Gallaudet University, Washington US
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Johanna Mesch	Stockholm University, Stockholm SE

Workshop Programme Committee

Richard Bowden	University of Surrey, Guildford GB
Penny Boyes Braem	Center for Sign Language Research, Basel CH
Annelies Braffort	LIMSI/CNRS, Orsay FR
Christophe Collet	IRIT, University of Toulouse, Toulouse FR
Kearsy Cormier	Deafness Cognition and Language Research Centre, London GB
Onno Crasborn	Radboud University, Nijmegen NL
Svetlana Dachkovsky	University of Haifa, Haifa IL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Stavroula-Evita Fotinea	Institute for Language and Speech Processing, Athens GR
John Glauert	University of East Anglia, Norwich GB
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Alexis Heloir	German Research Centre for Artificial Intelligence, Saarbrücken DE
Jens Heßmann	University of Applied Sciences Magdeburg-Stendal, Magdeburg DE
Julie Hochgesang	Gallaudet University, Washington US
Trevor Johnston	Macquarie University, Sydney AU
Reiner Konrad	Institute of German Sign Language, University of Hamburg, Hamburg DE
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Lorraine Leeson	Trinity College, Dublin IE
Petros Maragos	National Technical University of Athens, Athens GR
John McDonald	DePaul University, Chicago US
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Christian Rathmann	Institute of German Sign Language, University of Hamburg, Hamburg DE
Adam Schembri	National Institute for Deaf Studies and Sign Language, La Trobe University, Melbourne AU
Rosalee Wolfe	DePaul University, Chicago US

Table of contents

Author Index

Editors' Preface

This collection of papers stems from the Sixth Workshop on the Representation and Processing of Sign Languages, held in May 2014 as a satellite to the Language Resources and Evaluation Conference in Reykjavik.
While there has been occasional attention for sign languages at the main LREC conference, the main focus there is on spoken languages in their written and spoken forms. This series of workshops, however, offers a forum for researchers focussing on sign languages. For the fourth time, the workshop had sign language corpora as its main topic. This time, the focus was on any aspect beyond the manual channel. Not surprisingly, most papers deal with non-manuals on the face.
Once again, the papers at this workshop clearly identify the potentials of even closer cooperation between sign linguists and sign language engineers, and we think it is events like this that contribute a lot to a better understanding between researchers with completely different backgrounds.

The contributions composing this volume are presented in alphabetical order by the first author. For the reader's convenience, an author index is provided as well.

We would like to thank all members of the programme committee who helped us reviewing the submissions to the workshop within a very short timeframe!

Finally, we would like to point the reader to the proceedings of the previous workshops that form important resources in a growing field of research:

- O. Streiter & C. Vettori (2004, Eds.) *From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*. [Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon.] Paris: ELRA. Available online at http://www.lrec-conf.org/proceedings/lrec2004/ws/ws18.pdf
- C. Vettori (2006, Ed.) *Lexicographic Matters and Didactic Scenarios.* [Proceedings of the 2nd Workshop on the Representation and Processing of Sign Languages. 5th International Conference on Language Resources and Evaluation, LREC 2006, Genova.] Paris: ELRA. Available online at http://www.lrec-conf.org/proceedings/lrec2006/workshops/W15/Sign_Language_Workshop_Proceedings.pdf
- O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood (2008, Eds.) *Construction and Exploitation of Sign Language Corpora.* [Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation, LREC 2008, Marrakech.] Paris: ELRA. Available online at http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf
- P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz & A. Schembri (2010, Eds.) *Corpora and Sign Language Technologies*. [Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. 7th International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta.] Paris: ELRA. Available online at http://www.lrec-conf.org/proceedings/lrec2010/workshops/W13.pdf
- O. Crasborn, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Kristoffersen, J. Mesch (2012, Eds.) *Interaction between Corpus and Lexicon*. [Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages. 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey.] Paris: ELRA. Available online at http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf

The Editors

Mouth features as non-manual cues for the categorization of lexical and productive signs in French Sign Language (LSF)

Antonio Balvet[1], Marie-Anne Sallandre[2]
[1]Université Lille Nord de France, F-59000 Lille, France
UdL3, STL, F-59653 Villeneuve d'Ascq, France
CNRS, UMR 8163
Domaine Universitaire du Pont de Bois 59653 Villeneuve d'Ascq
[2]UMR SFL 7023 et Université Paris 8
UPS Pouchet, 59 rue Pouchet, 75017 Paris
[1]antonio.balvet@univ-lille3.fr, [2]marie-anne.sallandre@univ-paris8.fr

Abstract

In this paper, we present evidence from a case study in LSF, conducted on narratives from 6 adult signers. In this study, picture and video stimuli have been used in order to identify the role of non-manual features such as gaze, facial expressions and mouth features. Hereafter, we discuss the importance of mouth features as markers of the alternation between frozen (Lexical Units, LU) and productive signs (Highly Iconic Structures, HIS). Based on qualitative and quantitative analysis, we propose to consider mouth features, *i.e.* **mouthings** on the one hand, and **mouth gestures** on the other hand, as markers, respectively, of LU versus HIS. As such, we propose to consider mouthings and mouth gestures as fundamental cues for determining the nature, role and interpretation of manual signs, in conjunction with other non-manual features. We propose an ELAN annotation template for mouth features in Sign Languages, together with a discussion on the different mouth features and their respective roles as discourse and syntactic-semantic operators.

Keywords: LSF, mouth features, productive signs

1. Introduction

Non-manual features are an integral facet of sign languages (SL). Their relevance has been stressed by different authors, from different theoretical and descriptive backgrounds: (Boyes Braem, 2001), (Boyes Braem & Sutton-Spence, 2001), (Ebbinghaus & Hessmann, 2001), (Fontana, 2008) and (Sutton-Spence, 2007) to name but a few.

In this paper, we present evidence from a LSF case study, with narratives from 6 adult signers. In this study, picture stimuli (the Horse Story) as well as video stimuli (Tom & Jerry cartoons) have been used in order to identify the role of non-manual features such as gaze, facial expressions and mouth features. Hereafter, we will discuss the importance of mouth features as markers of the alternation between **Lexical Units** (LU) and **Highly Iconic Structures** (HIS),[1] also called **Productive Signs** (Johnston & Schembri, 2007) and **Classifier Constructions** in the literature (Emmorey, 2003).[2]

Based on our qualitative and quantitative analysis, we propose to consider mouth features, *i.e.* **mouthings** on the one hand versus **mouth gestures** on the other hand, as markers, respectively, of LU versus HIS.[3] As such, we propose to consider mouthings and mouth gestures as fundamental cues for determining the nature, role and interpretation of manual signs, in conjunction with other non-manual features. We propose a typology of mouth features, together with an ELAN annotation template for such non-manual features. Finally, based on our corpus and on the model presented in (Cuxac, 2000), (Sallandre, 2003) and (Garcia & Sallandre, 2014), we propose a discussion of the role of mouth features found in HIS as predicate modifiers.

2. Terminology and concepts for the study of SLs from a semiological perspective

In this section, we provide terminological and conceptual elements for the study of SLs from a semiological perspective. These elements are given so as to overcome terminological divergences stemming from different traditions in the study of SLs. A more detailed account of equivalences and discrepancies between the semiological model and other approaches can be found in (Garcia & Sallandre, 2014).

The classes discussed in Table 1 below are restricted to **Transfer Units** (TU) in Cuxac's terminology, *i.e.* non-lexical units. In the semiological approach, TUs are considered as belonging to the overall linguistic system of Sign Languages, on a par with LUs (or frozen signs). They are seen as the manifestation of an illustrative intent, which aims at conveying meaning "by showing", whereas the use of frozen signs falls under a non-illustrative intent, where meaning is conveyed "without showing".

1 See (Cuxac, 2000) and (Cuxac & Sallandre, 2007) for a thorough presentation of the semiological model.

2 These distinctions are further discussed in section 2 below.

3 In other words, we consider mouthings as markers of frozen signs, whereas mouth gestures are associated with Productive Signs and Classifier Constructions.

With this crucial dichotomy in mind (illustrative *vs.* non-illustrative intent), the distinctions proposed by C. Cuxac's model can be seen as elaborations on categories used in the literature. This model also provides a sound and consistent framework for categories which are still debated in SL linguistics, thanks to its semiological foundation.

SL Structures in the general literature	*SL Structures in the Semiological Model*
Frozen or Lexical Signs	Lexical Units
Classifier Handshapes	Proforms
Classifier Constructions	Transfers of Size and Shape (TSS) Situational Transfers (ST)
Role Shifts Constructed Actions	Personal Transfers (PT)
Constructed Dialogues	Personal Transfers with reported speech
Multiple References	Double Transfers (PT + ST)

Table 1: Terminological equivalences and discrepancies in different SL linguistics frameworks[4]

As Table 1 shows, one of the salient features of the semiological model is to provide a uniform and consistent way of classifying seemingly unrelated structures, under the illustrative intent (signing by showing). It should be noted that this model posits a continuum between Highly Iconic Structures (non-lexical units) and Lexical Units. Non-lexical units such as Classifier Constructions, Role Shifts, Constructed Actions, Constructed Dialogues, and Multiple References are thus seen as instances of the general Transfer category, which further distinguishes between core constructions and composed ones: TSS, ST and PT are core constructions. These constructions can, in turn, be combined with each other, as in the case of DT (PT + ST), or in the case of reported speech in a PT mode. In the latter case, the overall PT structure can integrate reported speech realized either with LU, or TU.
Even though this paper is focused on mouth-feature, it is worth noting that in the semiological model, non-manual parameters in general are an integral part of the theory. For example, eye gaze direction (towards hands *vs.* towards the interlocutor) is a very salient indicator of the LU *vs.* TU boundary.

3. Mouthings and mouth gestures in LSF narratives

(Petitta, Sallandre & Rossini, 2013) present a comparative case study on LSF and Italian Sign Language (LIS), using narratives based on picture as well as video stimuli. The main outcomes of this initial corpus-based study were the following:

- mouth features exhibit similar functions and roles in both LSF and LIS, their overall distribution in the corpus advocates in favour of their being a fundamental aspect of Sign Languages;
- a fundamental distinction can be drawn between **mouthings** and **mouth gestures**.

In the work presented here, we focus on the relationships between mouth features and the different structures proposed by (Cuxac, 2000) and (Sallandre, 2003) for LSF.
Mouthings can be defined as the (semi)articulation of lexical units ("words") from a given spoken language, in conjunction with manual and other non-manual features (gaze, facial expression, body movement).
It should be noted that mouthings serve essentially as non-manual visual cues associated with the manual ones. As such, they sometimes exhibit Gestalt-like properties: the actual fine-grained and complete articulation is not necessary; the most salient or conventionalized aspects of the "word" are sufficient. For example, in LSF, such mouthings will include actual French words, such as in LU CHEVAL + "cheval" (horse).[5]

Figure 1: LU HORSE with associated mouthing [ʃəv] (LS-Colin corpus, the Horse Story), (Cuxac et. al., 2002)

Figure 1 above gives an example of a mouthing associated with a LU.

Mouthings are not restricted to semantic words (Verbs, Nouns) borrowed from spoken languages: they can encompass grammatical words or even (parts of) broader constituents, and they can also be found in association with full-fledged manual signs, as well as other manual units, such as pointings (Figure 2). Mouthings are highly dependent on the particular oral language the signer is most familiar with, therefore, these mouth features are language-specific.

4 Adapted from (Garcia & Sallandre, 2014).

5 Important variations can be observed in mouthings's realizations by different signers: from truncated forms [ʃəv], or [val], to the complete form [ʃəval].

Figure 2: Pointing associated with mouthing [la] (there), (CREAGEST corpus), (Sallandre & L'Huillier, 2011)

Mouth gestures, on the other hand, can be defined as mouth features which are not associated with actual words in any given spoken language. They include motions and actions stemming from the mouth:
puffed cheeks, with or without air expulsion, as in Figure 3;

- expelling air in a "whistling" fashion, with stretched lips;
- vibrating lips while expelling air, with or without vibrating vocal cords;
- moving tongue sideways while signing;
- moving mouth in a downwards fashion, etc.

Figure 3: HIS description (Transfer of Size and Shape) with associated mouth gesture "a big bump is forming on Jerry's head" (CREAGEST corpus), (Sallandre & L'Huillier, 2011)

Contrary to mouthings, mouth gestures appear widely spread throughout sign languages,[6] and are not related to any given oral language.

Alongside these main mouth features, we propose to further distinguish **idiosyncratic mouth features**. These mouth-features are complementary to mouthings and mouth gestures insofar as they seem associated strictly with a subset of LUs, of which they appear to function as incorporated and mandatory parameters. Furthermore, they form a very limited paradigm of mainly plosive consonant-like configurations (*eg.*: [pi], [pæ], [po]), and they do not appear to commute with one another. An example of such an idiosyncratic mouth feature is the lexical unit: TYPIQUE + [pi] (typical of someone or something).[7]

4. The distribution of mouthings, mouth gestures and idiosyncratic mouth features

4.1. Mouthings and Lexical Units

In our study, mouthings appeared clearly associated with Lexical Units: actual mouthings appear in the context of a LU being signed; they can also appear in non-LU contexts: with pointings and Personal Transfers with reported speech.

Mouth Features	*S1*	*S2*	*S3*	*S4*	*S5*	*S6*	*Avg.*
Mouthing	27	33	24	22	35	21	27
Mouth gesture	13	16	25	28	14	38	22
Nothing	54	44	49	49	45	38	46
Idio-syncratic	2	1	3	2	5	2	2
Unsure	3	6	0	0	2	1	2
Total	**100**	**100**	**100**	**100**	**100**	**100**	**100**

Table 2: Distribution of mouth features in the Tom & Jerry cartoon retelling sub-corpus in LSF by six signers

Table 2 above gives an overall view of the distribution of mouth patterns in the Tom & Jerry data, by six different signers. The main conclusions we can draw from our observations are the following:

- most of the time (46%), units are realized without any mouth feature whatsoever;
- when mouth features are realized, they divide almost evenly between mouthings (27%) and mouth gestures (22%);
- idiosyncratic mouth features are marginal.

Other quantitative elements from our study indicate a clear association between lexical units and mouthings, with the added parameter of text grammar: as mentioned

6 (Petitta, Sallandre & Rossini, 2013) show similar distribution patterns of mouthings and mouth gestures in LSF and Italian Sign Language (LIS).

7 Lexical Units associated with such idiosyncratic mouthings correspond to LSF idioms.

before, our study is based on narratives, not elicited corpora or dialogues, which implies a common overall structure for all narratives. More precisely, mouthings are used for introducing actants (characters in the stories) and new topics,[8] while HIS and other mouth features are used throughout the narratives to elaborate on each actant's behaviour. The only cases of mouthings found in HIS are associated with reported speech.

4.2. Mouth gestures and Highly Iconic Structures

Based on the observations made on the LSF corpus described above, we can posit that mouth gestures are highly correlated with HIS, which are non lexical units. These observations appear consistent with the very foundation of Cuxac's semiological model, which lays the emphasis on the notion of semiotic intent. The clear-cut distinction we have identified in our corpus could be explained in the terms of the semiologic model: mouthings and lexical units belong to the "signing without showing" intent, while mouth gestures and HIS belong to the "signing by showing" one.

4.3. Annotating mouth features with ELAN

In this section, we discuss an annotation template for mouth features for ELAN, a multimodal corpus annotation software.[9]

As presented above, mouth features fall into two main categories: mouthings and mouth gestures. Alongside these two main categories, **idiosyncratic mouthings** can also be found. We therefore propose an annotation template for the different types of mouth features mentioned above, which associates the different structures in Cuxac's model.

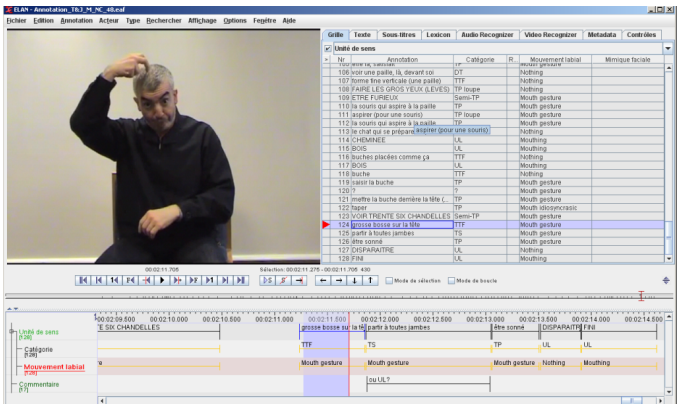


Figure 4: Annotating an HIS+mouth gesture with ELAN

In this template, mouthings and mouth gestures are children of a topmost "Sense Unit" node, and they are further distinguished according to the main dichotomy posited in the model between LU and HIS. Figure 4 shows an annotated example of an HIS associated with a mouth gesture. In the example above, the proposed hierarchy for annotating mouth features is the following:

- Sense Unit
  - Category
  - Mouth feature
    - Mouthing
    - Mouth gesture
    - Idiosyncratic
    - None
    - Unsure
- Comments.

This hierarchy is designed for the specific annotation of mouth-features, but these elements are a part of a more general template (manual/non manual params).

The Sense Units tier holds basic glosses (here in French), which delimit the different units, regardless of their category (LU/TU, pointings, etc.). The Categories tier further specifies the nature of any given Sense Unit: LU, pointing, PT, TSS, DT, etc. Each Category element is symbolically associated to its Sense Unit mother-node, which entails that Sense Units and Categories share the same exact boundaries. Mouth-feature elements are: mouthings, mouth gestures and idiosyncratic elements, as presented above. The mouth-feature tier can also be instantiated by a "Nothing", or an "Unsure" tag. As presented in Table 2, the majority of LUs are realized without any associated mouth-feature, thus making it necessary to annotate both their presence (and nature) and their absence, in order to provide quantitative as well as qualitative elements. In some occasions, annotators are not able to reliably detect mouth features, either for very material reasons (one hand is located in front of the signer's face) or because lip-reading is essentially error-prone. In those cases, dubious mouth features are marked as Unsure. Finally, the Comments tier, which is on the same level as the Sense Unit one, holds all doubts, questions and transient annotation considerations.

The hierarchy outlined above is currently being used in experimental annotations for LSF corpora (both adult and children productions), as presented here, as well as other SL corpora.[10] It was designed so as to be easily integrated into other templates, even ones not in accordance with the semiological model mentioned above.

5. Discussion

We have presented elements from narratives in LSF which show the fundamental role of mouth features as visual cues for the distinction between lexical units and non lexical signs. Based on the distribution observed in our corpus, which appears consistent with the model presented in (Cuxac, 2000) and subsequently updated in (Sallandre, 2003) and (Garcia & Sallandre, 2014), we have proposed a template for the annotation of such non-manual parameters.

8 This distribution is consistent with earlier studies, such as (Sallandre, 2003).

9 Elan is distributed by the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands). It is available at http://tla.mpi.nl/tools/tla-tools/elan/. See (Crasborn & Sloetjes, 2008) for more details on annotating SLs with ELAN.

10 See (Sallandre & Garcia, 2013) for an NGT mouth-features annotation example with the proposed hierarchy.

Mouth-gestures patterns encountered in TSS (Transfers of Size and Shape), a subtype of HIS, are typically used for descriptions: the hands depict the overall shape of a given object, while mouth and facial expression provide information on the dimensions of the object: mouth gestures and facial expression tell us how fine, broad, big, small etc. the depicted object is, according to the signer. As such, they could be considered fundamentally as modifiers, comparable to Adverbs for spoken languages: Figure 3 shows how the "mouth downwards+puffed cheeks" pattern can be associated to a Proform (C handshape) and a facial expression to convey the meaning that, after Tom hit himself with a log (in order to get at Jerry who was stealing cream from him with a straw) a really BIG bump is forming on his head.

Mouth gestures can also be used with PT (Personal Transfers). In these structures, the signer typically embodies the main participant, using his hands and body to describe both manner and path.[11] In these structures, mouth gestures such as the "whistling" pattern, or the "vibrating lips" one can be used to convey aspectual properties on the main predicate: the action can be depicted as swift, or slow, durative, punctual, bounded *vs.* unbounded, or even iterative, with the help of different mouth patterns. As such, mouth gestures not only give fundamental cues as to the nature of the structure being signed (LU *vs.* HIS), but it also can be considered as the equivalent of modality and aspectual markers at the predicative level. It is worth noting that the distribution of mouth features presented in Table 2 for narratives can also be observed among different genres (descriptive, prescriptive), sign languages,[12] as well as in children's productions. Mouthings appear as very salient non-manual cues, marking not only the realization of a Lexical Unit, but also a change of topic or the introduction of actants in narratives. The distinction between mouthings (truncated or complete coarticulation of spoken words in conjunction with LUs) *vs.* mouth gestures seems therefore to open new perspectives for both the manual annotation and automatic processing of sign language structures. From the manual annotation viewpoint, focusing specifically on mouthings and mouth gestures would provide annotators a clear-cut criterion for detecting non-lexical units and topic changes. From the SL automatic processing perspective, mouth-features could boost the automatic recognition of manual signs: if a word contour is detected with some level of confidence, then the current unit is bound to be a Lexical Unit. Conversely, if no word contour can be reliably detected despite a clear mouth movement, then the current unit is likely to be a non-frozen sign, and should therefore be marked as such for later human processing.

11 See (Sallandre *et al.*, 2010).

12 Similar distribution patterns have been observed in (Petitta, Sallandre & Rossini, 2013), a contrastive study on mouth features in LSF and LIS.

6. References

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. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Boyes Braem, P. (2001). The function of the mouthings in the signing of Deaf early and late learners of Swiss German Sign Language (DSGS). In *The hands are the head of the mouth: The mouth as Articulator in Sign Languages*, ed. by P. Boyes Braem and R. Sutton-Spence. Hamburg, Signum Press: pp. 99—132.

Boyes Braem, P., Sutton-Spence, R. (2001). *The hands are the head of the mouth*. Hamburg, Signum Press.

Cuxac, C. (2000). *La langue des signes française (LSF). Les voies de l'iconicité*. Bibliothèque de *Faits de Langues*, n. 15-16. Paris: Ophrys.

Cuxac, C., Braffort, A., Choisier, A., Collet, C., Dalle, P., Fusellier, I., Jirou, G., Lejeune, F., Lenseigne, B., Monteillard, N., Risler, A., and Sallandre, M-A. (2002). Corpus LS-COLIN, http://cocoon.tge-adonis.fr/exist/crdo/meta/crdo-FSL-CUC021_SOUND.

Crasborn, O., Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In: *Proceedings of LREC 2008*, Sixth International Conference on Language Resources and Evaluation.

Cuxac, C., Sallandre, M-A. (2007). "Iconicity and arbitrariness in French Sign Language : Highly Iconic Structures, degenerated iconicity and diagrammatic iconicity." In Pizzuto, E., P. Pietrandrea, R. Simone (eds.): *Verbal and Signed Languages : Comparing Structures, Constructs and Methodologies*. Berlin : Mouton de Gruyter: pp. 13—33.

Ebbinghaus, H., Hessmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In *The hands are the head of the mouth*, ed. by P. Boyes Braem and R. Sutton-Spence. Hamburg, Signum Press: pp. 133—153.

Emmorey, K. (ed.). 2003. *Perspective on Classifier Constructions in Sign Languages*. Mahwah: Lawrence Erlbaum Assoc.

Fontana, S. (2008). Mouth Actions as gestures in sign language, in *Gesture*, 8/1: pp. 104—123.

Garcia, B., Sallandre, M.-A. (2014). Reference resolution in French Sign Language. In Patricia Cabredo Hofherr & Anne Zribi-Hertz (eds.), *Crosslinguistic studies on Noun Phrase structure and reference*. *Syntax and semantics series*, volume 39. Leiden: Brill, pp. 316—364.

Johnston, T., Schembri, A. (2007). *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge: Cambridge University Press.

Liddell, S. K. (2003). *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

Petitta, G., Sallandre, M.-A., and Rossini, P. (2013). Mouth Gestures and Mouthing in Two Sign Languages (LIS and LSF), *Theoretical Issues in Sign Language Research Conference (TISLR 11)*, University College London, DECAL (Deafness, Cognition and Language Research Centre), 10-13 July 2013.

Sallandre, M.-A. (2003). *Les unités du discours en Langue des Signes Francaise. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité*. PhD thesis, Université Paris 8.

Sallandre, M.-A., Courtin, C., Fusellier Souza, I., and L'Huillier, M-T. (2010). « L'expression des déplacements chez l'enfant sourd en langue des signes française ». In Marie-Anne. Sallandre and Marion Blondel (eds), LIA1:1 (*Langage Interaction Acquisition*), Amsterdam : John Benjamins, pp. 41—66.

Sallandre, M.-A., L'Huillier, M.-Th. (2011). *Corpus CREAGEST*. Projet ANR Creagest, sous-projet 1 "Acquisition de la LSF enfantine".

Sallandre, M-A. and Garcia, B. (2013). "Epistemological issues in the semiological model for the annotation of sign language". In Laurence Meurant, Aurélie Sinte, Mieke Van Herreweghe & Myriam Vermeerbergen (eds.), *Sign Language research, uses and practices, Crossing views on theoretical and applied sign language linguistics* (Sign Language and Deaf Communities), Berlin/Boston : Mouton De Gruyter and Nijmegen: Ishara Press, pp. 159—177.

Sutton-Spence, R. (2007), Mouthings and simultaneity in British Sign Language. In *Simultaneity in signed languages*, ed. by M. Vermeerbergen, L. Leeson, O. Crasborn. Amsterdam, Benjamins: pp. 147—162.

Segmenting the Swedish Sign Language Corpus: On the Possibilities of Using Visual Cues as a Basis for Syntactic Segmentation

Carl Börstell, Johanna Mesch, Lars Wallin

Dept. of Linguistics, Stockholm University
S–106 91 Stockholm, SWEDEN
calle@ling.su.se, johanna.mesch@ling.su.se, wallin@ling.su.se

Abstract

This paper deals with the possibility of conducting syntactic segmentation of the Swedish Sign Language Corpus (SSLC) on the basis of the visual cues from both manual and nonmanual signals. The SSLC currently features segmentation on the lexical level only, which is why the need for a linguistically valid segmentation on e.g. the clausal level would be very useful for corpus-based studies on the grammatical structure of Swedish Sign Language (SSL). An experiment was carried out letting seven Deaf signers of SSL each segment two short texts (one narrative and one dialogue) using ELAN, based on the visual cues they perceived as boundaries. This was later compared to the linguistic analysis done by a language expert (also a Deaf signer of SSL), who segmented the same texts into what was considered syntactic clausal units. Furthermore, these segmentation procedures were compared to the segmentation done for the Swedish translations also found in the SSLC. The results show that though the visual and syntactic segmentations overlap in many cases, especially when a number of cues coincide, the visual segmentation is not consistent enough to be used as a means of segmenting syntactic units in the SSLC.

Keywords: Swedish Sign Language, corpus linguistics, segmentation, nonmanuals

1. Introduction

1.1. Segmenting sign language

Previous studies have shown that nonmanual markers in sign language (e.g. eye blinks, eyebrow movement, gaze, head movement etc.) readily have syntactic functions, but they also have prosodic functions (see Pfau and Quer (2010) for an overview). For instance, Wilbur (1994) argues that eye blinks can mark units with a variety of different functions in American Sign Language (ASL), such as syntactic, prosodic, discourse, and narrative units. Nonmanual markers have been shown to work in complex patterns, and boundaries between prosodic units are often aligned with those between syntactic units, with boundaries usually characterized by a change in several of the nonmanual features (Nespor and Sandler, 1999).

Other studies have tried to investigate the possibility of using prosodic and/or intonational information as a means of reliably segmenting certain linguistic units in sign language (e.g. clauses or sentences). A small-scale study on Auslan[1] compared the alignment of so-called *Intonation Units* (cf. Chafe (1994)) with syntactic units, and found that IUs often align with a single clause, although there are also cases of multiple IUs within a single clause as well as a single IU spanning multiple clauses (see Ferrara (2012) for a summary). In another study, Fenlon et al. (2007) found that signers and non-signers alike accurately identify sentence boundaries in sign language texts, and because several visual cues can coincide with each other, some boundaries were stronger (i.e. more visual cues coinciding) than others. This was shown to be true when the subjects viewed both a familiar (British Sign Language (BSL)) and an unfamiliar one (SSL). However, a study on German Sign Language (DGS) investigated whether certain formal boundary markers accurately coincide with sentence boundaries, but found that though the markers often align with the boundaries, they do so in a non-consistent and non-exclusive fashion (Hansen and Heßmann, 2007).

For corpus purposes, the idea of having syntactically segmented sign language texts is still under investigation. The annotation guidelines for the Auslan Corpus use the label *clause-like unit* (CLU) as a tentative equivalent of a ”potential clause”, corresponding to more traditional types of clause units as well as segments containing sign language specific strategies of describing or ”showing” meaning (Johnston, 2013, 50–51).

1.2. The Swedish Sign Language Corpus

The Swedish Sign Language Corpus—henceforth SSLC—consists of a collection of video recordings of pairs of Deaf signers, spanning various text types, e.g. semi-spontaneous dialogues, narratives, and elicitation tasks. The SSLC consists of approximately 25 hours of video data comprising 42 different signers (male and female; ages 20–82). The recordings have accompanying ELAN annotation files, which are being published as they are produced (see Mesch et al. (2014) for the current version). The ELAN annotation files currently consist of six tiers: four for sign glosses (two tiers for each signer; one for each of a signer’s hands), and two for written Swedish translations (one for each signer). Signs are annotated in individual cells with glosses corresponding to Swedish translations of each sign, with additional suffixed tags for some types of signs, such as fingerspellings or gestures (see Wallin and Mesch (2014) for the current guidelines for annotating SSL). To date, the SSLC features annotation files (with

[1] Australian Sign Language

glosses and translations) for about 19% of the total amount of recorded data (Mesch et al., 2012). However, there is no segmentation done above the lexical level, i.e. cells for individual signs, which complicates data exporting and concordance viewing by not being searchable within syntactic units (e.g. clauses or sentences).

The addition of a clausal/sentential segmentation is a natural first step toward analyzing—and annotating—e.g. semantic roles or syntactic functions. Automatic annotation of such categories would be facilitated by an existing linguistic segmentation. A first attempt at an automated induction of word classes was done using the segmentation for Swedish translations as utterances (Sjons, 2013), but a segmentation done independently of another language should prove more viable.

2. Methodology

2.1. Data

For the experiment, we selected two separate texts from the SSLC data: one narrative text, 1:35 minutes long; and one dialogue text, 2:08 minutes long. In the narrative text, only one signer in the pair is signing, with the other signer acting as a receiver of the signing, in the dialogue text, both signers are signing.

2.2. Experiment

The experiment consisted of two parts: in the first part, a number of Deaf signers individually segmented sign language texts based on visual cues; in the second part, a SSL expert segmented the same texts based on syntactic information. The segmentations were then layered on top of each other, together with the pre-existing glosses and translations. From this, we analyzed the data with regard to the number of boundaries marked within and across participants, as well as the amount of overlap in the alignment of the identified boundaries across participants and the syntactic segmentation. The two parts of the experiment are further described below.

2.2.1. Visual segmentation

For the first part, we asked seven Deaf signers (three female) to participate in a segmentation task on the two selected texts. The participants (labeled A–G) have all completed higher education in sign language linguistics. The participants first saw the video on a computer screen, and were then asked to segment the text into units based on the visual cues they interpreted as boundaries. The participants were asked to mark any occurrence of a boundary by pressing a key on the keyboard, but they were allowed to stop the video in order to go back in the video and mark the boundary's exact location. The test was conducted directly in the ELAN window, but all other annotation tiers were hidden during the experiment session. The experiment was run twice for the dialogue data, in order for the participants to segment the text in two tiers—one for each of the two signers in the video.

2.2.2. Syntactic segmentation

For the second part, we had a Deaf sign language researcher analyze the two texts and segment the texts according to available linguistic information regarding semantics and syntax, thus marking segments that semantically and/or syntactically could constitute a clause (although short turns or individual feedback signs would also be segmented as separate units in the dialogue text).

2.2.3. Analyzing overlaps

After the tiers from all participants and the language expert were layered on top of each other, it was clear that there was a certain amount of overlap across participants and the syntactic segmentation. However, in order to establish which segmentations should be grouped together—especially in sequences with several close consecutive segmentations—some criteria had to be defined. We assumed a point of overlap if (1) any two segment boundaries were less than 300 ms apart, and (2) the total span of the overlap point was less than 1000 ms. The latter criterion was quite generous, but deemed necessary in order to allow for cases of longer holds/pauses where some participants marked the end of the previous sign as a boundary, some in the middle of the hold/pause, and others at the beginning of the following sign. If a participant had marked two adjacent boundaries very close to each other such that both were within the scope of a single point of overlap, one boundary was included in the overlap and the other was excluded from it.[2]

3. Results

3.1. Number of segmentations

The results show that the participants differ on the number of segmentations they made for both texts. The number of boundaries marked range from 19 to 52 (average 39.7) for the narrative text, and 44 to 109 (average 73.7) for the dialogue text. Table 1 shows the number of boundaries marked by each participant as well as the boundaries in the syntactic segmentation made by the language expert (labeled "LE").

Participant	Narrative	Dialogue
A	45	76
B	32	79
C	52	109
D	19	44
E	49	62
F	41	85
G	40	61
LE	51	97

Table 1: Number of marked boundaries per participant.

Looking at the actual video with the aligned segmentation tiers, it is clear that while some participants (e.g. participant C) have segmented many single nonmanuals (e.g. a single eye blink), others have focused on the more extensively marked segments (e.g. eye blink, nod, and hands hold/drop at once). Comparing the participants' segmentation to the syntactic segmentation, there is one obvious difference in the task that was given to the participants compared to that

[2] If a participant had made a double segmentation with an interval of <200 ms it was counted as a single boundary.

of the researcher: while the participants task was based on marking the presence of a boundary with the help of visual cues, the researcher segmented cells that contain a syntactic clausal unit. Thus, in the researcher's segmentation, the cells are sometimes interspaced by a pause in the signing, but the boundaries marked by the participants are usually punctual and often coincide with either the start or the end of such an interspace (i.e. marking boundaries either at the end of one unit or the start of another). An example of this is given in figure 1 below, in which Signer 1 in the dialogue text puts his hands together for a short pause after a sign.

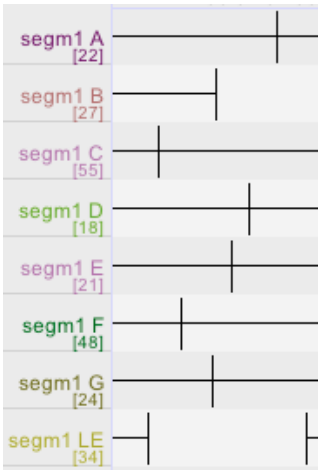


Figure 1: Differences in boundary markings.

As figure 1 illustrates, some of the participants (e.g. C and F) chose to mark the retraction of the hands as the boundary, whereas others (e.g. A and D) chose to mark the middle of the pause as the boundary. Since the language expert (LE) only annotated signing, the pause (about 1000 ms long) is thus marked by the absence of a cell. This difference in tasks could be one explanation as to why the number and placement of segmentations may vary between the language expert and the participants.

3.2. Amount of overlap

Turning to the points of overlap, it is visible from the data that some locations in the texts are characterized by a high number of overlap across participants and the language expert. Using the criteria described in section 2.2.3., the total number of unique boundaries was 78 for the narrative text, and 167 for the dialogue text. For some of these, only a single participant had marked it as a boundary, but the majority of boundaries are shared with at least one other participant and/or the language expert, thus constituting a point of overlap. Figures 2 and 3 below show the distribution of overlaps across participants, and whether or not these align with the syntactic boundaries made by the language expert, for each of the two texts. The X axis corresponds to the number of participants sharing a point of overlap, and the Y axis corresponds to occurrences. The different colored columns show whether or not the occurrences also overlap with a syntactic boundary.

Figure 2 shows that there is a tendency for the more agreed upon points of overlap to be marked also as a syntactic boundary by the language expert, although there are a couple of cases for which a syntactic boundary has not been marked by a single participant. However, there is a clear idiosyncrasy among (some of) the participants' segmentations, resulting in a very high number of boundaries marked by a single participant.

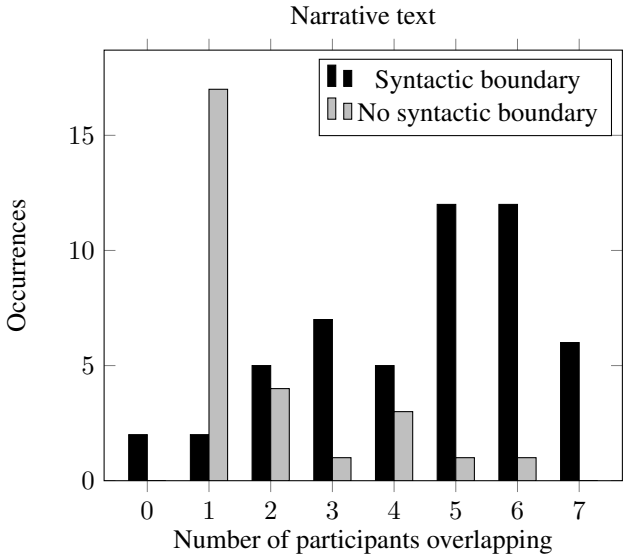


Figure 2: Overlap between visual and syntactic boundaries.

Figure 3 shows a similar pattern, again demonstrating the high amount of idiosyncrasy among the participants resulting in a high number of boundaries identified by a single participant. However, it also shows a global trend of syntactic boundaries being marked most often if many participants also marked it as a boundary, and vice versa.

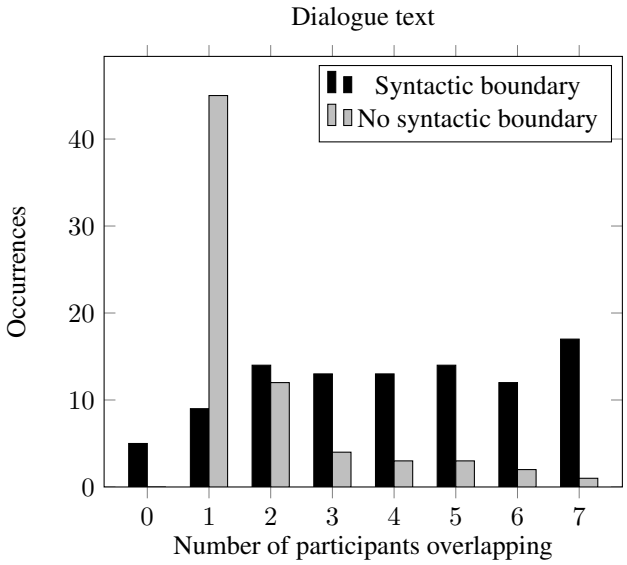


Figure 3: Overlap between visual and syntactic boundaries.

3.2.1. Comparison with the translations

For the narrative text, the translator used for the SSLC—a hearing native signer (i.e. CODA[3])—made a translation without access to any of the segmentations made by either the participants or the language expert. Looking briefly at the alignment of cells in this translation, it is clear that although many of the endpoints overlap with those made in the syntactic segmentation, the number of segmentations is not the same. The translation tier generally has longer segments, often spanning over several syntactic segmentations. While the endpoints of most translation cells do align with the endpoints of some syntactic segmentations, the range of syntactic segmentations within the scope of a single translation cell ranges between 1 and 9 (average 3). Thus, the translation tier cannot be considered an accurate segmentation of the SSLC on a clausal level.

4. Conclusion

This minor study is a first step toward adding a linguistic segmentation to the SSLC, which could prove useful in future work of annotating e.g. word classes or syntactic functions in the corpus. The purpose of the study was two-fold: first, we wanted to see how well different signers' segmentation of sign language texts based on visual cues correspond to each other; second, we wanted to see how well a segmentation based on visual cues corresponds to a segmentation based on a more in-depth analysis of linguistic units.

Our investigation demonstrated that while signers show some agreement in their segmentation of sign language texts based on visual cues, it is not completely reliable as a means of segmenting syntactic units. For instance, the number of segmentations varies across participants doing the same task—segmenting the text based on visual cues—and these do not always align with syntactic boundaries, as suggested by Hansen and Heßmann (2007). When layering the visual segmentations, they do pattern in a way that shows that a high degree of overlap often corresponds to the presence of a syntactic boundary, but using a high number of participants for visual segmentations might prove more time-consuming than a few language experts segmenting the SSLC for syntactic units. Thus, a visual-based segmentation would be neither the most accurate nor the most practical way of adding a syntactic segmentation to the SSLC.

5. Acknowledgments

The Swedish Sign Language Corpus project was funded by a grant from Riksbankens Jubileumsfond (In2008-0276-1-IK).

6. References

Chafe, Wallace. (1994). *Discourse, Consciousness, and Time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, Chicago, IL.

Fenlon, Jordan, Denmark, Tanya, Campbell, Ruth, and Woll, Bencie. (2007). Seeing sentence boundaries. *Sign Language & Linguistics*, 10(2):177–200.

Ferrara, Lindsay. (2012). *The grammar of depiction*. Ph.D. dissertation, Dept. of Linguistics, Macquarie University.

Hansen, Martje and Heßmann, Jens. (2007). Matching propositional content and formal markers: Sentence boundaries in a DGS text. *Sign Language & Linguistics*, 10(2):145–175.

Johnston, Trevor. (2013). Auslan Corpus Annotation Guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University.

Mesch, Johanna, Wallin, Lars, Nilsson, Anna-Lena, and Bergman, Brita. (2012). Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University.

Mesch, Johanna, Rohdell, Maya, and Wallin, Lars. (2014). Annotated files for the Swedish Sign Language Corpus. Version 2. Sign Language Section, Department of Linguistics, Stockholm University.

Nespor, Marina and Sandler, Wendy. (1999). Prosody in Israeli Sign Language. *Language and Speech*, 42(2-3):143–176, June.

Pfau, Roland and Quer, Josep. (2010). Nonmanuals: their grammatical and prosodic roles. In Brentari, Diane, editor, *Sign languages: A Cambridge language survey*, pages 381–402. Cambridge University Press, New York, NY.

Sjons, Johan. (2013). Automatic induction of word classes in Swedish Sign Language. M.A. thesis, Dept. of Linguistics, Stockholm University.

Wallin, Lars and Mesch, Johanna. (2014). Annotation guidelines for sign language texts. Version 5. [In Swedish]. Sign Language Section, Department of Linguistics, Stockholm University.

Wilbur, Ronnie B. (1994). Eyeblinks & ASL Phrase Structure. *Sign Language Studies*, 84(1):221–240.

[3]Child of Deaf Adult

Synthesizing Facial Expressions for Sign Language Avatars

Yosra Bouzid, Oussama El Ghoul, Mohamed Jemni
Research Laboratory of Technologies of Information and Communication & Electrical Engineering
University of Tunis
yosrabouzid@hotmail.fr, oussama.elghoul@rnu.tn, mohamed.jemni@fst.rnu.tn

Abstract

Sign language is more than just moving the fingers or hands; it is a visual language in which non-manual gestures play a very important role. Recently, research has paid increasing attention to the development of signing avatars endowed with a set of facial expressions in order to perform the actual functioning of the sign language, and gain wider acceptance by deaf users. In this paper, we propose an effective method to generate facial expressions for signing avatars basing on the physics-based muscle model. The main focus of our work is to automate the task of the muscle mapping on the face model in the correct anatomical positions and the detection of the jaw part by using a small set of MPEG-4 Feature Points of the given mesh.

Keywords: facial expressions, signing avatars, MPEG-4, feature points

1. Introduction

Thanks to the advances in virtual reality and human modeling techniques, signing avatars have become increasingly common elements of user interfaces for a wide range of applications such as interactive e-learning environments and mobile phone services, with a view to improving the ability of hearing impaired people to access information and communicate with others. In order to ensure maximum comprehension and clarity to these signers, digital humanoids are required to perform not just broad hand movements, but also many subtle clues and features like face movements and expressions, which must be clearly seen in order to understand the meaning. Eyebrow height, mouth shape, and other facial gestures are linguistically required in sign language, and identical hand movements can have different meanings depending on the facial expressions performed during the sentence (Neidle et al., 2000).

Different approaches have been taken to animate a three dimensional synthetic human face, but most require a significant effort and time-consuming to adjust animation parameters. For example, the process of rigging requires many hours of manual work to set up the bone structure for an entire face. Even simple method like shapes blending needs the intervention of an artist to create large libraries of key shapes. On the other hand, the production of expressive and realistic animations involves a high computational complexity for simulating the physical property of the underlying facial structure which includes the skeletal, different muscles forms and the subcutaneous fatty tissues.

To deal with these problems, we present in this paper an effective method capable of deforming a 3D mesh of an arbitrary synthetic human face to generate emotional expressions without considerable amount of manual intervention and artistic skill. This approach relies on the physics-based muscle model proposed by Waters (1987) to emulate the contraction of the muscle onto the skin surface. Our contribution consists essentially of the automatic construction of mimetic muscles as well as the jaw mesh detection using only MPEG-4 feature points of the given mesh.

2. Background

This section presents a brief description of the most popular techniques and approaches which are generally utilized in 3D facial animation today.

2.1 Blend Shapes

The blend shape animation method, also known as morph target animation or shape interpolation is the most intuitive and commonly used technique in the field since it is quite straightforward and easy to accomplish. The basis for this method is that during the animation, the interpolated facial model is created from a specific set of key facial poses called blend shapes through interpolation over a normalized time interval. Typically, a blend shape model is the linear weighted sum of a number of topologically conforming shape primitives. Varying the weights of this linear combination allows the representation of facial motions with little computation. However, it is important to note here that the generation of a significant range of highly detailed expressions usually implies the creation of large libraries of blend shapes which can be very time-consuming. Moreover, if the topology of the model needs to be changed, all the shapes must be redone (Ping et al., 2013).

2.2 Facial Rigging

Rigging is the process of setting up a group of controls to operate a 3D model, analogous to the strings of a puppet. It plays a fundamental role in the animation process as it eases the manipulation and editing of expressions, but rigging can be very laborious and cumbersome for an artist. This difficulty arises from the lack of a standard definition of what a rig is and the multitude of approaches on how to set up a face (Orvalho et al., 2012).

2.3 Parameterization

In parameterization, facial geometry is broken into parts where each part is exposed properly to its parameter sets or control points. This allows the animators to have control of the facial configurations (Ping et al., 2013). By combining different parameters, a large range of facial expressions can be produced. Facial Action Coding System (Ekman & Friesen, 1977) and the MPEG-4 Facial Animation standard are the most famous parametrizations that can be included in this category. The advantage of these approaches is that once control parameters are determined, they provide a detailed control over the face. But determining this is hard: complexity of creating an animation with these control parameters is related to the number of control parameters, as is the possible range of expressions (Ilie et al, 2012). Furthermore, the animation does not seem to respond to basic physical deformations of human faces since direct parameterizations make no attempt to represent the detailed anatomical structure. They model only the changes visible on the skin surface.

2.4 Physics-based muscle Modelling

Physics-based muscle methods have the potential to produce natural 3D animations by precisely simulating the real effects of the facial muscular tissues. They can generally fall into three different categories: mass spring systems, layered spring meshes and vector representations. Mass-spring systems (Platt & Badler, 1981) are designed to propagate muscle forces in an elastic spring mesh that models the elastic properties of the skin, while, layered spring meshes (Terzopoulos & K. Waters, 1990) extend the mass spring structure to three connected mesh layers to simulate the anatomical aspects of the face more faithfully. In vector representations, the actions of muscles upon the skin are modelled using motion fields in delineated regions of influence. The most successful muscle models were proposed by Waters (1987) and Kähler (2007) who proposes a muscle structure composed by quadric segments. Although the physics-based techniques are the most scientifically based, they are also among the most difficult to implement. The construction of the anatomical facial structure is an extremely tedious task which requires artistic skills and massive computation.

3. MPEG-4 Facial Animation

A widely used and validated parametrization for synthetic characters is the one defined inside the MPEG-4 specification, namely MPEG-4 Facial and Body Animation (FBA). Such a standard makes use of three main sets of parameters to specify a face model in its neutral state (Pandzic & Forchheimer, 2002).

The Facial Animation Parameters (FAPs) are used to direct control the face movements. They are based on the study of minimal perceptible actions (MPA) and are closely related to muscle actions, such as movements of lips, jaw, cheeks and eyebrows. They make up a complete set of basic facial actions that represent the most natural facial expressions. FAPs define 68 parameters. The first 2 are high level parameters representing visemes and facial expressions. Viseme is the visual counterpart of phonemes in speech while facial expressions consists of a set of 6 basic emotions for anger, joy, sadness, surprise, disgust and fear as prototypes. The rest of the low level FAPs deal with specific regions on the face, like right corner lip, bottom of chin, left corner of left eyebrow.

The Facial Definition Parameters (FDPs) are needed for the calibration of a synthetic face. These parameters are scalable; they can define the shape, texture or even the whole facial polygon mesh.

The Feature Points (FPs) are used to describe and define the shape of a standard face. There are a total of 84 feature points in a head model. They are subdivided in groups, mainly depending on the particular region of the face to which they belong. Each of them is labelled with a number identifying the particular group to which it belongs and with a progressive index identifying them within the group. A subset of these points can be affected by the facial animation parameters (FAPs) to control the animation.

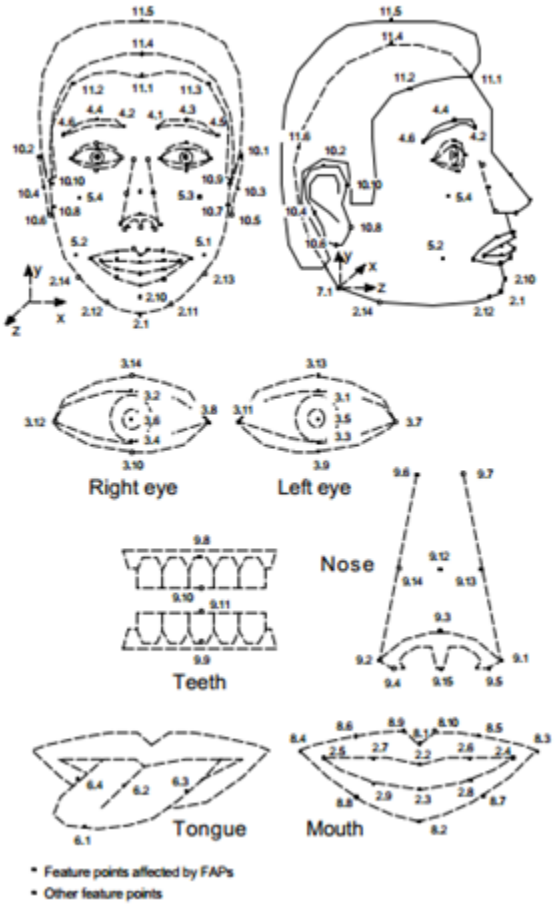


Figure 1: The 84 Feature Points (FPs) defined on a neutral face (Pandzic & Forchheimer, 2002)

4. Approach

Our approach aims to create plausible facial animation with a variety of facial expressions by automatically contracting a group of individual muscles and rotating the jaw mesh. To elaborate the mimic musculature, we have relied on Water's muscle model in which two types of muscles are defined: linear muscles that pull and sphincter muscles that squeeze. The mapping of such musculature to the face model is achieved by identifying the key nodes of each muscle with the appropriate set of MPEG-4 features points. The selection of FPs is done according to the anatomical properties that characterize the given muscle. In fact, the use of MPEG-4 features points as key nodes of each muscle will certainly reduce the amount of work that must be done manually by animators.

4.1 Muscle Modeling

To emulate the behaviour of muscles upon skin, Waters presents one of the most popular and complete parametric muscle models that are based on the human facial anatomy. This model is computationally cheap and easy to implement, it includes two types of muscles, linear and sphincter, independent of the bone structure. Each of these muscles can be defined by two key nodes, an area of influence which presents a skin portion affected by the contraction, and a deformation formula for all influenced vertices.

The linear muscle is modeled as a vector from a bony attachment point that remains static, to an insertion point which is embedded in the soft tissue of the skin. Its influence area is represented by a cone shape (Fig.2). For example the Zygomaticus Major which acts to draw the angle of the mouth up and back to smile or laugh, is a linear muscle.

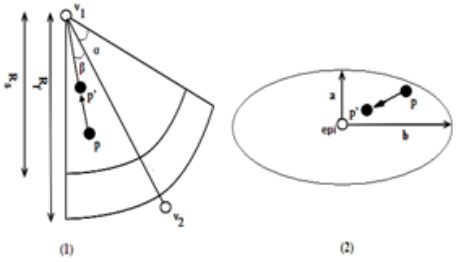


Figure 2: Linear and sphincter muscles

When a linear muscle contracts, all points in its influence area are displaced towards its point of attachment. The displacement of a point p affected by the muscular action is given by the Equation 1 (k is a fixed constant representing the elasticity of the skin).

$$p' = p + a\,r\,k\,\overrightarrow{pv_1} \quad (1)$$

$$a = \frac{(\cos\beta - \cos\alpha)}{(1 - \cos\alpha)}$$

$$r = \begin{cases} \cos\left(\frac{1 - \|v_1p\|}{R_s} \cdot \frac{\pi}{2}\right), \|v_1p\| < R_s \\ \cos\left(\frac{\|v_1p\| - R_s}{R_f - R_s} \cdot \frac{\pi}{2}\right), \quad else \end{cases}$$

For sphincter muscles, we can identify only two instances in a human face: the Orbicularis Oculi muscle around each eye and the Orbicularis Oris which circles the mouth. This kind of muscle attaches to the skin both at the origin and at the insertion. Its influence area has an elliptical shape defined by a virtual center and two semi-axes (Fig. 2). When a sphincter muscle contracts, the points in its influence area are displaced towards the center of the spheroid. The displacement of a point p affected by the action of muscle is given by the Equation 2 (k is a fixed constant representing the elasticity of the skin).

$$p' = p + r\,k\,\overrightarrow{po} \quad (2)$$

$$r = \cos\left(\left(1 - \frac{\sqrt{p_x^2 b^2 + p_y^2 a^2}}{ab}\right) \cdot \frac{\pi}{2}\right)$$

It is important to note that Waters combined the muscle actions sequentially by applying the displacements caused by them on a vertex one by one. However, such a process can produce undesirable effects especially when a mesh vertex is under the influence of multiple muscle actions: the vertex will be shifted outside the influence area of adjoining muscle vectors. To avoid the undesired effects, we have used the Wang approach (Wang, 1993) which summarizes the displacements and then applies it to the vertex.

4.2 Muscle Construction

Our facial musculature comprises essentially 31 muscles including three sphincter muscles that are used to represent the orbicularis oris and orbicularis oculi, and 11 pairs of linear muscles that are placed symmetrically through the face to accomplish the major face movements. The remaining linear muscles, namely Cheek Sup, Cheek Center and Cheek Inf, are located on each cheek and don't exist in a real human face. They have been added to our model to simulate specialized expressions in sign languages like cheek movements. The complete facial muscle structure is shown in Fig.3. The face model is represented as a single layered mesh with no skeleton, it is expected to consist of triangular or quadratic polygons, a high-poly or low-poly mesh.

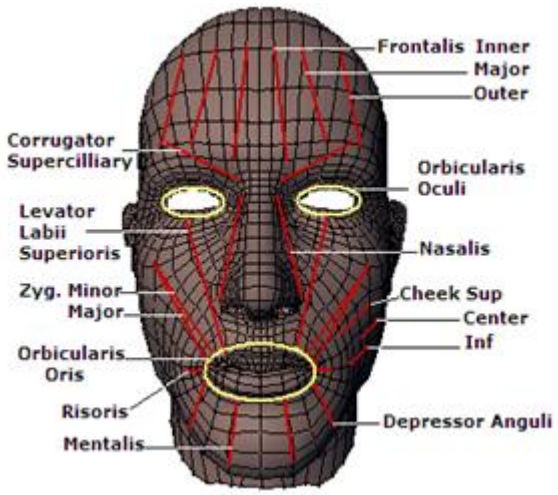


Figure 3: Facial musculature

The construction of the proposed musculature involves two basic steps. At first, the anatomical positions of the muscle control points will be defined with the suitable MPEG-4 Feature Points. Second, the set of vertices that belong to each influence area will be detected.

For a linear muscle, three points are needed to define its location on the input face mesh: an attachment point AP, an insertion point IP and a reference point RP which was not used in Waters model, we have added it in our method to facilitate the determination of the mesh part affected by the muscle action. The obtained properties of muscle vectors are given in Table 1. For instance, the Left Nasalis which depresses the cartilaginous part of the nose is characterized by:

- an AP which coincides with the feature point 9.7 located in the left upper edge of the nose bone
- an IP which coincides with the feature point 9.1 located in the left nostril border.
- an RP which coincides with the feature point 9.3 located in the nose tip .

Muscle name	Attachment point	Insertion point	Reference point
Frontalis Inner	11.1+$\frac{1}{6}$(11.2 -11.1)	4.2	4.2 +$\frac{1}{2}$(4.4 - 4.2)
	11.1+$\frac{1}{6}$(11.3 -11.1)	4.1	4.1 +$\frac{1}{2}$(4.3 - 4.1)
Frontalis Major	11.1+$\frac{2}{3}$(11.2 -11.1)	4.2 +$\frac{1}{2}$(4.6 - 4.2)	4.6
	11.1+$\frac{2}{3}$(11.3 - 11.1)	4.1 +$\frac{1}{2}$(4.5 - 4.1)	4.5
Frontalis Outer	11.2	4.6	4.4
	11.3	4.5	4.3
C.S	4.2 + $\frac{1}{2}$(4.2 - 4.4)	4.4	4.2 +$\frac{1}{2}$(4.2 –3.8)
	4.1 + $\frac{1}{2}$(4.1 - 4.3)	4.3	4.1 + $\frac{1}{2}$(4.1 -3.11)
Nasalis	9.6	9.2	9.3
	9.7	9.1	9.3
Levator Labii Superioris	3.10	2.7	8.9+$\frac{1}{3}$(8.6 – 8.9)
	3.9	2.6	8.10+$\frac{1}{3}$(8.5 -8.10)
Z. Minor	9.2	2.7	8.4
	9.1	2.6	8.3
Z. Major	5.4	2.9	9.15
	5.3	2.8	9.15
Risoris	5.2	2.5	2.5+$\frac{1}{3}$(2.5 – 9.2)
	5.1	2.4	2.4+$\frac{1}{3}$(2.4 – 9.1)
Depressor Anguli Oris	8.4 + (8.4 – 8.6)	8.6	8.8
	8.3 + (8.3 – 8.5)	8.5	8.7
Mentalis	2.1 + $\frac{1}{2}$(2.1 – 2.12)	2.9	8.2
	2.1 + $\frac{1}{2}$(2.1 – 2.11)	2.8	8.2
Cheek Inf	K + $\frac{1}{2}$(K – 2.11)	K= 5.1 + (5.1 – 8.2)	5.1
Cheek Center	K + $\frac{1}{2}$(K – 2.13)	K= 5.1 + (5.1 – 8.2)	5.1
Cheek Sup	K + $\frac{1}{2}$(K – 2.11)	K= 5.1 + (5.1 – 8.3)	5.3

Table 1: Linear Muscle Properties (Right and Left)

Similarly, each sphincter muscle is defined by three key points: the epicenter of the spheroid EP, a semi-major axis SJ and a semi-minor axis SN. The obtained properties of sphincter muscles are shown in Table 2. For instance, the Orbicularis Oculi Left which closes the eyelids of the left eye is characterized by:

- an EP which coincides with the midpoint of the line segment formed by the two points 3.7 and 3.11
- SJ is equal to the length of the line segment formed by the two points 3.7 and EP
- SN is equal to the length of the line segment formed by the two points 3.9 and EP.

Muscle name	Semi-major axis	Semi-minor axis	Epicenter
Orbicularis Oculi	3.12	3.10	$\frac{1}{2}$(3.8+3.12)
	3.7	3.9	$\frac{1}{2}$(3.7+3.11)
Orbicularis Oris	8.3	8.2	$\frac{1}{2}$(8.3+8.4)

Table 2: Sphincter Muscle Properties

Once the positions of muscle control points are computed and mapped on the facial mesh, we can determine then the set of vertices which will be influenced by the muscle contraction. For a linear muscle, all influenced vertices should match the following conditions (see Fig. 2):

$$\|\overrightarrow{pv_1}\| > 0,\ \|\overrightarrow{pv_1}\| \leq \boldsymbol{Rf},$$

$$\boldsymbol{\cos\beta \geq \cos\alpha}$$

Similarly, all influenced vertices of a circular muscle should be within its spheroid (see Fig. 2).

$$\left(\frac{\boldsymbol{p_x - epi_x}}{\boldsymbol{a}}\right)^2 + \left(\frac{\boldsymbol{p_y - epi_y}}{\boldsymbol{b}}\right)^2 + \left(\frac{\boldsymbol{p_z - epi_z}}{\sqrt{\boldsymbol{a^2 - b^2}}}\right)^2 < 1$$

Fig. 4 shows the obtained result after applying the proposed algorithm to the Nasalis muscle. The face on the left illustrates the set of vertices having a distance from AP less than the length of muscle fiber. The second face illustrates the set of vertices that will be displaced when muscle contracts.

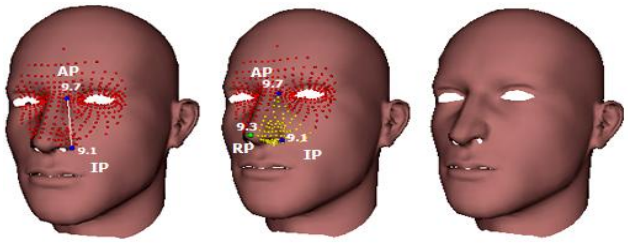


Figure 4: The set of vertices that belong to the influence area of Nasalis Left is colored in yellow

4.3 Jaw Articulation

As we mentioned above, the face model does not have a skull, so the jaw is not a particular mesh, it will be rather detected automatically from the initial mesh. To do so, we have used some features points for approximating the vertices of the chin and lower lip that are affected by the jaw rotation.

4.3.1. Jaw Detection

To define the chin vertices, we have taken the following steps: first, we project the facial mesh on the plane P passing through the midpoint of the segment joining the FPs 10.8 and 10.7, with the normal vector defined by this midpoint and the vertex 10.8 in order to get a profile view. Second, the projections of 2.14, 10.8 and 8.3 respectively p1, p2 and p3 are marked on the projected mesh. The vertices whose projects are inside the angular sector $(p1\widehat{,p2,}p3)$ are considered to be in the chin influence (Fig. 5).

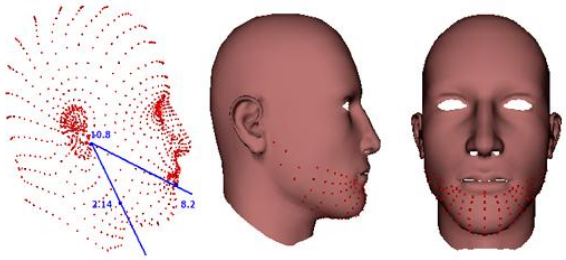


Figure 5: Detecting chin vertices

The process is about the same for the lower lip but by using other feature points. It is important to note here that the vertices located on the inner contour of the lower lip should also be taken into account since the projection is incapable of detecting them. The extraction of these vertices is done as follows: the algorithm will browse all the edges of the mesh to find those that belong only to one surface. The selected edges depict the contours of the facial model such as the openings of the eyes, nose as well as the space between the two lips. Using this set of edges and some MPEG-4 FPs, we can distinguish the inner contour of the lower lip. All we have to do is finding the closest edge to the point 2.3, and then its neighboring segments (Fig. 6). For each new segment found, we perform the same process until finding edges closest to the points 8.3 and 8.4 which define the corners of the lips.

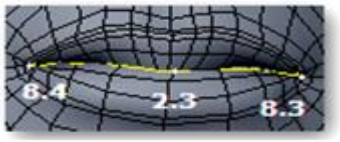


Figure 6: The inner lower lip contour

4.3.2. Jaw Rotation

To rotate the jaw, the vertices of the chin and the lower lip are rotated around a line passing through the feature point 10.8 and parallel to the X axis (Fig. 7). The final positions are calculated by the following equation where φ represents the degrees of rotation and (x1, y1, z1) the coordinates of 10.8.

$$\begin{pmatrix} x \\ y' + y1 \\ z' + z1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix} * \begin{pmatrix} x \\ y - y1 \\ z - z1 \end{pmatrix} \quad (3)$$

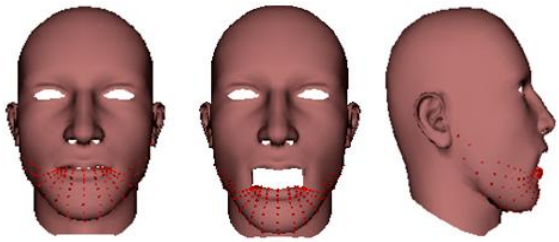


Figure 7: Jaw rotation

5. Results & Evaluation

The production of emotions is the result of a contraction or relaxation of one or more facial muscles. Fig. 8 shows some examples of basic facial expressions (happiness, anger, sadness, surprise, disgust and fear) on different face models, while Fig. 9 illustrates the contraction of Cheek Sup, Cheek Center and Cheek Inf which are used to emulate the tongue motion on the right cheek.

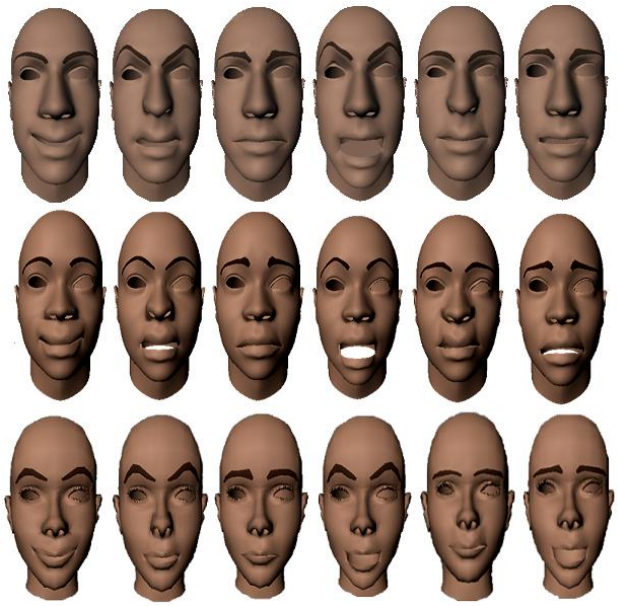


Figure 8: The simulation of some basic facial expressions on different face models

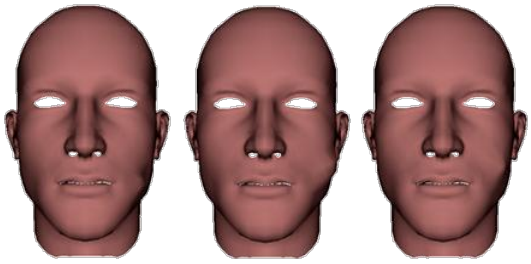


Figure 9: The contraction of cheek Sup, Cheek Center and Cheek Inf

In order to validate our approach, we have tested the performance of facial muscles in different face models. The goal of this evaluation is to check the choice accuracy of features points in the definition of muscle key nodes as well as the fitting of those features with anatomically correct positions on the face. Fig. 9 shows the recognition rate of each muscle motion is calculated with 50 models.

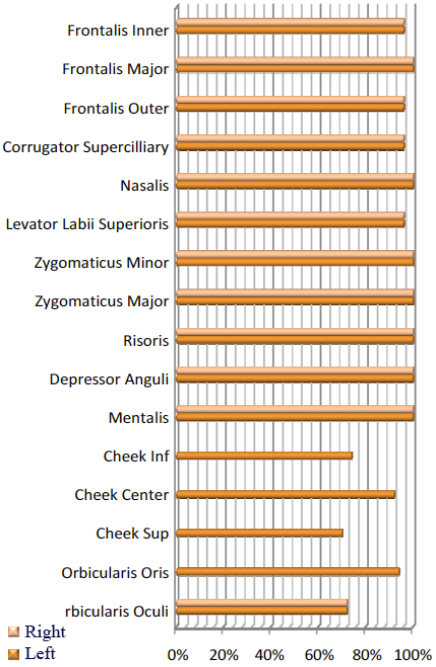


Figure 10: The recognition rates of muscle motion on 50 face models

We can notice that the contraction effects of the vast majority of muscles have been simulated. For the Frontalis Major, Nasalis, Zygomaticus Major, Zygomaticus Minor and Risoris, our approach achieves 100% recognition rate, whereas, the Cheek Inf, Cheek Sup and Orbicularis oculi achieve recognition rates ranging from 70% to 75%. This is can be explained by the fact that muscular activities do not give the desired animations for some meshes.

On the other hand, in order to study the sensitivity of muscle performance on high-poly and low-poly meshes, we have used three sets of face models that have the same appearance, the same polygonal resolution, but with different number of vertices: less than 1000 vertices, between 1000 and 4000 vertices and over than 4000 vertices. The obtained result is drawn in Fig.11. It is clear that most muscles are insensitive to the changes in the number of vertices, with the exception of Orbicularis Oculi. This is can be explained by the fact that in low-poly meshes, the eyelids and eyebrows have common polygons.

It should be noted that the proposed method depicts one module of the WebSign project (ElGhoul & Jemni, 2008) (Othman, ElGhoul & Jemni, 2011) which renders sign language animations in real time using a virtual avatar, from a writing text or a SignWriting notation (Bouzid & Jemni, 2013). The control of the muscles and jaw articulator is done via a scripting language called SML (Sign Modeling Language). Examples of face movements rendered by our virtual avatar are illustrated in Fig. 12.

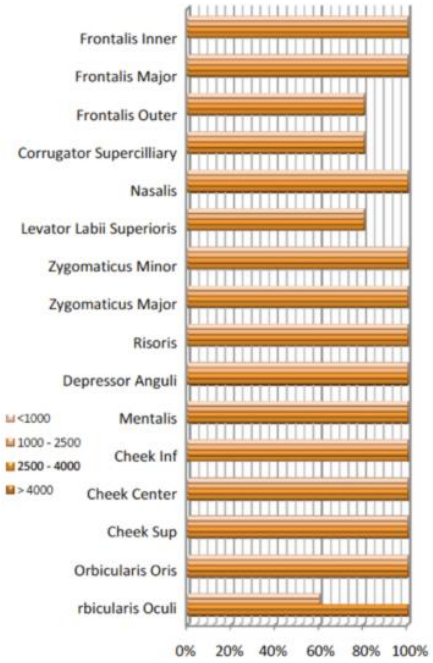


Fig. 11 The recognition rates of muscle motion on low and high poly meshes

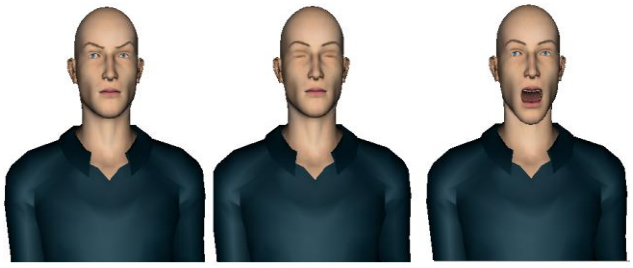


Fig. 12 Examples of face movements

6. Conclusion

We have presented an automatic facial animation method based on Waters vector model and some MPEG-4 feature points. The experimentation shows that more than 90% of tested facial actions can be animated without any human intervention. However, for some low-poly meshes we need to adjust the influenced area of sphincter muscles. To this end, we aim in our future work to modify the sphincter muscle model by ameliorating the algorithm used to detect its influence area.

7. References

Bouzid, Y., ElGhoul, O., Jemni, M. (2013). Synthesizing Facial Expressions for Signing Avatars using MPEG-4 Feature Points. In Proceedings of the 4th International Conference on Information and Communication Technology & Accessibility, Tunisia

Bouzid, Y., Jemni, M. (2013). An avatar based approach for automatically interpreting a sign language notation. In Proceeding of the 13th IEEE International Conference on Advanced Learning Technologies, China

Deng, Z., Noh, J. (2008). Computer facial animation: a survey. In Data-Driven 3D Facial Animation, London: Springer-Verlag, ch. 1, pp. 1-28.

Ekman, P., Friesen, W. (1977). Manual for the Facial Action Coding System. Consulting Psychologist, Palo Alto, California

Elghoul, O., Jemni, M. (2009). Automatic generation of sign language's facial expression. In Proceeding of ICTA 2009, pp143-146, Tunisia

Fratarcangeli, M. (2005). Physically based synthesis of animatable face models. Workshop on Virtual Reality Interaction and Physical Simulation

Ilie, M.D., Negrescu, C., Stanomir, D. (2012). Energy minimization tool for generating composite facial expressions in 3D facial animations, International Journal of Computer Theory and Engineering

Kähler, K. (2007). 3D Facial animation-Recreating human heads with virtual skin, bones, and muscles. VDM Verlag

Lee, Y., Terzopoulos, D., Waters, K. (1993). Constructing physics-based facial models of individuals. In Proceedings of Graphic Interface 1993, pp. 1-8

Liu, C. An analysis of the current and future state of 3d facial animation techniques and systems. (2009). M. S. thesis, School of Interactive Arts and Tech., Simon Fraser Univ., Burnaby, BC, Canada

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee, R. (2000). The syntax of American Sign Language: functional categories & hierarchical structure. Cambridge: MIT Press.

Orvalho, V., Bastos, P., Parke, F., Oliveira, B. (2012). A Facial Rigging Survey. In EUROGRAPHICS 2012

Othmen, A., Elghoul, O., Jemni, M. (2012). An automatic approach for facial feature points extraction from 3D head. In GRAPP/IVAPP 2012, pages 369-372, Italy

Parke, F. I. (1972). Computer generated animation of faces. In Proceedings of the ACM Annual Conference, vol. 1, pp. 451-457

Parke, F. I., Waters, K. (2008).Computer Facial Animation, Wellesley, MA: AK Peters Ltd, pp. 85-362.

Pandzic, I. S., Forchheimer, R. (2002). MPEG-4 facial animation: the standard, implementation and applications, New York

Ping, H. Y., Abdullah, L. N., Sulaiman, P. S., Abdul Halin, A. (2013). Computer Facial Animation: A Review. International Journal of Computer Theory and Engineering.

Platt, S. M., Badler, N. I. (1981). Animating facial expressions. In Proceedings of the 8th Annual Conference on Computer Graphics and Interactive Techniques, pp 245-252. ACM New York, USA

Terzopoulos, D., Waters, K. (1990). Physically-based facial modeling, analysis, and animation. Journal of Visualization and Computer Animation, vol. 1, no. 4, pp. 73-80

Wang, C. L. (1993). A hierarchical spline based facial animation system with simulated muscles. Master's thesis, University of Calgary

Waters, K. (1987). A muscle model for animation three-dimensional facial expression. In Proceedings SIGGRAPH Computer Graphics.

Eye Gaze Annotation Practices: Description *vs.* Interpretation

Annelies Braffort
LIMSI-CNRS
Campus d'Orsay, bat 508, BP 133, F-91403 Orsay cx, France
E-mail: annelies.braffort@limsi.fr

Abstract

If sharing best practices and conventions for annotation of Sign Language corpora is a growing activity, less attention has been given to the annotation of non-manual activity. This paper focuses on annotation of eye gaze. The aim is to report some of the practices, and begin a discussion on this topic, to be continued during the workshop. After having presented and discussed the nature of the annotation values in several projects, and explain our own practices, we examine the level of interpretation in the annotation process, and how the design of annotation conventions can be motivated by limitations in the available annotation tools.

Keywords: eye gaze annotation, sign language corpora

1. Introduction

For Sign Languages (SL), reception of linguistic information is primarily conducted through the eyes. The addressee usually fixates his gaze at the signer's face, particularly the area around the signer's eyes. Eyes are also one of the body components that convey linguistic information, together with other non-manual and manual ones. Several eye aspects can be considered: blinking, eye aperture, and eye gaze. In this paper, we focus on eye gaze.

Eye gaze has a number of different linguistic functions. Some of these functions have been pointed out in the literature (Engberg-Pedersen, 1999): Certain lexical signs may require a specific gaze direction, some iconic constructions require a gaze directed at the hands or at signing space, and gaze is also involved in role shifts. In some theoretical models (Cuxac, 2000), gaze have a semiotic function, allowing distinguishing between two modes of expression: without or with an illustrative process (gaze toward addressee or not).

Analysing SL corpora, by looking at the eye gaze values, their durations, and the co-occurring or surrounding events conveyed by other manual and/or non-manual components, can provide evidences for the definition of formal descriptions linked to functional categories.

SL corpus linguistics is a recent field, and if some practices begins to be promoted and shared, e.g. the use of a database for the lexical signs for annotation consistence and reliability (Hanke, 2008; Johnston, 2008), or even some kind of standardisation (Shembri, 2010; Crasborn, 2012), less attention has been given to the annotation of non-manual activity.

The aim of this paper is to report some of the practices related to eye gaze annotation (section 2), including ours (section 3), and begin a discussion on this topic, to be continued during the workshop (section 4).

2. Eye gaze annotation practices

This section reports the practices used to annotate eye gaze in five projects, for Auslan, ASL and various European SLs. They have been selected to illustrate the various practices.

2.1. Annotation conventions for the Auslan corpus

The Auslan corpus annotation guidelines (Johnston, 2013), designed using the Elan annotation software, is regularly updated as the annotations progress.

The current version of the annotation scheme includes a tier to code eye gaze with four possible values: **a** for "addressee", **t** for "target", **o** for "other", and **z** for "cannot be coded".

These four values code the target of eye gaze.

2.2. Annotation conventions for the ASL Linguistic Research Project in Boston

The ASLLRP project includes the development of annotation software (SignStream) and documentations on the conventions used for the annotation.

The annotation scheme includes an eye gaze tier, with the following values:

- Direction of eye gaze: **left**, **right**, **up**, **down**.
- These values can be combined: **up/lf**, **up/rt**, **dn/lf**, **dn/rt**.
- **addressee** is used to code when eye gaze is directed toward the addressee.
- **track-hand** is used to code when eye gaze follows the hand.
- It is also possible to code eye gaze directed at a specific location, such as **i** (**i** is an index for a given location), or **"under table"**, or even **indef** in the case of an indefinite reference.

In this scheme, the eight first values give the direction of eye gaze from the perspective of the signer. The other values give the target of eye gaze.

2.3. Annotation conventions for the ECHO project

The ECHO European project included a case study devoted to SL. A comparable corpus of three European SL was constituted (the SL of Sweden, United Kingdom, and the Netherlands), together with a common annotation scheme. This annotation scheme includes a gaze tier, with the following values:

- **l-90**: left close to 90 d° (of midsaggital plane)
- **l**: left, close to 45 d° (of midsaggital plane)
- **r-90**: right, close to 90 d° (of midsaggital plane)
- **r**: right, close to 45 d° (of midsaggital plane)
- **u**: upward
- **d**: downward
- Combinations are possible, e.g. **ru** (right and upward)
- **lh**: to the left hand
- **rh**: to the right hand
- **bh**: to both hands
- **p**: toward a person present
- **c**: toward the camera

In this system, the six first values and the associated combinations code the eye gaze direction, while the others code the target of eye gaze.
For the direction values, a different granularity is used depending on the plane: The horizontal plane is segmented into four values, the vertical one into two values.

2.4. ViSiCAST European project

The ViSiCAST European project didn't include a task on corpus annotation and the design of an associated annotation scheme, but some work has been done on a computing representation of signed utterances. An XML system called SiGML, based on HamNoSys, has been designed. This is a timed multi-tier representation where each tier encodes one of the parallel information channels. One of the tiers is used to represent eye gaze, with the following values:

- **AD**: toward addressee
- **FR**: far
- **HD**: towards the signer's own hands
- **HI**: Towards the signer's own dominant hand
- **HC**: Towards the signer's own non-dominant hand
- **UP, DN, LE or RI**: up, down, left or right
- **NO**: no target, unfocussed
- **RO**: rolling eyes

Here also, there is a mix between directional type and target type values. Moreover, a new kind of value is used, which is dynamic: "rolling eyes".
Another particularity of this system is that it is considered that head movement and eye gaze can be linked. This is represented in the head tier, not in the eye gaze one, and here also, this is a dynamic value.

2.5. Intersign network

The Intersign European network[1] aimed at developing standards and guidelines for the study of European SL. Six SLs were represented.
One of the contributions was related to eye gaze in Danish SL with considerations about notations issues for forms and functions (Engberg-Pedersen, 1999). In this contribution, three levels of interpretation in the notations are proposed, from pure formal to pure functional:

1. At the formal level:
   - eye contact with the receiver;
   - some other direction than the receiver;
   - eye blink.
2. At an intermediate level, when eye gaze is directed at the signing space:
   - Are instances of eye gaze in some other direction than the receiver in a meaningful direction or not?
   - If the direction is meaningful, is it in the direction of a locus or in the direction of a configuration?
3. At the functional level, where there are five categories, based on a distinction between two types of signing depending on who the signer's locus represents: the signer as sender of the current utterance (i.e. the sender level) or one of the individuals talked about (i.e. the character level). The following category definitions are extracted from (Engberg-Pedersen, 1999):
   - the narrator's eye contact with the receiver (sender level),
   - avoidance of eye contact at major boundaries by blinking or by looking away in no particular direction (sender level),
   - reference-tracking eye gaze in the direction of a locus just before a predicate or with a topical nominal or a resumptive pronoun (sender level),
   - imitative eye gaze with constructed action, thoughts or dialogue, imitates the holder of the point of view or the quoted person (character level),
   - configurational eye gaze with polymorphemic predicates (it can be the sender or the character level).

In all of these levels, the values code the eye gaze target.
Something particular in this system is the presence of eye blink, which is not a target value. In other annotation schemes, eye blink is considered as one of the possible values of eye lid or eye aperture tiers, or even as a specific tier (Braffort & Chételat, 2011).

2.6. Main trends and particularities

From this report, we can notice that:

- three of these five projects propose as annotation values a combination of directional and target values, and two of them only target values;
- the directional values, based on a segmentation of the signing space from the perspective of the signer, are more or less the same, with in one case a different segmentation of the signing space (more than two values in one plane);
- the target values are quite different; only the "addressee" value is common to all the annotation scheme; some values are more or less detailed, some are present only in one scheme;
- four schemes includes additional values with no equivalent in the other studies: a parameterised value when the gaze is directed toward the signing space, and two dynamic values that doesn't code a direction (blink and rolling eyes).

[1] http://www.sign-lang.uni-hamburg.de/intersign/intersign.html

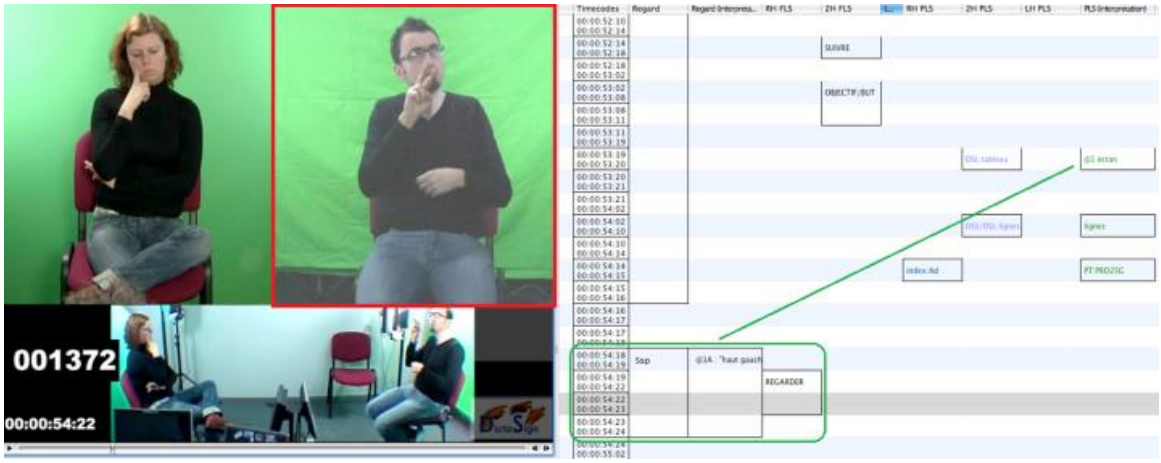


Figure 1: Extract from the annotation of the LSF part of the Dicta-Sign corpus

3. Gaze annotation in the French Sign Language part of the Dicta-Sign corpus

This section reports the practices used to annotate eye gaze in the French Sign Language (LSF) part of the Dicta-Sign corpus (Matthes et al, 2010), which was a comparable corpus created during a European project including studies on four SLs (German, Greek, English, and French).

3.1. Annotation scheme

First, we used only one type of values, in order to facilitate the design of analysis requests that could be more complex in case of mixed values.

Then we used target values, because this allows saving time for analysis. Moreover, this avoids using an arbitrary segmentation of the signing space. We based our controlled vocabulary on the Auslan annotation guideline, with additional details for the cases where eye gaze is directed toward a target, being virtual or real.

Finally, we distinguished two levels of annotation, a more formal one, to code the target kind, and a more interpretative one, to code the supposed target itself in case of target in the signing space. For that, we used two tiers, called Gaze and Gaze interpretation.

The tier "Gaze" allows identifying the target, with the following values:

- **ad**: addressee
- **ssp**: signing space
- **hd**: hand or part of hand
- **real**: real object (e.g. elicitation material, such as paperboard and computer screen) or other person than the addressee
- **x**: far (e.g. the signer is thinking or is looking away without a given target)
- **?**: cannot be coded

The tier "Gaze interpretation" is used to code more information in the case of a *ssp*, *hd* or *real* value in the Gaze tier, with the following values:

- ***@code***: associated with a *hd* value; *code* refers to hands or fingers, identified more or less precisely using a code (e.g. @I_PAD(r) means the index pad of the right hand)
- ***@id:txt***: associated with a *ssp* value; *id* refers to a previously annotated entity located in the signing space, and *txt* id a textual description of the referred entity (e.g. "@2" refers to the localised entity number 2)
- ***code:txt***: associated with a *hd* value; *code* can take one of the three values hands, hand(r) or hand(l), and *txt* described the referred entity hold by the hand(s) (e.g. "hand(r):billet" refers to the right hand holding a ticket (*billet*))
- ***txt***: associated with a *hd* or *real* value; *txt* is a textual description of the referred entity (e.g. "Bottle", or "top right corner of the screen"), or the real object (e.g. "screen") or person (e.g. "moderator").

With this organisation, we can have detailed information on the way eye gaze is used in constructions requiring a gaze directed at the hands or at the signing space.

3.2. Detailed example

Figure 1 illustrates an example of eye gaze annotation with our annotation scheme. The annotation software used is iLex (Hanke, 2008). In this view, time flows from top to bottom, and tiers are vertical. In this example, we have annotated the eye gaze that is associated with the lexical sign REGARDER that means "to look at":

- The first two tiers in the figure are used for eye gaze.
- The third tier is used for the lexical signs performed by the right hand, here **REGARDER**.
- Notice also the eighth tier, which is used to add interpreted information in case of depicting signs. In our example, it has been used to attribute an index to an entity that has been located in the signing space: **@1 écran** means entity number 1, which is a screen (*écran*).
- The value for the Gaze tier (Regard) is **ssp** for signing space.
- The value for the gaze interpretation tier is **@1A: "haut gauche"**. This means that the target is a sub-part of the entity number 1, this sub-part being interpreted as the top left corner of the screen, from the perspective of the signer.

By using this method, we can design requests that allow us to automatically link values related to spatial annotation in different tiers.

4. Discussion

This section proposes thinking about the various practices, their pros and cons, as a start for more interactive discussion during the workshop.

4.1. Description vs. interpretation

A first point is the identification of the level of interpretation in the annotation process, and all the possible biases that annotators do not realize that they have, as they will have common knowledge on the grammar of written language.

As much as possible, the annotations should intend to be descriptive, rather than to express particular theoretical beliefs. But coding of pure descriptive information is sometimes impossible, or even useless.

This is the case for eye gaze, where a "pure description" would be anatomical (e.g. the relative position of the iris regarding a given landmark), or, less directly, mathematical (e.g. a 3d vector). We can imagine that these data could be computed automatically, using image processing tools, providing by this way purely objective annotations. But anyway, segmentation would remain to be done, and more computation would be needed to help interpretation and analysis of the data.

Of course, we can attribute a direction value to eye gaze directly, as this has been done in some of the reported studies here. But this necessitates segmenting the signing space into arbitrary zones, because direct 3d annotation is not possible in the current annotation tools. And also here, interpretation of the target remains to be done.

Then, a more "interpreted description" for eye gaze is to code the target kinds, as we have done in our project. In this case, it is not easy to define objective criteria, and the choice relies on the subjectivity of the annotator. This saves time for the next step of annotation and analysis, at the price of the risk of more errors and less annotator agreement.

4.2. Dependence on the available tools

Another point to consider is that it is very difficult to design conventions that are completely independent of the limitations in the annotation tools. For example, the use of index to allow links to be established automatically during analysis between eye gaze and discourse entity that are located in the signing space is due to the fact that the used tool doesn't allow to create a list of no temporal entities with associated identifiers. This kind of process is possible with the Anvil tool, but on the other hand, Anvil doesn't allow using a lexical database such as in iLex, which is an essential part of annotation tools for SL.

It is likely that our conventions, guidelines and methods will continue to evolve in the following years, as the tools available for annotation become more sophisticated.

Ideally, and this is a call toward the image processing community, the field would greatly benefit of computed descriptions and representations associated with segmentation capabilities. Conversely, progress in the linguistic field would help automatic processing by providing more knowledge on the phenomena to be processed (Gonzalez et al, 2012).

5. References

Braffort, A. & E. Chételat-Pelé (2011). Analysis and description of blinking in French Sign Language for automatic generation, in Gesture and Sign Language in human-computer interaction and embodied communication, LNCS/LNAI 7206, Springer, p.173-182.

Crasborn, O. & Windhouwer, M. (2012) ISOcat data categories for signed language resources. In Gesture and Sign Language in Embodied Communication and Human-Computer Interaction, LNCS/LNAI, Springer, p.118-128.

Cuxac, C. (2000). La langue des signes française (LSF) : les voies de l'iconicité, Ophrys, Paris.

Engberg-Pedersen, E. (1999). Eye gaze in Danish Sign Language monologues: Forms, functions, notation issues. Intersign project papers, v. 3, p.33 -37. URL: http://.../engbergpedersen.html

Gonzalez, M., Filhol, M. & Collet, C. (2012). Semi-automatic Sign Language corpora annotation using lexical representations of signs, in proc. of the 8th int. conf. on Language Resources and Evaluation LREC, ELDA, p.741-745.

Hanke, T. (2002). ViSiCAST Deliverable D-: Interface Definitions. URL: http://.../ViSiCAST_D5-1v017rev2.pdf

Hanke, T. & Storz, J. (2008). iLex - A database tool integrating sign language, in proc. of the 3rd workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, LREC conference, ELDA, p.64-67.

Johnston, T. (2008). Corpus linguistics and signed languages: no lemmata, no corpus, in proc. of the 3rd workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, LREC conference, ELDA, p.82-87.

Johnston, T. (2013). Guidelines for the annotation of the video data in the Auslan corpus. URL: http://.../AuslanCorpusAnnotationGuidelines_Johnston.pdf

Matthes, S., T. Hanke, A. Regen, J. Storz, S. Worseck, E. Efthimiou, A.-L. Dimou, A. Braffort, J. Glauert and E. Safar (2012). Dicta-Sign - Building a Multilingual Sign Language Corpus, in proc. of the 5th workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Proceedings LREC-2012, ELDA p.23-27.

Neidle, C. (2002). SignStream Annotation: Conventions used for the American Sign Language Linguistic Research Project. URL: http://.../asllrpr11.pdf

Nonhebel, A., Crasborn, O. & van der Kooij, E. (2004). Sign language transcription conventions for the ECHO Project, Version 9. URL: http://.../ECHO_transcr_conv.pdf

Schembri, A. & Crasborn, O. (2010) Issues in creating annotation standards for sign language description, in proc. of the 4th workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, ELDA., p.212-216.

An annotation scheme for the linguistic study of mouth actions in sign languages

Onno Crasborn & Richard Bank
Radboud University Nijmegen, Centre for Language Studies
P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl, r.bank@let.ru.nl

Abstract

This paper describes the annotation scheme that has been used for research on mouth actions in the Corpus NGT. An orthographic representation of the visible part of the mouthing is supplemented by the citation form of the word, a categorisation of the type of the mouth action, the number of syllables in the mouth action, (non)alignment of a corresponding sign, and a layer representing some special functions. The scheme has been used for a series of studies on Sign Language of the Netherlands. The structure and vocabularies for the annotation scheme are described, as well as the experiences in its use so far. Annotations will be published in the second release of the Corpus NGT annotations in late 2014.

Keywords: sign language, annotation scheme, transcription, non-manual features, mouth actions, mouthings, mouth gestures

1. Goal

This paper aims to describe the annotation scheme that has been developed for a series of studies of mouth actions in Sign Language of the Netherlands (NGT), based on the Corpus NGT (Crasborn, Zwitserlood & Ros, 2008; Crasborn & Zwitserlood, 2008). These studies are targeted at achieving a better understanding of the role of the mouth as an articulator in NGT, with a focus on mouth actions that consist of or are derived from spoken language words ('mouthings'). While it is clear that such mouthings form a case of simultaneous code mixing, dubbed 'code blending' by Emmorey et al. (2005), it has only recently been argued that mouthings form an integral part of deaf communication in the Netherlands (Bank et al., 2013). They are used in virtually every utterance by every user of the language (Bank et al., submitted).

Psycholinguistic studies have demonstrated that deaf people are proficient lip-readers (e.g., Auer & Bernstein, 2007), and it is likely that this information contributes to successful interaction between deaf people also when they use sign language as their primary and preferred mode of communication. While the nature and function of mouth actions have received considerable attention in the sign language literature for a variety of (primarily European) languages (cf. the contributions to Boyes Braem & Sutton-Spence, 2001), no large-scale corpus studies had been performed until recently.

To be able to study the various properties of mouth actions in a corpus, we devised an annotation scheme that systematically separates form from meaning, and that aims to increase efficiency by using Dutch orthographic representations rather than a visual phonetic representation in terms of 'visemes' for the basic transcription layer.

2. The annotation scheme

In this paragraph, we describe the six tiers that we use for every signer in an ELAN annotation file. The transcription (par. 2.1) is independently aligned, while the other tiers containing annotations to the transcription are dependent on this parent tier. This leads to the tier structure displayed in Figure 1.

Mouth	par. 2.1
MouthLemma	par. 2.2
MouthType	par. 2.3
MouthSpr	par. 2.4
MouthSyll	par. 2.4
MouthAdd	par. 2.5

Figure 1: Tier structure for mouth actions

In section 3, we will further discuss how this structure is further implemented in the Corpus NGT.

2.1 Transcription

2.1.1. Preliminary considerations

The start of any investigation into mouth actions will be based on a description of their forms. This immediately leads to problems, as there is no standard transcription system that can be used. One option is to focus purely on the visible properties of articulations, using a classification of the amount of lip rounding, lip opening, and visibility of the tongue, for instance. This appears attractive as it is these properties that are accessible in deaf communication, any possible acoustic accompaniments not being perceivable to deaf people. Although proposals for such 'viseme' categories have been proposed in the literature (see Massaro, 1998; Cappalletta & Harte, 2012; Nonhebel et al., 2004), they lead to a description that in a sense is true to the function of the forms, but that is hard to read. The same holds for a detailed articulatory transcription of mouth actions by use of the action units available in the Facial Action Coding System (FACS; Ekman & Friesen, 1978).

As has become clear from earlier research, the majority of mouth action tokens are mouthings, articulations that consist of (parts of) spoken words. It is thus attractive to somehow use knowledge of speech in the transcription of mouth actions, if only for mouthings.

We know however that any attempt at speech reading involves a lot of interpretation, all aimed at reconstructing words from a spoken language from a small number of visible contrasts. Only a subset of the phonological distinctions in a spoken language has a visible correlate. For vowels, for instance, lip rounding and to a limited extent also tongue/jaw height can be visually perceived, but front-back distinctions in vowels are almost impossible to perceive visually. Thus, if we would use a phonetic or orthographic transcription of a spoken language, we need to make a lot of inferences about what the signer might be saying, on the basis of relatively little phonetic evidence. Comparing the meaning of the perceived mouthings with the co-occurring sign may help in deciding on the transcription, but it may also be misleading.

A different problem with using a transcription system that is based on a representation of the spoken language is that not all mouth actions can be related to spoken language words. In most, if not all sign languages, not only mouthings but also mouth gestures are used (papers in Boyes Braem & Sutton-Spence, 2001; Crasborn et al., 2008). These mouth gestures are by definition not composed of (parts of) spoken words, and may include a variety of articulations (see Crasborn et al., 2008, and Woll, 2001 for discussion). Transcribing them by using a system that is made for speech creates the false suggestion that mouth gestures have always somehow evolved from spoken language words.

Despite these drawbacks, we decided to use an orthographic representation of the spoken language (primarily Dutch, in our case) to transcribe mouth actions. The most powerful argument in favour of this choice is efficiency: not using (a visual version of) a phonetic notation like IPA but using spoken language orthography saves enormous amounts of time during the annotation phase, and the same holds for the exploitation phase. Because of the good readability of orthographic transcriptions as compared to regular phonetic (let alone visual phonetic) transcriptions, the chances that the information about mouth actions will be taken into account in a variety of future studies based on our corpora, orthographic transcriptions are also to be preferred from the point of view of the general user of corpus data. Based on our research findings for NGT that will be briefly discussed in section 4 below, we argue that in addition to glosses and a sentence-level translation, a transcription of mouth actions should be a basic layer of annotation that is needed for any sign language corpus.

The arguments relating to efficient annotation and efficient exploitation are rather similar in nature to the arguments for using a gloss representation for manual signs. Although spoken language glosses have all kinds of disadvantages (including the representation in another language), they are unrivalled in their usability (Johnston, 2010).

Aside from these practical considerations for the corpus annotator and corpus user, filling in details of spoken language articulations that cannot be perceived visually is not all that unnatural: it is what deaf speechreaders do all the time, and are highly proficient at (Woll, 2012). Where (deaf and hearing) communicators are constantly using limited visual information to arrive at an interpretation of what is being said (a process not unlike auditory speech perception in noisy circumstances or in the case of fast speech, for instance), it is important to keep the task of transcription in mind when we annotate mouth actions for corpus annotation. The goal here is not to correctly lemmatise the spoken word, but merely to transcribe the parts of spoken language words that the annotator observes, or in the case of mouth gestures, to arrive at a consistent written representation of the visible form irrespective of any possible spoken language origin. More concretely, what we propose to use for the transcription of mouthings is to only include the segments or syllables that are actually produced, and not any deleted segments or syllables. Reference to the spoken language lemma that the articulation is hypothetically an instance of can be made on the Lemma tier (see section 2.2 below).

2.1.2. Conventions

Mouth action transcriptions are made on a tier called 'Mouth'. Articulations that are perceived as being (fragments of) spoken language words (mouthings) are written in lowercase without any special markers. All other mouth actions (any type of mouth gesture) are put between single quotation marks ('…'). If a mouth gesture cannot be easily described in terms one or more spoken language segments, we use a phonetic description of the mouth articulation between pipes (|…|). This set of descriptors was based on what was developed for the ECHO project (Nonhebel et al., 2004), and adapted on an ad hoc basis.

Acoustic correlates of the mouth action such as phonation are not annotated. We acknowledge that for studies on code mixing, for instance, this could be important information. We suggest that this type of information could best be annotated on a separate tier, with conventions to be established in accordance with the purpose of a specific research goal.

As on other tiers used in the Corpus NGT, uncertainty about the correct representation can be labelled with a single question mark following the transcription. As with manual signs, false starts are prefixed with a tilde symbol (~).

Especially in the case of mouth gestures, the nature of the transcriptions will be influenced by the research findings on this topic for the language at hand (whether in linguistic publications or implicit in dictionary representations or teaching materials). While consistency will be difficult to achieve in the absence of a vocabulary of mouth gestures, the creation of such a vocabulary can be the result of multiple revisions of the set of transcriptions created by a variety of annotators in a first annotation pass. The ECHO conventions for mouth gestures referred to above may serve as a basis for this, but are in need of an evaluation and possibly adaptation,

as they have never been used for a large-scale corpus, as far as we know.

2.2 Lemma

As was already referred to above, the MouthLemma tier is a child tier of the transcription of the Mouth tier, and is the place where the presumed uninflected lemma can be notated of which the observed mouthing is an instance. By using a lemma rather than a full (inflected) form of the spoken word, we stay clear from any overinterpretation of (the morphological specificity of) the mouthing.

The lemma information allows for the searching for mouth actions based on a spoken word type, and will thus facilitate the extraction of various instantiations of the word, whether inflected or not inflected and no matter how reduced or repeated (see section 2.4 below) a Mouth token may be. For this reason, it would be advisable to include a lemma annotation for all mouth annotations, also when they do not differ.

2.3 Classification

On the tier MouthType, we classify the mouth action transcribed on the Mouth tier. We adopt the five-part classification proposed in Crasborn et al. (2008), distinguishing the following categories:

M	Mouthing
E	'Empty' mouth gesture: a lexicalised phonological component of a sign that is not derived from a spoken word
A	Adverbial mouth actions, lexicalised independently of a manual sign
4	'Mouth for mouth' actions: instances where the mouth represents the mouth (as in pantomiming drinking or chewing)
W	Whole-face actions that include a specific mouth articulation, as in affective facial expressions

Figure 2: Types of mouth actions

In addition to these five main types, the Mouthing category is further specified into five subtypes, presented in Figure 3.

M	Regular mouthing
M-back	Mouthing used as backchannel signal
M-add	Mouthing that is not related to a manual sign but temporally overlaps with manual signs.
M-solo	Mouthing that does not overlap with manual signs
M-spec	Mouthing that is co-articulated with a manual sign that serves to specify the semantics of the manual sign

Figure 3: Types of mouth actions for different uses of mouthings

This latter subdivision has arisen in the context of our investigations into NGT mouthings, briefly discussed in section 4. A similar investigation into mouth gestures is likely to lead to a further specification of the four types of mouth gestures listed in Figure 2 (see e.g. Sandler's (2009) category of 'iconic mouth gestures').

2.4 Phonetic properties

Two types of phonetic properties are encoded each on their own tier. First of all, the alignment of the mouthing with the manual glosses is characterised on the MouthSpr tier ('Mouth spreading', following the description of spreading as a prosodic process in Sandler, 2006). As in feature spreading in spoken language segmental phonology, spreading refers to the phenomenon that certain articulatory features may be lengthened to co-occur not only with their source, but also with neighbouring elements. In the case of spreading mouthings, mouthings that have a clear 'source' sign with which the mouthing semantically overlaps are articulated in such a way that they also overlap with the preceding or following sign(s).

The annotation on the MouthSpr tier contains information on the glosses that overlap with the mouth annotation. Angled brackets are used to encode the direction of spreading (< for regressive, > for progressive). For example, the MouthSpr annotation 'BIER > DRINKEN', together with the Mouth annotation *bier* 'beer', means that the mouthing that accompanies the manual sign BEER is either lengthened or maintains it final state so long as to also cover the manual sign DRINKEN 'to drink'. Signers are usually not maximally synchronised in their articulation of sign/mouth pairs, so MouthSpr annotations should not be applied every time that there is a single-frame difference in start or end, irrespective of the duration of the actions and/or the signing speed, for instance. In our own investigations, a mouthing is categorised as spreading over an adjacent sign when it overlaps that sign with at least 50% or 10 or more video frames, whichever applies first.

A second type of phonetic information can be encoded on the MouthSyll tier. It is used to specify the number of syllables of the observed mouth articulation. For mouthings, the number of syllables of the visible word would be transcribed, while for mouth gestures, if countable, the number of cycles of the articulation would be encoded. We have not yet used this tier for our ongoing investigations, but it is devised to study the alignment of manual and oral actions. There are cases in our data where the first syllable of mouthings is reduplicated, seemingly to correspond to the number of movement cycles (syllables) in the manual sign. To investigate the hypothesis that 'the hand drives (the prosody of) the mouth', systematic annotation of the MouthSyll together with the number of movements on the 'NOM' tier (a child of the gloss tiers in the Corpus NGT) will be needed.

2.5 Semantic role

While in our data most mouthings appear to be clearly linked to manual signs both in terms of their semantics (typically overlapping with, if not equal to, that of the sign) and in terms of their timing (typically being co-articulated), there are also mouthings that cannot be analysed as linked to a manual sign. We call these 'added mouthings', as they add an element to the semantics of the whole utterance (rather than specifying the semantics of an individual sign). Solo mouthings (specified as such on the MouthType tier, see Figure 3), have the same function as added mouthings but do not overlap with manual signs. They occur often at the start or end of a signed phrase, before the signing starts or after the signing has ended.

In order to efficiently analyse these utterances, the annotations on the MouthAdd tier consist of a string of manual glosses (ignoring differences between one-handed and two-handed signs and various types of two-handed constructions) followed by a string of mouthings.

Although these annotations are made on sentence level or phrase level, they can still be rather short. For example, utterances like *BEGINNEN begin maar* 'START start go-ahead' are not uncommon.

3. Application of the scheme to the Corpus NGT

We are using the tier structure described above for annotating the Corpus NGT with the ELAN annotation tool. In order to systematically separate annotations for the two signers in the dialogues, we create a double set of tiers, one set per participant in the dialogue. The tiers are suffixed by "S1" and "S2" for the two signers, a system that is used throughout the Corpus NGT and that could easily be adapted for multilogues. A participant tag (S001, S002, ..., S092) for each tier makes it possible to uniquely link each annotation to an individual signer.

The two tiers are 'linked' by having the same 'linguistic type' property in the ELAN documents. This linguistic type is an obligatory specification for each tier, and is in turn specified among other things for its independent or child status, and in the latter case, for the name of the parent tier and the nature of the relation of (one or more) annotations on the child tier to an annotation on the parent tier. In the tier hierarchy outlined in Figure 1 above, the Mouth tiers are independent tiers, not having a parent tier to which they are associated, while all other tiers are child tiers of a Mouth tier. The linguistic types of these child tiers are all specified with the restriction 'symbolic association', meaning that there is a one-on-one relation between child annotation and parent annotation, and that the child annotations cannot be independently aligned with the time axis. Figure 4 presents the names of the linguistic types for the six mouth tiers. Following the conventions for the Corpus NGT, tier names have initial capitals for each word, while linguistic types only use lowercase in combination with underscores to separate words. These conventions help to highlight the distinction between tiers and types both in ELAN and when working with the XML code in the ELAN document.

Tier name	**Linguistic Type**
Mouth	mouth
MouthLemma	mouth_lem
MouthType	mouth_type
MouthSpr	mouth_spr
MouthSyll	mouth_syll
MouthAdd	mouth_add

Figure 4: Tiers and their linguistic types in ELAN

4. Use of the annotation scheme in recent and on-going research

The above annotation scheme has been developed for a series of studies on mouth actions in NGT, with a focus on mouthings. A small subset of the Corpus NGT of over 94 minutes (40 sessions containing data from 40 signers) was fully annotated for the Mouth tiers at the time of writing. In the whole corpus, over 250 sessions contained some Mouth annotations, counting almost 12,000 tokens for a total of 70 different participants. These Mouth tier annotations were all classified according to type on the MouthType tier, and formed the basis of all our studies. Depending on the specific research goal, data from the whole corpus were used or from the smaller subset identified above.

In a first study (Bank et al., 2011), we investigated the variation in Dutch lexical items used as mouthings for twenty highly frequent signs. We used the MouthLemma tier to find all tokens of a certain type, and the MouthType classification to make a distinction between mouthings and mouth gestures. The main source of variation turned out to be between using a mouthing versus a mouth gesture, rather than between different spoken words occurring with the same manual sign. This dichotomy between mouthings and mouth gestures was established by using the MouthType tier.

We continued to investigate mouthings by looking at their spreading behaviour, encoding this information on the MouthSpr tiers (Bank et al., 2013). This allowed us to easily classify regressive and progressive spreading, as well as determining the scope of spreading by counting the number of angled brackets in an annotation. The finding here confirmed the findings of Crasborn et al. (2008) for the ECHO fable stories, namely that spreading is a frequent phenomenon: more than one in ten mouthings are spread out over two or more signs. The MouthSyll tiers could be used in future investigations on spreading that aim to analyse the phonological length of words, comparing those with the length of signs. Although we report some findings on this subject, we did not systematically annotate the number of syllables in each mouthing.

While in this study on spreading, no sociolinguistic differences were found based on distinctions in gender,

age, or region, we continued to look at sociolinguistic differences in the use of mouth actions more generally. In Bank et al. (submitted) we report the finding that while no group differences were found based the variables region, gender, or age, what does stand out is the high frequency of mouthings in comparison to the various types of mouth gestures. Depending on the signer, between 65 and 100% of all mouth actions are mouthings. We concluded that spoken language is an important element of deaf interaction in the Netherlands, even for native signers signing to other native signers whom they know well. Although the semi-spontaneous interaction was recorded in a lab setting, the further conclusion appears warranted that there simply is no 'pure' NGT in the sense of not being accompanied by elements of the spoken language, even though we consider NGT to be a language with its own lexicon and its own grammar.

In a final study, we are building on this conclusion by making use of the MouthAdd tiers (Bank et al., forthcoming). The MouthAdd tier is the only place where oral and manual information is combined, information that cannot otherwise be retrieved in an automated search in ELAN. In this study, we will analyse the structure of utterances where mouthings do more than contribute redundant information to manual signs or specify the semantics of manual signs.

The data for all of these studies will be published in the second release of the Corpus NGT annotations foreseen for the autumn of 2014.

5. Conclusion

We hope to have described an annotation scheme for mouth actions that could benefit a large number of sign language corpora. Many of the phenomena at its basis have been observed for many sign languages, albeit often on the basis of rather small data sets. We recommend the transcription of mouth actions on the Mouth tier as a basic element of corpus annotation for all sign languages, especially ones in which mouthings are not uncommon.

Admittedly, the validity of the distinctions that we propose to some extent remains to be confirmed by more research. As with other types of sign language corpus annotation, the annotation and analysis of many elements of signed interaction remains a constant process of improvement and revision based on new research methods and new insights into the functioning of sign languages and deaf interaction more generally. This should not withhold us from striving towards annotation standards (cf. Schembri & Crasborn, 2010).

Unlike the validity, the inter-annotator and intra-annotator reliability of the various elements of the annotation scheme is something that could be established relatively easily by dedicated studies. This is one of the steps we plan to take next.

References

Auer, E. T., Jr., & Bernstein, L. E. (2007). Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *Journal of Speech, Language, and Hearing Research, 50*(5), 1157-1165.

Bank, Richard, Crasborn, Onno, & Hout, Roeland van. (2011). Variation in mouth actions with manual signs in Sign Language of the Netherlands (NGT). *Sign Language and Linguistics*, 14(2), 248–270.

Bank, Richard, Crasborn, Onno, & van Hout, Roeland. (2013). Alignment of two languages: The spreading of mouthings in Sign Language of the Netherlands. *International Journal of Bilingualism*. doi: 10.1177/1367006913484991

Bank, Richard, Crasborn, Onno, & van Hout, Roeland. (Submitted). The prominence of spoken language elements in a sign language.

Bank, Richard, Crasborn, Onno, & van Hout, Roeland. (forthcoming). Bimodal code-mixing: speech supported signing is the norm in NGT signers. Ms., Radboud University Nijmegen.

Boyes Braem, P., & Sutton-Spence, R. (Eds.). (2001). *The hands are the head of the mouth. The mouth as articulator in sign languages*. Hamburg: Signum Verlag.

Cappalletta, Luca, & Harte, Naomi (2012). *Phoneme-to-viseme mapping for visual speech recognition. Proceedings of the International Conference on Patter Recognition, Applications and Methods*, pp. 322-329.

Crasborn, Onno, Kooij, Els van der, Mesch, Johanna, Waters, Dafydd, & Woll , Bencie (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language and Linguistics, 11*(1), 45-67.

Crasborn, Onno, & Zwitserlood, Inge. (2008). The Corpus NGT: an online corpus for professionals and laymen. In Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst Thoutenhoofd & Inge Zwitserlood (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Marrakech, Morocco: ELRA, pp. 44-49.

Crasborn, Onno, Zwitserlood, Inge, & Ros, Johan. (2008). The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands (Video corpus). from Centre for Language Studies, Radboud University Nijmegen http://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6

Ekman, Paul, & Friesen, Wallace V. (1978). *The facial action coding system. Investigator's guide.* Palo Alto, CA: Consulting Psychologists Press.

Emmorey, Karen, Borinstein, H.B., & Thompson, Robin. (2005). Bimodal bilingualism: Code-blending between spoken English and American Sign Language. In J. Cohen, K.T. McAlister, K. Rolstad & J. MacSwan (Eds.), *ISB4: Proceedings of the 4th International Symposium on Bilingualism*. Somerville, MA:

Johnston, Trevor. (2010). From archive to corpus: Transcription and annotation in the creation of signed

language corpora. *International Journal of Corpus Linguistics*, 15(1), 104-129.

Massaro, Dominic W. (1998). *Perceiving talking faces. From speech perception to a behavioral principle*. Cambridge, MA & London: The MIT Press.

Nonhebel, Annika, Crasborn, Onno, & Kooij, Els van der. (2004). Sign language transcription conventions for the ECHO Project: BSL and NGT mouth annotations. Ms, Radboud University Nijmegen.

Sandler, Wendy. (2006). From phonetics to discourse: the nondominant hand and the grammar of sign language. In Louis Goldstein, D.H. Whalen & Catherine Best (Eds.), *Laboratory Phonology 8* (pp. 185-212). Berlin: Mouton de Gruyter.

Sandler, Wendy. (2009). Symbiotic symbolization by hand and mouth in sign language. *Semiotica*, 174(1), 241-275.

Schembri, Adam, & Crasborn, Onno. (2010). *Issues in creating annotation standards for sign language description*. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Valletta, Malta, pp. 212-216.

Woll, Bencie. (2001). The sign that dares to speak its name: echo phonology in British Sign Language (BSL). In Penny Boyes Braem & Rachel Sutton-Spence (Eds.), *The hands are the head of the mouth* (pp. 87-98). Hamburg: Signum Verlag.

Woll, Bencie. (2012). Speechreading revisited. *Deafness & Education International*, 14(1), 16-21.

Implementation of an Automatic Sign Language Lexical Annotation Framework based on Propositional Dynamic Logic

Arturo Curiel[†] **, Christophe Collet**
Université Paul Sabatier - Toulouse III
118 route de Narbonne, IRIT,
31062, Toulouse, France
E-mail: curiel@irit.fr, collet@irit.fr

Abstract

In this paper, we present the implementation of an automatic sign language (SL) sign annotation framework based on a formal logic, the Propositional Dynamic Logic (PDL). Our system relies heavily on the use of a specific variant of PDL, the Propositional Dynamic Logic for Sign Language ($\mathrm{PDL}_{\mathsf{SL}}$), which lets us describe SL signs as formulae and corpora videos as labeled transition systems (LTSs). Here, we intend to show how a generic annotation system can be constructed upon these underlying theoretical principles, regardless of the tracking technologies available or the input format of corpora. With this in mind, we generated a development framework that adapts the system to specific use cases. Furthermore, we present some results obtained by our application when adapted to one distinct case, 2D corpora analysis with pre-processed tracking information. We also present some insights on how such a technology can be used to analyze 3D real-time data, captured with a depth device.

Keywords: sign language framework, automatic annotation, propositional dynamic logic

1. Introduction

Research in sign language (SL), both from the point of view of linguistics and computer science, relies heavily on video-corpora analysis (Dreuw et al., 2008). As such, several methods have been developed over time for the automatic processing of both video or other sensor-based corpora (Ong and Ranganath, 2005). Even though these kind of research efforts are usually geared toward recognition, few work has been done in relation to the unification of raw tracked data with high level descriptions (Cooper et al., 2011; Bossard et al., 2004). This calls to a reflection on how we represent SL *computationally*, from the most basic level.

SL lexical representation research is focused on sign synthesis before than recognition. Works like (Filhol, 2009; Losson and Vannobel, 1998) present the use of geometric lexical descriptions to achieve animation of signing 3D avatars. While their approach is well suited for synthesis, it is not completely adapted for sign identification. Recognition tasks in both natural language processing and computer vision are well known to be error-prone. Also, they are highly susceptible of bumping into incomplete information scenarios which may require some kind of inference, in order to effectively resolve ambiguities. In addition, SL linguistic research has consistently shown the existence of common patterns across different SLs (Aronoff et al., 2005; Meir et al., 2006; Wittmann, 1991) that may be lost with the use of purely geometrical characterizations, as the ones needed in synthesis. This limits the application of these kind of sign representations for automatic recognition, especially since we would want to exploit known linguistic patterns by adding them as properties of our descriptions. Works like (Kervajan et al., 2006; Dalle, 2006) have acknowledged the necessity of introducing linguistic information to enrich interaction, in an effort to help automatic systems bear with ambiguity. Moreover, the use of additional linguistic data could simplify connections between lexical information and higher syntactic-semantic levels, hence pushing us closer to automatic discourse analysis. However, this has long been out of the scope of synthesis-oriented description languages.

On the side, research in SL recognition has to deal with other important drawbacks not present in synthesis, namely the use of very specialized tools or very specific corpora. This alone can severely impact the portability of a formal, computer-ready, representation out of the original research context, as it complicates the use of the same techniques across different information sources and toughens integration with new tools.

The framework described here is based on previous work presented by (Curiel and Collet, 2013) on the Propositional Dynamic Logic for Sign Language ($\mathrm{PDL}_{\mathsf{SL}}$). $\mathrm{PDL}_{\mathsf{SL}}$ is a formal logic created with the main purpose of representing SL signs in a computer-friendly way, regardless of the specific tools or corpora used in research. Such a representation can potentially reduce the overhead of manually describing SL signs to a computer, by establishing well-known sets of rules that can be interpreted by both humans and automatic systems. This could, incidentally, reduce dependency on thoroughly geometrical descriptions. Moreover, the flexibility of $\mathrm{PDL}_{\mathsf{SL}}$ lets us combine any kind of information in our descriptions; for example, we can integrate non-manual markers if we have sight and eyebrow tracking, or we can add 3D movements if we are using a depth camera.

In general, we propose an automatic SL lexical annotation framework based in $\mathrm{PDL}_{\mathsf{SL}}$ descriptions. Ideally, the system will:

- simplify the application of logical inference to recognize $\mathrm{PDL}_{\mathsf{SL}}$-described signs;

[†] Supported by CONACYT (Mexico) scholarship program.

- characterize and analyze corpora in terms of PDL_{SL} models;
- represent SL with different degrees of granularity, so as to adapt the formulae to the specific technical capabilities available in each use case.

Our framework aims to ease the integration of PDL_{SL} with various corpora and tracking technologies, so as to improve communication between different SL research teams. We expect that this will, in turn, enable the construction of both research and user-level applications in later stages.
The rest of the paper is divided as follows. In section 2., we introduce the basic notions of our formal language, applied to 2D SL video-corpora analysis. Section 3. shows how we can describe SL lexical structures as verifiable PDL_{SL} formulae. Section 4. gives a detailed description of the system's architecture. Finally, sections 5. and 6. present some preliminary results and conclusions, respectively.

2. Sign Language Formalization with Logic

The Propositional Dynamic Logic (PDL) is a multi-modal logic first defined by (Fischer and Ladner, 1979) to characterize computer languages. Originally, it provided a formal framework for program descriptions, allowing them to be interpreted as modal operators. PDL_{SL} is an specific instance of PDL, based on the ideas of sign decomposition by (Liddell and Johnson, 1989) and (Filhol, 2008). In general, PDL_{SL}'s modal operators are movements executed by *articulators*, while static postures are interpreted as propositional states reachable by chains of movements.
A propositional state will be none other than a set of distinct atomic propositions. These can be used to represent articulators' positions with respect to one another; specific configurations; or even their spatial placement within a set of *places of articulation*. Table 1 shows a brief summary of the atomic propositions defined to analyze 2D corpus data.

Symbol	Meaning
$\beta_1{}^{\delta}_{\beta_2}$	articulator β_1 is placed in relative direction δ with respect to articulator β_2.
$\mathcal{F}_c^{\beta_1}$	articulator β_1 holds configuration c.
$\Xi_\lambda^{\beta_1}$	articulator β_1 is located in articulation place λ.
$\mathcal{T}_{\beta_2}^{\beta_1}$	articulator β_1 and β_2 touch.

Table 1: Atomic propositions for PDL_{SL}

Basic movements can be described by atomic actions codifying either their direction, speed or even if they follow a particular trajectory. This is exemplified by the definitions on Table 2, which presents some of the operators used to characterize 2D corpus movements.
Both atomic propositions and actions presented in this case were chosen specifically to capture information that we are able to detect with our tracking tools. Different sets of atoms can be defined depending of the technical capabilities available to assert their truth values (*e.g.* sight direction, eyebrow configuration, hand movement, etc).

Symbol	Meaning
δ_{β_1}	articulator β_1 moves in relative direction δ.
$\leftrightsquigarrow_{\beta_1}$	articulator β_1 *trills*, moves rapidly without direction.
skip	denotes the execution of any action

Table 2: Atomic actions for PDL_{SL}

Atoms form the core of the PDL_{SL} language, which is presented below in Backus–Naur Form (BNF) by way of definitions 1 and 2.

Definition 1 (Action Language for SL Body Articulators $\mathcal{A}_{\mathsf{SL}}$)**.**

$$\alpha ::= \pi \mid \alpha \cap \alpha \mid \alpha \cup \alpha \mid \alpha;\alpha \mid \alpha^*$$

where π is an atomic action.

Definition 2 (Language PDL_{SL})**.**

$$\varphi ::= \top \mid p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\alpha]\varphi$$

where p denotes an atomic proposition and $\alpha \in \mathcal{A}_{\mathsf{SL}}$.

A more formal presentation of the model basis can be found in (Curiel and Collet, 2013).

3. Extending PDL_{SL} formulae to Describe Sign Language Lexical Properties

The presented PDL_{SL} language lets us easily codify individual signs by way of our logic formulae. However, during implementation, we noticed the need to extend the original formalism in order to develop a better suited characterization of more general properties. We wanted to represent lexical structures common across multiple signs. With this in mind, we extended PDL_{SL} to include *lambda expressions*, explained in (Barendsen, 1994), for variable binding. The introduced syntax is presented in definition 3.

Definition 3 (Extended PDL_{SL})**.**

$$var ::= \langle \texttt{uniqueID} \rangle \mid var, var$$

$$\varphi_f ::= \varphi \mid var \mid \neg\varphi_f \mid \varphi_f \wedge \varphi_f \mid \lambda\, var.(\varphi_f) \mid var = \varphi_f$$

where $\varphi \in \text{PDL}_{\mathsf{SL}}$.

The rules of quantification and substitution remain the same as in classic lambda calculus.
Lambdas let us describe properties over sets of PDL_{SL} atoms instead of one. For example, Figure 1 shows two french sign language (FSL) signs, $\texttt{SCREEN}_{\mathsf{FSL}}$ and $\texttt{DRIVE}_{\mathsf{FSL}}$. Both can be described as instances of the same underlying common structure, characterized by both hands holding the same morphological configuration while being positioned opposite from one another.
Their common base can be described by way of a lambda expression as shown in example 1.

Example 1 (**opposition** lambda expression)**.**

$$hands\_config = \lambda c.(\mathcal{F}_c^{\texttt{right}} \wedge \mathcal{F}_c^{\texttt{left}})$$
$$\textbf{opposition} = \lambda c.(\texttt{right}^{\leftarrow}_{\texttt{left}} \wedge hands\_config(c))$$

SCREEN$_{\text{FSL}}$ DRIVE$_{\text{FSL}}$

COMMON STRUCTURE

Figure 1: Comparison of signs SCREEN$_{\text{FSL}}$ and DRIVE$_{\text{FSL}}$ sharing the same underlying structure

In example 1, $\mathcal{F}_c^{\texttt{right}}$ means that `right` holds configuration c. Atom $\mathcal{F}_c^{\texttt{left}}$ has the same meaning, but for the `left` hand. Atom $\texttt{right}^{\leftarrow}_{\texttt{left}}$ means that `right` hand lies in direction $\leftarrow$ with respect to `left`, from the annotator's point of view. In this case we called our expression **opposition**, because both hands are in opposite horizontal positions from one another.

Once we've defined the base structure, the SCREEN$_{\text{FSL}}$ and DRIVE$_{\text{FSL}}$ signs can be easily described in the database by passing the missing arguments to our lambda expression (as shown by example 2).

Example 2 (**opposition**-derived signs)**.**

$$\text{SCREEN}_{\text{FSL}} = \textbf{opposition}(\texttt{L\_FORM})$$
$$\text{DRIVE}_{\text{FSL}} = \textbf{opposition}(\texttt{FIST\_FORM})$$

In example 2, `L_FORM` is a morphological configuration of the hand where the thumb and the index fingers are held orthogonally. Similarly, `FIST_Form` is a configuration where hand is held as a closed fist. Here we just expressed that **opposition** will substitute each apparition of its first argument with either form, so as to define two distinct signs.

We could also have described both signs as standalone, independent formulae. However, by describing the common structures across different signs, we are able to cope better with incomplete information in recognition. For example, a generic **opposition** structure with free variables will correctly hit in states where we can recognize hand positions but no hand configurations (as it's often the case). This immediately derives into a list of possible signs that could be later reduced with either further processing or with user interaction. In this scenario, standalone formulae for SCREEN$_{\text{FSL}}$ and DRIVE$_{\text{FSL}}$ wouldn't be found, since only using position information isn't enough to tell them apart.

4. Detailed Framework Architecture

The objective of the system is to take an untreated SL video input, either in real time or not, and return a set of satisfied PDL$_{\text{SL}}$ formulae. Moreover, the system has to return a PDL$_{\text{SL}}$ model representing any relevant information contained in the video as a labeled transition system (LTS). This can only be fulfilled by adapting the modeling process on-the-fly to the specific characteristics of our data. To achieve this end, our framework generalizes the original architecture proposed by (Curiel and Collet, 2013), shown in Figure 2, so as to enable module swapping depending on the technical needs presented by the inputs.

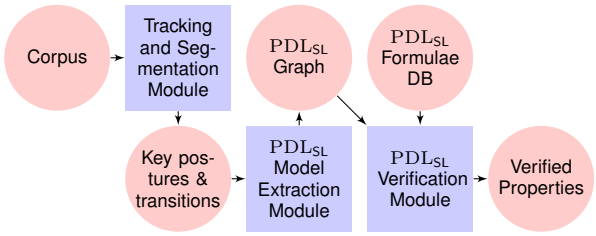


Figure 2: Block diagram of a generic PDL$_{\text{SL}}$-based SL lexical structure recognition system

In the original version, a *Tracking and Segmentation* module uses the raw data of an automatic hand-tracker on 2D corpora, like the one presented by (Gonzalez and Collet, 2011), and returns a list of time-intervals classified either as *holds* or *movements*. The aforementioned interval list is passed to the *Model Extraction Module*, which translates each *hold* and *movement* into a time-ordered LTS. In the LTS, *holds* correspond to unique propositional states and *movements* map to transitions between states. An example of the resulting LTS is shown in Figure 3.

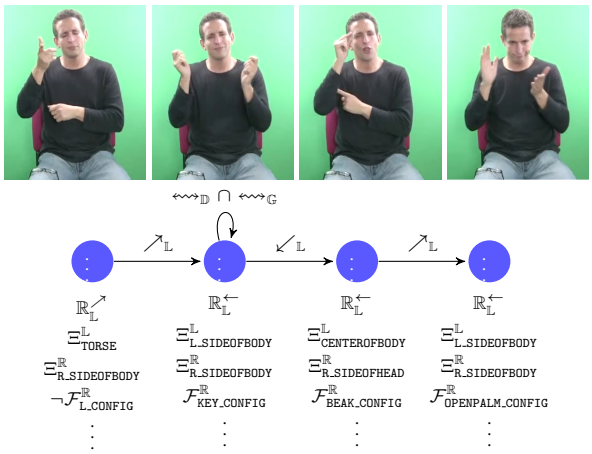


Figure 3: Example of modeling over four automatically identified frames as possible key postures

Finally, the *Verification Module* takes both the generated LTS and a database of PDL$_{\text{SL}}$ formulae to determine which of them are satisfied in the model. As each formula corresponds to a formal description of a sign or property, the module can use logical satisfaction to verify if the property is present or not in the video. The complete process is shown in Figure 4. Finally, the system maps each state

where a formula is satisfied to its corresponding frame interval, so as to generate an annotation proposition.

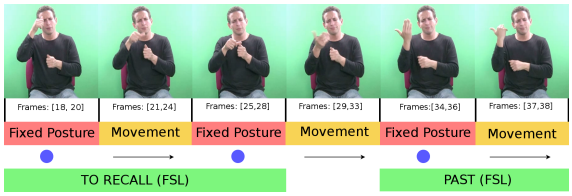


Figure 4: Example of the different layers processed by an automatic annotation system

4.1 Observer Architecture Design

In order to be able to adapt dynamically to the particular needs of the input data, we devised the observer architecture shown in Figure 5.
The main idea behind this design rests upon two axes:

- the possibility of using several tracking tools, adapted to different kinds of corpora;
- the generation of PDL_{SL} models consistent with the information generated by the different trackers.

Moreover, not only do models have to be consistent with every tracker but, as previously stated, not all trackers will give the same information nor track the same features. As such, the framework has to coordinate the loading of the proper modules depending on the corpus and the trackers. This process is entirely done by way of event-triggering. The same mechanism enables communication between modules by implementing multiple-reader/single-writer (MRSW) buffers, which allow every module to read their information but let only one of them write modifications. Each time a new modification is written in a MRSW register, an event is issued system-wide to notify of the existence of new information. This event is then available to every module listening to that register's notifications. For the sake of compatibility, modules are obliged to implement an internal listening thread which can be subscribed to the communication channels of any other module.
In general, the framework establishes development guidelines for the modules of the basic architecture, the one shown on Figure 2, so we can adapt them to specific cases without breaking compatibility. This is achieved by way of generic templates that implement the most basic functionalities of every module. These templates can later be extended to cover the specific cases arising in research; a developer can simply override the critical functionality in each template with their own code. Additionally, modules can register new events within the framework, so as to convey further information (if needed) for particular cases. As such, the system is capable of distributing self-contained, interchangeable, modules that can adapt to different situations.
The execution process is also fairly straightforward. At the beginning a *Start* event is fired-up, prompting to load both a video stream and a tracker. This corresponds to the *Tracking and Segmentation Module* on the basic architecture (Figure 2). The system chooses between the compatible video inputs and pairs the selection with the proper tracker. This is done by reading the events sent out by the loading functions. Likewise, the model construction rules are loaded after a compatible set of video/tracking inputs has been selected. In this way, we can assure that the modeling algorithm will only take in account pertinent rules, those relying on the specific features we are tracking. This mechanism avoids generating models based on hand positions, for example, if our tracker is only capable of detecting non-manuals. Once a compatible set of modules is activated, the process can continue as proposed by (Curiel and Collet, 2013).

5. Experimental Results

We obtained some preliminary results on the proposed framework by implementing the system's core and a set of minimal templates for each of the modules on Figure 2. The core contains the necessary data structures to represent both PDL_{SL} models and formulae, alongside the semantic rules necessary to acknowledge logical satisfaction.
For the creation of the module templates, we considered two possible scenarios:

- the system is being used to annotate previously captured video corpora;
- a camera as going to be used as input for real-time sign recognition.

Furthermore, we had to consider two distinct cases when treating video; whether we had 2D or 3D information available for determining relationships between hands and body. For simplicity, we worked only with hand-tracking data. Nevertheless, the addition of non-manual trackers is also a possibility, since introducing new modeling rules for non-manuals follow the same principles of the 2D to 3D transition.
Once all the framework tools were in place, we created a specific implementation for the 2D case, when tracking features over existing corpora.

5.1 Automatic Annotation in 2D Corpora

To obtain some initial results over real-world data, we developed the first modules based on the atoms originally presented with the PDL_{SL} language. Additionally, we created a property database made of PDL_{SL} formulae, adapted to be used with our tracking device. The database position in the architecture is shown in Figure 5, as the node *Lexical Formulae*. The formulae were exclusively constructed for the 2D case; this means that, for any other kind of tracking information, we would need to define new PDL_{SL} database with different properties. For tracking, we used the tracker developed by (Gonzalez and Collet, 2011), which is capable of finding 2D positions of the hands and head over SL video corpora. As for the SL resources, we used an instance of the *DictaSign* corpus (DictaSign, 2012) as video input for our system.
Since the used tracking tool is not adapted for real-time processing, the implemented tracking module just recuperates

the previously calculated information from an output file. This is done sequentially, after each successful querying of a new video frame, to simulate real-time.

To calculate the posture segmentation we used the method proposed by (Gonzalez and Collet, 2012), which is based on measuring speed-changes.

Our $\mathrm{PDL}_{\mathrm{SL}}$ description database contains four structures:

opposition. $\lambda c.(\mathtt{right}^{\leftarrow}_{\mathtt{left}} \wedge hands\_config(c))$. Hands are opposite to each other, with the same configuration.

tap. $\lambda s, w.(\neg\mathcal{T}^s_w \rightarrow [moves(s) \cup moves(w)]\mathcal{T}^s_w \rightarrow [\mathbf{skip};\mathbf{skip}]\neg\mathcal{T}^s_w)$. Hand touches briefly the other hand, only for a single state.

buoy. $\lambda s, posture.(posture \wedge [moves(s)^*]posture)$. The state of one hand remains the same over several states, regardless of the movements of the other hand.

head anchor. $\lambda s, w, posture.(\mathbf{buoy}(s, posture) \wedge \mathcal{T}^{\mathtt{head}}_w)$. One of the hands remains within the head region while the other signs.

The $posture$ variable denotes the propositional state of an articulator. The $moves(s)$ function can be interpreted as any action executed by articulator s. We omit the complete, formal definition of this operator for the sake of readability.

To measure the *hit* ratio of the system, we manually annotated the apparition of the described properties in one video within the corpora. Table 3 shows the quantity of observed apparitions of each property over the chosen video.

φ	**oppos.**	**buoy**	**tap**	**h. anch.**
Total	76	40	33	74

Table 3: Manually annotated apparitions of property formulae on one video

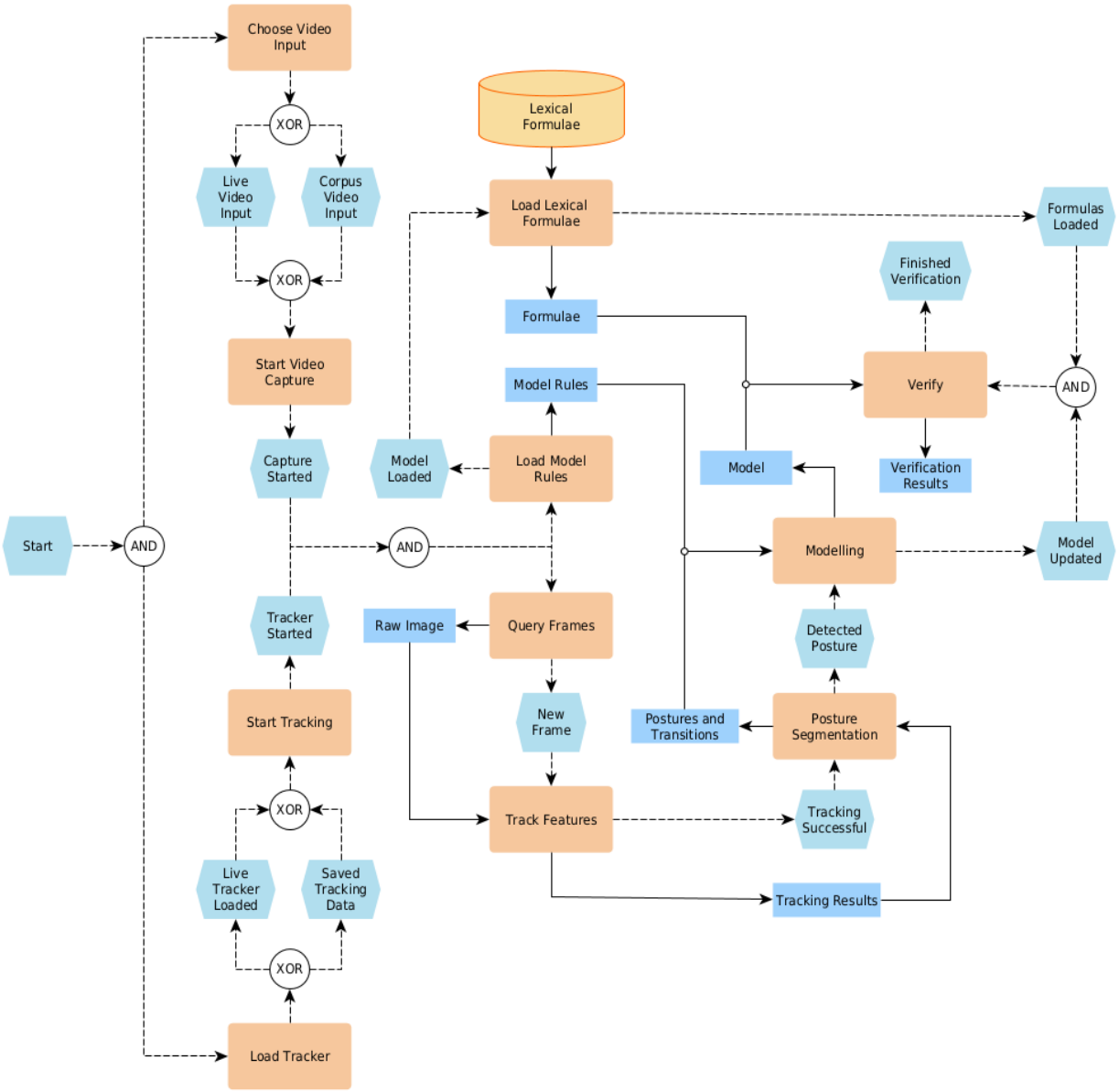


Figure 5: Information and control flow in the SL annotation framework

For each signer, the system creates a model based only on the atoms specified by the modeling rules. It then uses the created model to verify every formula on-the-fly. The execution of our algorithm over the same video rendered the results shown in Table 4.

φ	**oppos.**	**buoy**	**tap**	**h. anch.**
Total	164	248	79	138

Table 4: Total reported hits of property formulae on one video

On Table 4 we can see the total number of times each of the formulae were verified on the video, as returned by the system. We compare the human annotations with these results on Figure 6.

TOTAL OBSERVATIONS

φ	**oppos.**	**buoy**	**tap**	**h. anch.**
By hand	76	43	33	74
Automatic	164	245	79	138

Human \ Automatic	φ	**oppos.**	**buoy**	**tap**	**h. anch.**
	oppos.	67	64	10	33
	buoy	22	40	7	17
	tap	15	24	25	11
	h. anch.	23	50	13	44
	False P.	37	67	24	33

Figure 6: Formulae verification results

Figure 6 shows data from both Tables 3 and 4, alongside a *matching table* where, for each property formula, we count the quantity of times it was verified on previously human-annotated frames. Each row represents the total number of human observed apparitions of one property, while each column represents the quantity of positive verifications reported by the system for each formula. For example, cell (**opposition**, **opposition**) shows the total number of times the **opposition** formula was correctly verified on human-annotated **opposition** frames. The next cell, (**opposition**, **buoy**), holds the number of times the **buoy** property was verified on human-annotated **opposition** frames. Positive verifications can overlap, *i.e.* the system could have verified two or more formulae over the same states of the model. Therefore, a single annotation could belong to different classes. The cells on the last row of the table correspond to *false positives*, reported detections that don't overlap with any human observation.

Further analysis on the matching table is represented on Table 5, which shows the total number of correctly and incorrectly classified formulae, as well as the total mismatches.

The results show a high recognition rate for **opposition**, **buoy** and **tap**, but also a high quantity of misclassification hits and false positives. Most of the erroneous hits are due to the definitions of the properties themselves. Take, for example, **opposition** and **buoy** properties. In the video, some of the states satisfying a **buoy** could easily be classified as **opposition**. When this happens, the only thing that differentiates them, if we only have tracking information, is their movement: if a hand is moving is a **buoy**, otherwise is an **opposition**. Even though this is not always the case, sometimes the situation arises and the system confuses these properties for one another; if some of the movements of the hands are too fast, or not ample enough, when performing a **buoy**, the system interprets them as a static posture, therefore classifying some of the internal states of the **buoy** as **opposition**. This, however, doesn't impede finding the buoy, since the definition of **buoy** specifies, from the beginning, an arbitrary number of internal states, hence not affected by having found one instead of two distinct states. The opposite case might also arise, when a short involuntary movement, is interpreted by the system as an intended action instead of noise, hereafter classifying an **opposition** as a **buoy**, or even as two sequential **oppositions**. Similar arguments can be made for **tap** and **head anchor**, where movement thresholds alone can affect the form and the quantity of states on the LTS. In the future, we expect that adding new information will reduce the quantity of misclassified data, specially because this will result in a more fine-grained model from the beginning.

φ	HUMAN OBS. HIT	HUMAN OBS. MISS	ERRONEUS MATCH
opposition	67	9	107
buoy	40	3	46
tap	25	8	50
h. anchor	44	30	86

Table 5: Per-formula summary of the total number of observations found, missed and erroneously classified observations

At this stage, the system returns a list of proposed properties as result of the verification phase. What the numbers on Table 5 mean is that, in most cases, the proposed annotation will almost never return single properties but rather sets of properties. This may not be a problem with simple formulae like the ones described, but would be problematic with complete sign descriptions; there is such thing as too much information. In that case, we would need a human being to complete the classification process. This points out the need or a higher level module in charge of cleaning the annotation proposal by way of machine learning techniques.

Finally, most of the false positives that don't correspond to any overlap with human observations were caused by signer's movements without communication intent. For example, some **opposition** properties were found when a signer crossed his arms, when his hands were posed over his knees or when he assumed other natural repose positions. Similarly, some co-articulatory movements created chains of states that satisfied the formulae for **buoy** or **tap**. These cases could also be reduced with help of a higher level module or a human expert.

5.2 Extending to 3D

Currently, we are extending the system to model features tracked in 3D. We have already extended the framework to process data returned by the Kinect (Microsoft, 2013),

a motion sensing device capable of tracking 3D positions on several body articulations. Figure 7 shows the points that can be tracked by using the Kinect with it's official development kit.

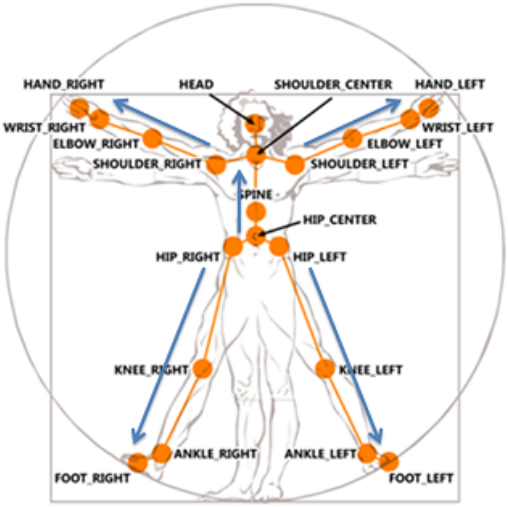


Figure 7: Points tracked by the Kinect device (Microsoft, 2013)

For the moment, we have been able to reuse the same modeling rules that we implemented for the 2D case; mainly, we have used the Kinect tracker to obtain 3D position data of hands and head, and we have projected this information in 2D. This lets us create the same kind of models we build from corpora. However, the variety of the tracked articulations and the 3D capabilities of the sensor, call for the definition of more complex atoms and lambda properties, as well as 3D descriptions of individual signs. As of 2014, work is still ongoing on the matter and has not been properly evaluated. Nevertheless, we considered important to point out that we can already exchange trackers if needed, so as to showcase the flexibility of our framework.

6. Conclusions

Here we have presented an automatic annotation framework for SL based on a formal logic. The system lets us represent SL video inputs as time-ordered LTSs by way of $\mathrm{PDL}_{\mathsf{SL}}$, a multi-modal logic. We have shown that it is possible to use the resulting graph to verify the existence of common lexical structures, described as logical formulae. Furthermore, the framework gives us the necessary tools to adapt the model generation for different corpora and tracking technologies.

From the point of view of recognition, we noticed that the quality of the tracking tools is of utmost importance for both formula definition and model generation. The low presence of information and high levels of noise immediately took a toll on verification; in some cases, we lacked enough information to distinguish between intended movements and noise. In turn, this resulted on high rejection rates of what would otherwise be considered *hit* frames.

Similarly, we noticed that modeling can be affected by the presence of low information, which can render states indistinguishable. For instance, without hand configurations every state satisfying **opposition** is, effectively, the same state. Therefore, every formula sharing the same **opposition** base would be satisfied on that single state. This could gravely affect the system's performance; in the worst case, all states could satisfy all formulae. On the other hand, a too fine-grained model can lead to a LTS that replicates the same problems we have in synthesis-oriented descriptions. In that case, we would need very specific formulae (with near to perfect SL corpora) to achieve any identification at all. Similarly, formula creation can't be neither too broad nor too specific, if we want to minimize the quantity of imperfect matches. Anyhow, one of the advantages we have by using a logical language is that we can control the granularity of information simply by defining or discarding atoms, which opens the door to the use of algorithmic techniques to control information quantity.

From the point of view of the implementation, the results of the 2D experiments show that further effort has to be put on integrating new sources of information to the system, especially if we want avoid false positives. Even though the system is in place and works as expected, the high quantity of erroneous hits reflects the gravity of the problems we can have with indistinguishable states. Further comparisons have to be done once the system completely incorporates 3D modeling, so as to measure the effective impact of additional information on verification.

Future work in recognition will be centered on implementing machine learning techniques to improve verification. Using data analysis to find relationships between detected structures, could lead us to better results even in suboptimal environments. Additionally, we would like to integrate communication with user level software like the one presented by (Dubot and Collet, 2012), a manual annotation tool. This could lead to other possible uses of the framework as engine for higher applications, such as dictionary searching or even for automatic creation of sign description databases from SL videos.

Further analysis will also target the building blocks of the language, by changing the semantic definitions of $\mathrm{PDL}_{\mathsf{SL}}$ operators to better suit SL. Changes to its syntax are also to be expected, in an effort to ease the development of extensions for different trackers and simplify descriptions. Finally, we want to steer further into 3D representation and the inclusion of non-manual features, important stepping stones towards higher level language processing.

7. References

Aronoff, M., Meir, I., and Sandler, W. (2005). The paradox of sign language morphology. *Language*, 81(2):301.

Barendsen, H. B. E. (1994). Introduction to lambda calculus.

Bossard, B., Braffort, A., and Jardino, M. (2004). Some issues in sign language processing. In Camurri, A. and Volpe, G., editors, *Gesture-Based Communication in Human-Computer Interaction*, volume 2915 of *Lecture Notes in Computer Science*, pages 90–100. Springer Berlin Heidelberg.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In Moeslund, T. B., Hilton, A.,

Krüger, V., and Sigal, L., editors, *Visual Analysis of Humans*, pages 539–562. Springer London.

Curiel, A. and Collet, C. (2013). Sign language lexical recognition with propositional dynamic logic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 328–333, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dalle, P. (2006). High level models for sign language analysis by a vision system. In *Workshop on the Representation and Processing of Sign Language: Lexicographic Matters and Didactic Scenarios (LREC), Italy, ELDA*, page 1720.

DictaSign. (2012). http://www.dictasign.eu.

Dreuw, P., Stein, D., Deselaers, T., Rybach, D., Zahedi, M., Bungeroth, J., and Ney, H. (2008). Spoken language processing techniques for sign language recognition and translation. *Technology and Disability*, 20(2):121–133.

Dubot, R. and Collet, C. (2012). Improvements of the distributed architecture for assisted annotation of video corpora. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon.*, pages 27–30. Language Resources and Evaluation (LREC).

Filhol, M. (2008). *Modèle descriptif des signes pour un traitement automatique des langues des signes*. Ph.D. thesis, Université Paris-sud (Paris 11).

Filhol, M. (2009). Zebedee: a lexical description model for sign language synthesis. Internal, LIMSI.

Fischer, M. J. and Ladner, R. E. (1979). Propositional dynamic logic of regular programs. *Journal of Computer and System Sciences*, 18(2):194–211, April.

Gonzalez, M. and Collet, C. (2011). Robust body parts tracking using particle filter and dynamic template. In *2011 18th IEEE International Conference on Image Processing (ICIP)*, pages 529 –532, September.

Gonzalez, M. and Collet, C. (2012). Sign segmentation using dynamics and hand configuration for semi-automatic annotation of sign language corpora. In Efthimiou, E., Kouroupetroglou, G., and Fotinea, S.-E., editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, number 7206 in Lecture Notes in Computer Science, pages 204–215. Springer Berlin Heidelberg, January.

Kervajan, L., Neef, E. G. D., and Véronis, J. (2006). French sign language processing: Verb agreement. In *Gesture in Human-Computer Interaction and Simulation*, number 3881 in Lecture Notes in Computer Science, pages 53–56. Springer Berlin Heidelberg, January.

Liddell, S. K. and Johnson, R. E. (1989). *American sign language: The phonological base*. Gallaudet University Press, Washington. DC.

Losson, O. and Vannobel, J.-M. (1998). Sign language formal description and synthesis. *INT.JOURNAL OF VIRTUAL REALITY*, 3:27—34.

Meir, I., Padden, C., Aronoff, M., and Sandler, W. (2006). Re-thinking sign language verb classes: the body as subject. In *Sign Languages: Spinning and Unraveling the Past, Present and Future. 9th Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil*, volume 382.

Microsoft. (2013). Tracking users with kinect skeletal tracking, http://msdn.microsoft.com/en-us/library/jj131025.aspx.

Ong, S. and Ranganath, S. (2005). Automatic sign language analysis: a survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873 – 891, June.

Wittmann, H. (1991). Classification linguistique des langues signées non vocalement. *Revue québécoise de linguistique théorique et appliquée*, 10(1):88.

Creation of a multipurpose sign language lexical resource: The GSL lexicon database

Athanasia-Lida Dimou[1], Theodore Goulas[1], Eleni Efthimiou[1], Stavroula-Evita Fotinea[1], Panayiotis Karioris[1], Michalis Pissaris[1], Dimitris Korakakis[1], Kiki Vasilaki[2]

[1]ILSP - R.C "Athena", [2]Aristotle University of Thessaloniki – Philology Department
[1]Artemidos 6 & Epidavrou, Maroussi, 15125 Athens, Greece
E-mail: {ndimou, tgoulas, eleni_e, evita, pkarior}@ilsp.gr, pissarakia@gmail.gr, korakakis79@gmail.com, kikivasilaki@yahoo.gr

Abstract

The GSL lexicon database is the first extensive database of Greek Sign Language (GSL) signs, created on the basis of knowledge derived from the linguistic analysis of natural signers' data. It incorporates a lemma list that currently includes approximately 6,000 entries and is intended to reach a total number of 10,000 entries within the next two years. The design of the database allows for classification of signs on the basis of their articulation features as regards both manual and non-manual elements. The adopted information management schema accompanying each entry provides for retrieval according to a variety of linguistic properties. In parallel, annotation of the full set of sign articulation features feeds more natural performance of synthetic signing engines and more effective treatment of sign language (SL) data in the framework of sign recognition and natural language processing.

Keywords: GSL, SL lexicon, manual feature, non-manual features, sign articulation, SL technologies, SL data acquisition

1. Introduction

Here we present the methodology followed in creating a multipurpose lexical data base of the Greek Sign Language (GSL) which currently incorporates approximately 6,000 sign entries and it is intended to reach a content of 10,000 entries in the next two years. The main effort is been placed on creation of an extensive resource of sign lemmas which may serve a variety of goals, including extraction of bilingual dictionaries/ glossaries, incorporation of lexical information in natural language processing (NLP) systems as in the case of machine translation (MT) from and into sign language, creation of training material for sign recognition technologies, and input to sign synthesis tools enabling signing by virtual signers (avatars).

Given the scope of the resource and the range of usability cases it is intended to serve, design criteria which had to be satisfied extend from naming conventions to coding of manual and non-manual elements of each sign for representation via synthetic signing and retrieval purposes.

The GSL lexicon database in its current status has been created by integrating two different available lexical resources after careful content evaluation and thorough revision of the previously available database structure design.

In the rest of the paper, we report on the methodological milestones and undertaken actions that the reported attempt required, as well as the procedures that are planned to be carried out next in order to extend the database content. In this framework, an initial study of available data has revealed considerable participation of non-manual features in GSL sign formation, while in many cases, non-manuals disambiguate the meaning of lemmas articulated by means of the same manual activity (see also Section 4 below). Thus, annotation of non-manual elements of signs becomes a central task in the current attempt, given the need to fully code articulation features of sign lemmas to equally support SL data computing and synthetic signing needs, parallel to the al use of the lexicon in communication and education context.

2. Exploited resources for the GSL lexicon database

The main resources used for the creation of the GSL lexicon data base derive from two different sources, i) the content of the bilingual (GSL-Modern Greek) multimedia dictionary NOEMA [1], and ii) the lemmatized GSL DICTA-SIGN[2] corpus. We provide next information on the structure of these two sources, which influenced the design of the GSL lexicon database.

2.1 Multimedia dictionary NOEMA

The NOEMA dictionary is the first electronic dictionary of GSL signs and contains 3,000 video lemmas of general language falling within the definition of basic lexicon content (Efthimiou & Katsoyannou, 2001). NOEMA is a bilingual dictionary which is aimed to provide structured knowledge of GSL lexicon to a large non-specialized audience. It is equally addressed to natural deaf GSL signers and to hearing individuals who are interested in learning GSL as a second language. Thus, the dictionary organization is intended to serve both groups of end users; to this end every sign has been categorized according to the thematic group it belongs to and is associated with a Greek translation, as well as synonyms and antonyms in

[1]http://www.ilsp.gr/en/services-products/products/item/item/2-noema
[2] http://www.dictasign.eu/

GSL. The NOEMA dictionary has actually been constructed as a tool to support an introductory course in GSL, providing paradigms of all handshapes recorded to be used by the language in basic vocabulary concepts.
One of the assets of NOEMA has been the search option in the dictionary content by means of a selected handshape or a combination of handshapes (Figure 1). The latter has been accomplished by annotation on the dictionary database for main as well as secondary handshape(s) used in sign formation for all its lemmas.

Figure 1: Interface for handshape based search option in the NOEMA dictionary

The video lemmas that comprise the NOEMA dictionary provided the content substructure of the GSL lexicon database; the 3,000 signs from the domain of general language constituted a significant core for the creation of the new lexicon. However, prior to transfer of the lemmas to the new database, a thorough evaluation study took place which pointed out a number of significant improvements needed to take place in order to optimize the data acquisition process, currently under development.
The list of enhancements in respect to the content available in NOEMA incorporates re-acquisition of lemmas by means of HD and Kinect cameras, corrections in lemma representation where necessary, addition of paradigms of use and coding of the manual and non-manual articulation elements of each sign.
As regards lemma correction, this involves two sets of corrections: i) while recording predicative lemmas any indication of declination (as to first person singular), which was often met with predicate GSL lemmas representation in NOEMA, is strictly avoided, and ii) a small number of lemmas which have been recognized to derive via interference from oral/written Greek but are not recognized as an integral part of the GSL vocabulary have been omitted from inclusion in the new vocabulary list.
All lemmas are acquired alongside with paradigms of use which aim at clarifying the represented concept and demonstrating all possible contexts of use of a specific lemma. Lemmatization of the utterances which serve as paradigms of use adds new lemmas to the initial lexicon which is significantly augmented via this process.

Another category of lemmas which is not transferred in the new database as it used to appear in NOEMA, involves association of classifiers with a specific equivalent lemma in Greek, as it has been i.e. the case of associating classifier C with the Greek lemma for PIPE. In the current framework, classifiers are treated as a class of entities associated with specific semantic properties and only those cases which are identified by native GSL signers as related to a specific concept without the need for associating their interpretation with information previously provided in their linguistic context, are treated as autonomous lemmas. Thus, in the currently adopted design, classifiers which have not been lexicalized are studied within their signed context and are treated in the lexicon either as bound morphemes or as semantic indicators with pronominal function.

2.2 Lemma extraction from an annotated corpus

Complementary to lemmas deriving from NOEMA, the GSL lexicon database has also been enriched by lemmas extracted from the annotated GSL segment of the Dicta-Sign corpus[3].
The corpus created during the Dicta-Sign project (Matthes et al., 2010; 2012) made available natural discourse productions in four SLs: Greek, German, French and English, to a significant extend fully annotated for the entailed lemmas in the ilex[4] (Hanke & Storz, 2008) annotation environment by means of the HamNoSys notation system (Hanke, 2004; Prillwitz et al., 1989). Lemma annotation of the Greek segment of the corpus enriched the GSL lexicon database with approximately 2,000 lemmas.
Creation of the Dicta-Sign corpus intended to elicit naturally produced signing, hence the elicitation procedures were carefully designed so as to promote naturalness of the acquired data. The outcome of the related data acquisition process was a corpus rich in continuous signing information markers, incorporating in-context lemma productions. In terms of the currently developed GSL lexicon, the sign lemmas deriving from the Dicta-Sign corpus need to be enhanced in respect to speed of production and co-articulation effects during the new acquisition process.
However, searching in the corpus for lemma extraction proved to be valuable for also providing a wide spectrum of use cases related to each lemma.
Furthermore, the study –currently in progress– on extraction and classification of the GSL classifiers system beyond the set of lexicalized classifier items referred to above, is heavily based on annotated data deriving from the same corpus.

3. Compilation of the GSL lemma list

In order to provide content to a common database, both

[3] http://www.sign-lang.uni-hamburg.de/dicta-sign/portal/lang_inform.html.
[4] www.sign-lang.uni-hamburg.de/ilex

sets of NOEMA and Dicta-Sign data had to be unified. The required lemmatization procedure ended with identification of 5,500 unique GSL lemma entries. The derived lemma list needed to be checked for corrections and undergo enhancements as indicated in 2.1 and 2.2 above, in order to ensure homogeneity during the new video recordings and the previously completed evaluation as to inclusion/exclusion criteria, applied to each lemma before its addition to the GSL lexicon database. Other decisions relate to the way compounds are treated depending on whether they are formed via combination of only free or free and bound morphemes, the provisions made with respect to GSL vs. oral Greek synonyms for the representation of a specific concept, and the coding of non-manual articulation features. Compounding has been decided to initially be addressed on the basis of a continuum from productive to lexical compounds approach (Liddell & Johnson, 1986), also adopbted by (Sandler & Lillo–Martin, 2006). Lemma corrections against intuitive GSL linguistic knowledge and selection of paradigms of use have been undertaken by two GSL natural signers, members of the development team supported by a team of three SL linguists.

Compilation of the "GLOSS" field of the database against a lemma list of Modern Greek revealed several one to many GSL to Greek alignments. Since within the scope of this lexicon is to provide for a wider semantic association of concepts and representations between GSL and Modern Greek, the need for the development of a linking mechanism that will enable proper lemma association in the two languages and will also effectively support lexical retrieval and sign language NLP applications has become obvious and related on-going experiments will be published in the next period.

4. Non-manual features

Work is SL linguistics has long recognised the importance of non-manual markers in the articulation of a sign. Non-manuals are considered to be an integral part of sign articulation when they participate along with manual activity in sign formation, and for this reason they have to be specified in the lexical entry of a sign (Pfau & Quer, 2010).

Figure 2: GSL sign LOVE –non-manual neutral articulation

Figure 3: GSL sign THANK-YOU – head movement and facial features differentiate the signed concept from the flat, with respect to non-manuals sign LOVE

Figure 4: GSL sign MINE (1-Sg-Poss) – head movement and facial features identify the signed concept as differing from concepts THANK-YOU and LOVE

There are two kinds of non-manual markers: facial and non facial. Facial non-manuals occur entirely on the face, while non-facial markers take the form of a particular head or body movement (Neidle et al., 2000).

When they form part of sign phonology, there is a strong tendency for non-manual markers to be synchronized with the manual part of the sign. For example, in articulating the GSL signs HAPPY, SAME and GET-BORED the signs' manual articulation is obligatorily accompanied by a particular facial expression performed in parallel. Moreover, non-manual markers in GSL may distinguish two (or more) otherwise identical signs, i.e. they can define minimal pairs. For instance, the signs LOVE, THANK-YOU and the first person singular of the possessive pronoun (MINE) are all identified by the different non manual signals accompanying the same hand activity as indicated in Figures 2, 3 and 4. Similarly, pairs of signs are very often distinguished by non-manual articulation elements, like the signs BE-CRAZY ABOUT and COMMIT SUICIDE which are minimally distinguished by facial expression.

Non-manual features are systematically addressed in respect to the lexicon under development as according to

the design specifications of the GSL lexicon database. They are dedicated a separate section in which the presence or absence of facial and body features are annotated and accordingly demonstrate critical alternations in the meaning of a manually signed or a classifier entity (Efthimiou et al., 2010).

Furthermore, coding of signs in respect to non-manuals is ranked as equally important for synthetic signing as manual features coding. Incorporation of non-manuals is directly related to the degree of achieved naturalness and related acceptance of synthetic signing by Deaf communities in general. In our case, it is a prerequisite for exploiting the reported resource in teaching and communication environments which consume language technologies.

Enrichment of lemmas with annotations for both manual and non-manual features is facilitated by a dedicated section in the Sis-Builder[5] tool (Goulas et al., 2010).

For the facilitation of assignment of HamNoSys notation symbols to manual activity involved in formation of a specific sign, the environment provides virtual keyboards for the marking of symmetries, handshape, hand position, hand location and motion actions, partly shown in Figure 5.

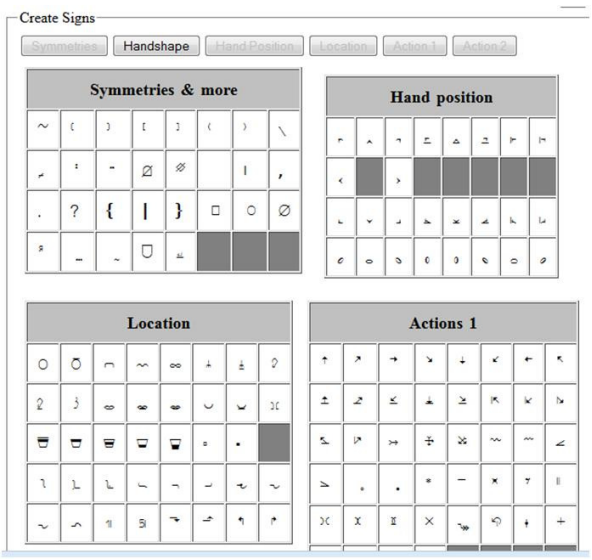


Figure 5: Virtual keyboards for the annotation of manual activity in sign formation

Non manual elements of sign formation are coded in SiS-Builder by selection from a drop-down menu of values for all possible facial and body features which participate in sign articulation parallel to manual activity. Figures 6a and 6b depict the set of non manual features taken into account for coding and the way coding takes place via selection from the available drop-down menus. Annotation of signs in respect to both their manual and non manual articulation parameters (Figure 7) provides the necessary information for their more natural synthetic representation. In fact, this information is crucial for a range of applications in the area of SL processing, focussing on improvement of retrieval and sign recognition results. Nonetheless, completeness in representation of articulation features of signs is also crucial in SL linguistics research and SL learning environments equally in the framework of treating SL as first or second language.

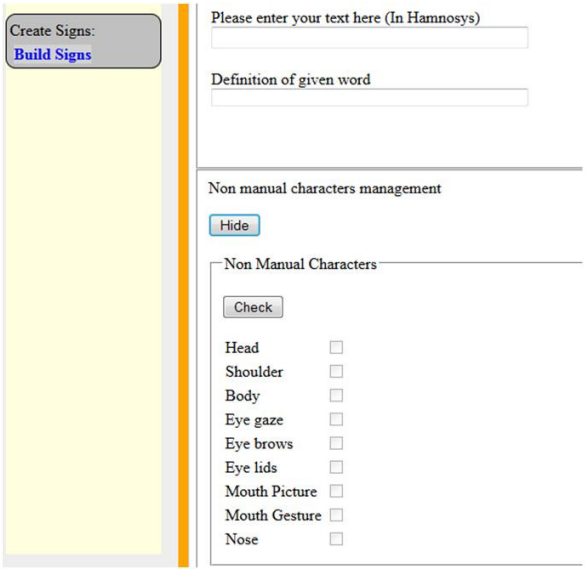


Figure 6a: The set of non manual features handled by the SiS-Builder environment

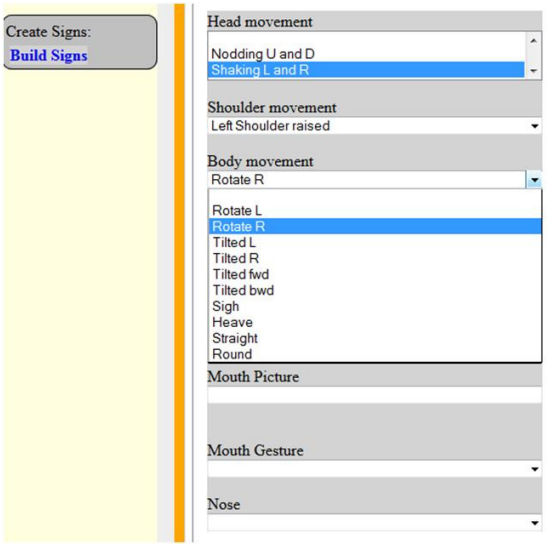


Figure 6b: Non manual feature values assignment via drop-down selection

5. Data acquisition methodology and set-up

Recording sessions follow a predefined script which includes the lemmas to be acquired each time along with the set of usage examples accompanying each lexical entry, which are selected on the basis of linguistic-lexicographic criteria to satisfy demonstration of semantic/syntactic properties of the lemmas.

[5] http://speech.ilsp.gr/sl

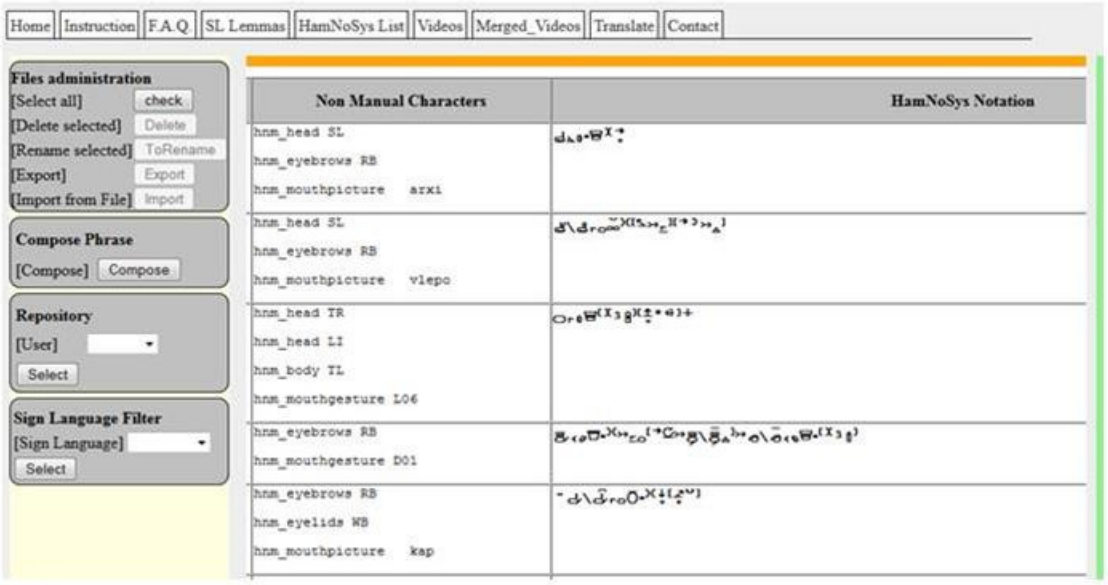


Figure 7: Manual and non-manual activity annotation on sign lemmas of GSL in the SiS-Builder environment

The data acquisition team is composed of the engineer who controls data flow from the acquisition devices, the studio officer who is in charge of the studio set-up and cameras control, an interpreter/facilitator who supports the informant, and a native signer who performs the scheduled lemmas and their paradigm of use phrases in three (3) repetitions each. Prior to recording, the team members need to study the lemmas to be captured and decide on their representative paradigms of use, if such paradigms are not already available in the GSL corpus. During capturing, the predefined list of lemmas which falls within the session's schedule is projected to the informant by means of a monitor.

The examples which are associated with each lemma are noted down in a note taking environment in the form of "written GSL", completely avoiding the use of subtitles in Greek language, in order to provide an easy to check list of all signs that are contained in the usage examples and also diminish oral language interference effects in the grammar of the paradigm utterances. Lemmatization of the newly produced paradigm of use utterances is intended to ensure that all signs used in the example phrases are also incorporated in the lemma list, thus using this qualitative control also as a means of augmenting the lexicon with new lemmas.

GSL lemmas are realized in isolations, in a clear, comprehensible manner. Examples of use are preferably small, simple phrases that demonstrate each sign's proper linguistic use. Examples need not be performed flat (in a dry manner), although non-manual markers that are related with a high degree of emotion demonstration on sentence level are advised to be left out for avoidance of confusion as to the proper sign articulation.

Recordings take place at a high-end technology studio (Figure 8) that provides all necessary facilities (lighting, storage media, microphones, cameras) for HD quality recording. In terms of data acquisition equipment, one HD camera (front view) and one Kinect camera (for depth information of sign articulation) are used. The synchronisation of these media is accomplished via clapping[6] as audio cue and flashing as visual cue.

Figure 8: Lexicon acquisition studio set-up

6. Conclusion

The GSL lexicon database is an ambitious project, opting for the creation of a multipurpose resource of at least 10,000 distinctive GSL lemma entries, mainly addressing SL processing needs in the framework of human language technologies applications and also in service of SL recognition and synthetic signing technologies. Thus, exhaustive coding of lemmas for their manual and non-manual features is a major task. In this context, association of lemmas within an appropriate ontology scheme is required to enable more efficient bilingual associations between GSL and Modern Greek, which will significantly augment accessibility of written Greek texts

[6]Microphones are typically used in multicamera data acquisition to capture clapping signals which are exploited in synchronization of the different video streams.

by Deaf individuals allowing for more effective language engineering solutions in a variety of communicative environments. These involve machine translation, meaning spotting and retrieval of information from a written text source, facilitation of visual processing and SL synthesis, the goal being to achieve proper linking of a sign with an equivalent word in Modern Greek, but also with all its available synonyms and the range of related hypernyms and/or hyponyms.

In parallel, systematic categorization of non-manual features of sign articulation is expected to lead to a more concrete definition of the linguistic function of non-manuals in GSL sign formation, as well as to higher acceptance of synthetic signing (Jennings et al., 2010), since sign synthesis engines which take non-manuals into account improve significantly in respect to naturalness of signing performance (Figure 9).

Finally, a resource providing the qualitative and quantitative range of information incorporated in the GSL lexicon, will be of value also to GSL language education both in respect to first and second language learning.

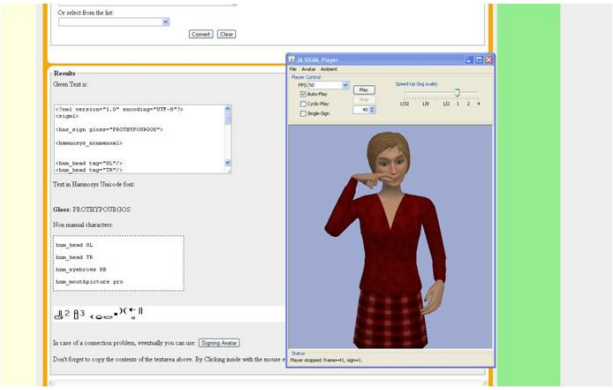


Figure 9: Avatar performance incorporating non manuals as annotated in the SiS-Builder environment for the GSL lemma PRIMEMINISTER

7. Acknowledgements

The research leading to these results has received funding from POLYTROPON project (KRIPIS-GSRT, MIS: 448306), based on initial data acquisition in the framework of the Dicta-Sign project (FP7/2007-2013 grant agreement n° 231135).

8. References

Efthimiou E., Fotinea S-E., Dimou A-L., Kalimeris C. (2010). Towards decoding Classifier function in GSL. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta.

Efthimiou, E. & Katsoyannou, M. (2001). Research issues on GSL: a study of vocabulary and lexicon creation. In *Studies in Greek Linguistics, Computational Linguistics* 2: 42-50 (in Greek).

Goulas, T., Fotinea, S-E., Efthimiou, E. and Pissaris, M. (2010). SiS-Builder: A Sign Synthesis Support Tool. In Dreuw, P. et al. (Eds.), LREC-2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 102--105.

Hanke,T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In O. Streiter and C. Vettori (Eds.), LREC-2004. *Proceedings of 1st Workshop on Representing and Processing of Sign Languages,* pp. 1--6.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Construction and Exploitation of Sign Language Corpora. Proceedings of 3rd Workshop on the Representation and Processing of Sign Languages*. ELRA, Paris, pp. 64--67.

Jennings, V., Elliott, R., Kennaway, R. and Glauert, J. (2010). Requirements for a Signing Avatar. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta, pp. 133--136.

Klima, E., and Bellugi, U., (1979). *The signs of language*. Harvard University Press, USA.

Liddell, S. and Johnson, R., (1986). American Sign Language Compound Formation Processes and Phonological Remnants. In *Natural Language and Linguistic Theory,* vol.4, Reideil Publishing Co, pp. 445--513.

Matthes S., Hanke T., Regen A., Storz J., Worseck S., Efthimiou E., Dimou A.-L., Braffort A., Glauert J. and Safar. E. (2012). Dicta-Sign – Building a Multilingual Sign Language Corpus. In Crasborn et al. (Eds.), LREC 2012, *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Istanbul, Turkey.

Matthes S., Hanke T., Storz J., Efthimiou E., Dimou A-L, Karioris P., Braffort A., Choisier A., Pelhate J., Safar E. (2010). Elicitation tasks and materials designed for Dicta-Sign's multi-lingual corpus. In Dreuw, P. et al. (Eds.), LREC 2010, *Proceedings of 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valetta, Malta.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., et Lee, R.G. (2000). *The Syntax of American Sign Language. Functional Categories and Hierarchical Structure*. Cambridge, MA: MIT Press.

Pfau, R., & Josep, Q. (2010). Nonmanuals: their grammatical and prosodic roles. Sign Languages, In D. Brentari (Ed), pp. 381--402. Cambridge: Cambridge University Press.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T. and Henning, J. (1989). HamNoSys. Version 2.0. *Hamburg Notation System for Sign Language: An Introductory Guide*. Signum Verlag, Hamburg.

Sandler, W. & Lillo – Martin, D. (2006) "*Sign Language and Linguistic Universals*", Cambridge University Press, UK, pp.72.

A hybrid formalism to parse Sign Languages

Rémi Dubot, Christophe Collet

IRIT
Université de Toulouse
France
dubot@irit.fr, collet@irit.fr

Abstract

Sign Language (SL) linguistics is dependent on the expensive task of annotation. Some automation is already available for low-level information (eg. body part tracking) and the lexical level has shown significant progresses. The syntactic level lacks annotated corpora as well as complete and consistent models. This article presents a solution for the automatic annotation of SL syntactic elements. It exposes a formalism able to represent both constituency-based and dependency-based models. The first enables the representation of structures one may want to annotate, the second aims at fulfilling the holes of the first. A parser is presented and used to conduct two experiments to test the solution. One experiment is on a real corpus, the other is on a synthetic corpus.

1. Introduction

To study Sign Languages (SLs), linguists need annotations. Currently, corpus annotation is done manually, it is time-consuming and suffers difficulties with inter and intra-annotator reliability. For this reason, efforts are carried out to automatize the annotation process. Early efforts focused on the very low-level non-linguistic information: body part tracking, activity detection. They finally reached the base of the linguistic level: detection of sign phases (Gonzalez and Collet, 2011), sub-lexical (Cooper et al., 2012) and lexical units (Curiel and Collet, 2013). Work on this last level has focused on manual gestures. The only exceptions were attempts to remove ambiguity on some lexical signs with the help of Non-Manual Gestures (NMGs) (Paulraj et al., 2008) or detection of NMG (Yang and Lee, 2011; Neidle et al., 2009). Now is the time to address the annotation of supra-lexical features. But when it comes to syntactic features, it is not possible to ignore the NMGs anymore.

The syntax SLs is complex and different from vocal languages(Cuxac, 2000; Dubuisson et al., 1999; Bouchard and Dubuisson, 1995; Bouchard, 1996). They use the multiplicity and the spatial abilities of the available articulators. It results non-sequential productions with complex temporal, spacial and articulatory synchronizations. The syntactic models developed for the processing of vocal languages are deeply based on the sequentiality of lexical units. Consequently, the processing of SL syntax requires the invention of new models or, at least, to deeply rethink and adapt the existing ones.

A recognition system always has an internal representation of the phenomena to recognize. However, there are multiple manners to obtain such a representation. From one extreme to another, it can be expert knowledge formalized into a model or it can be results of uninformed automatic learning on real data. The first requires experts to formalize a complete and consistent model from their knowledge. The second requires massive data and computer calculation. For the syntax of SLs, neither is available. The expert knowledge is sparse and sometimes inconsistent. Annotated SL corpora are too small and too heterogeneous for uninformed learning.

Our goal is to develop tools for the semi-automatic annotation. The general approach we adopt is to use supra-lexical/syntactic models for the annotation. It targets two objectives. First, it aims at producing annotations for all the structures of the model. Second, it aims to enhance the lower levels. Indeed, such models can improve two aspects of the quality of the lexical recognition: the results, by re-scoring the lexical candidates, and the efficiency, by informing the lexical layer and thereby reducing the search space. The models are used to propagate the information of the low-level detections.

This article exposes elements in favor of a hybrid parsing of SLs. It presents a formalism able to represent constituency-based structures as well as dependency-based structures. This formalism has been created to represent models combining transfered linguistic knowledge and automatically learned dependencies. The feasibility is demonstrated with a parser in two experiments. First, the parser is run on excerpts of the Dicta-Sign Corpus with a model composed of five structures. Second, synthetic dependency grammars are used to parse synthetic corpora.

Such a hybrid formalism is the solution we found for the lack of annotated corpora and the incompleteness of the available models. We aim at enabling the use of incomplete models transfered from the linguistic knowledge with learned data.

This work tries to avoid hypotheses that would simplify SL processing by making SLs closer of vocal languages but would be unrealistic. In particular, it makes no assumptions such as the predominance of the hands over the other articulators or the existence of a sequential skeleton of the SL locutions. It is based on the ideas introduced by Filhol (Filhol, 2009) to represent structures with the minimal constraints that make them recognizable. This approach enable to naturally represent the complex temporal synchronization mechanisms (Filhol, 2012) of SL simultaneity (Vermeerbergen et al., 2007).

This document is structured as follow. It starts with the presentation of the example used all along the article. The formalism is described jointly with its usage for constituency-based structures. The representation of dependency struc-

tures comes next. After the formalism, the parsing is presented with its general characteristics but without details on its internal algorithm. The last part presents the two experiments, their results and an analysis.

2. Formalism description

The first step toward the automatic annotation is the formal representation of a model. The representation we propose is similar to Context-Free Grammars (CFGs) in that it is a derivational grammar. But it differs from CFGs on three fundamental points. First, the right-hand side of a production rule is not a string of units but a set of units. Second, it introduces the possibility to express constraints between all the units of a production rule. Third, in CFGs, the left-hand side of a production rule is non-terminal symbol. We have no such thing as non-terminal and terminal symbols. We have instead detectable and non-detectable units, and both can be atomic (terminal) or not.

We target the representation of two types of models. In the first, a production rule represents a relation of constituency. It comes from the Phrase Structure Grammars (PSGs) of Chomsky (Chomsky, 1957). In the second, a production rule represents a relation of dependency. It comes from the dependency grammars of Tesnière.

2.1. Constituency structures

2.1.1. Example presentation

We illustrate the description of the formalism with the construction of a constituency-based model from an excerpt of a real corpus.

The excerpt comes from the French Sign Language (LSF) part of the Dicta-Sign corpus (Efthimiou et al., 2010) which is composed of spontaneous dialogs performed by deaf signers. In this excerpt, the informant relates a memory of a journey in Paris visiting the Louvre museum with a friend. In the studied part, he explains to his interlocutor the purpose of the journey –to visit the Louvre– and checks that they share the same sign for Louvre. Figure 1 summarizes the excerpt with a sequence of pictures.

2.1.2. Pattern decomposition

We call pattern a rule representing how a unit comes with others. It is similar to the production rules of CFGs. We usually draw these patterns as trees as shown in figure 2. In the present formalism, we make each pattern correspond to a unit (the inverse is false, it is not an equivalence relation). Consequently, a unit can be the root of at most one pattern for a given model. An atomic unit can be associated to a pattern with only a root. It is the single assumption make about units and patterns in a model. Aside from this, everything is possible. Units can appear several times in the same pattern. Patterns can be recursive, mutually recursive, etc.

The model we are about to introduce contains four patterns observed in the excerpt: a buoy pattern, a "sign check" pattern, a question pattern, and an acknowledgment pattern. These patterns are examples and do not rely on a strong linguistic basis. Stronger models remain to be developed with linguists.

The patterns are described in terms of constituents as shown in figure 2. Their internal arrangement is then described with constraints (section 2.1.4.).

The first described pattern is a buoy (Liddell, 2003). It is visible in figure 1, the left hand of the bi-manual sign TO-VISIT (fig. 1(a)) is maintained all along the excerpt. The pattern is decomposed into three sub-elements: two signs and one locution. The second pattern is an acknowledgment. It happens in figure 1 (g). It is decomposed into two sub-elements: a head node and a sign. The third pattern is a question. It also happens in figure 1 (g), but is less clear on this snapshot. It is decomposed as a marker (eyebrows up) and a locution. The "sign check" is a question and an acknowledgment.

As shown in figure 2, the pattern decomposition can be easily represented as a tree. The sub-elements are patterns which can be decomposed themselves or can be considered atomic in the model. Edges represent a relation of constituency. In a decomposition, multiple elements can be instances of a same pattern. When defining a model, one may need to introduce the same pattern multiple times in a same decomposition. This fact is of particular importance as it highlights that an element, in a decomposition, does not represent a pattern but an instance. As a consequence, the name of a pattern is not sufficient to designate elements without ambiguity. It is therefore necessary to associate each instance with a role name.

2.1.3. Alternatives

Patterns do not allow generalization as all their internal elements are mandatory. As patterns describe compositions, we define an other type of rule to explicitly express alternatives. The same restriction as for patterns applies to the use of a unit as root for an alternative. In the example model, we define a node *Locution* as an alternative between the four patterns (figure 2a). Alternatives appear as rectangle nodes in figures 2 and 4.

2.1.4. Constraints

Patterns and alternatives represent invariants in the composition. Invariants in the internal organization of the patterns are expressed with constraints.

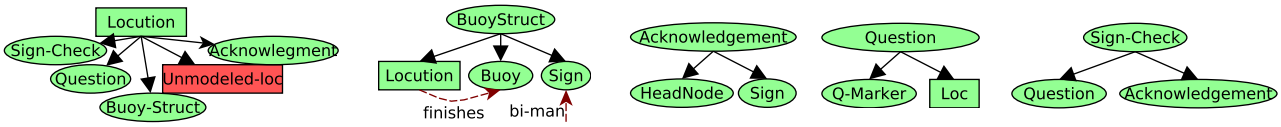
To come back to the example, we can extract several kinds of invariants. One may hypothesize that the sign beginning a buoy structure must be *bi-manual* (figure 2b). Another may want to describe the temporal structure of the patterns (Buoy *finishes* BuoyStruct, in figure 2b). It could also be useful to express global constraints, for instance constraints between one unit and all its descendants. All these invariants should be expressible formally.

We represent temporal, spatial and articulatory invariants as constraints. The constraints restrain the possible values for the attributes of pattern instances. The attributes, their encoding, and the logic formalisms – used to express the constraints – are a whole. Their choice strongly impacts the model. This is the reason why the formalism has to be independent of the logics and attributes.

Representing a complete model requires multiple logics, each addressing a different aspect: temporal, spatial, articulatory, etc. We showed examples of the temporal (*finishes*) and articulatory (*bi-manual*) aspects. In this article,

(a) VISIT (b) MUSEUM (c) L (d) FORGET (e) pointer (f) LOUVRE (g) LOUVRE

Figure 1: Decomposition of the excerpt



(a) Locution (b) Buoy (c) Ack (d) Question (e) "Sign Check"

Figure 2: Example of model with 4 patterns (b, c, d & e)

we focus on the formalism to describe the model. For this demonstration, only temporal constraints are used.

2.2. Edges of the models

Developing a complete model is, at best, very hard. We consider two solutions to work with incomplete models. As this work is developed for semi-automatic annotation, the first solution is to transfer the charge to the human operator. Such a system would ask something like "There might be a 'Question' there, is there an 'unmodeled-loc'? and which are its characteristics (attributes)?". This solution requires from the operator precisely what makes annotation difficult for humans: he/she is supposed to fulfill many attributes that are hard to measure for a human being. This problem leads to the second solution: coarse-grained models. Such models are not meant for the analysis of their results, they intend to produce a block with attribute values similar to what could have produced a complete model. Our solution combines these two approaches.
When a model is incomplete, edge nodes appear which are used but not modeled. Such an edge is present in the example model as "unmodeled-loc". The "unmodeled-loc" represents locutions built using non-modeled structures. We have built an experimental coarse model based on the sequence of lexical signs (because the annotation was already existing). The results, as expected, are not good. Depending on how constrained we make the model, we have far too much false-negatives or false-positives. The sequence model does not work well with the overlapping units: it includes units we don't want included and vice-versa. We expect dependency-based models to constitute better coarse-grained models.

2.3. Dependency structures

For the dependency grammar part, we present the formalism with a model which makes several simplistic hypotheses. The example model divides the units in two types: Manual Gestures (MGs) and NMGs, each one with its proper behavior. The units can represent a variety of forms: standard signs, other MGs (e.g. pointing MGs), facial gestures (e.g. qualifiers, quantifiers, modality markers), gaze gestures (e.g. references), etc. In SLs, articulatory constraints impact the syntactic level. Some units interact and some others are incompatible. In this example, the model emulates simplified articulatory interactions between its units:

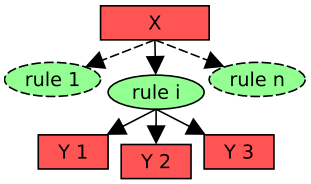


Figure 3: Representation of a dependency

- MGs never overlap. This is a simplification as it excludes the representation of yet described phenomenons (e.g. buoy structures, Cuxac's situational-transfers (Cuxac, 2000)).
- All NMGs can overlap. This is a simplification as some NMGs are articulatory impossible to produce simultaneously.

These simplifications allowed us to work with a slightly extended version of the Hays' formalism. Hays defines rules of the form $X(Y_{-n}, ..., Y_{-1}, *, Y_1, ..., Y_m)$ where X and Y_k are categories of units. Such a rule expresses that a unit of category X takes the place of the star in a sequence of dependents of categories Y_{-n} to Y_m. This formalism is sufficient to represent MGs (assuming the sequence simplification). But the NMGs requires to extend it, which is done with rules of the form $X(Y)$.
We have represented such dependency structures with the formalism with the construction shown in figure 3. The categories are described as alternatives between rules. The rules are described as patterns. The constraints work exactly as for constituency-based structures.

3. Parsing

The purpose of this work is the semi-automatic annotation of structures of models. The first step toward this objective was to formalize the model to recognize. The next step is the recognition itself. We give here an outline only of

the developed system. The detailed description will be the subject of a dedicated article.

In addition to the formalized model, the parser needs an input to parse. This input is an annotation of a subset of the units of the model. Units of this subset (they can be either pattern or alternatives) are said to be detectable. Their annotation can originate from manual annotation or third party detectors. These detectable units appear in red in figures 4 and 3. The parser is able to command the external detectors as it runs. In this mode, it does not receive the input annotation a priori, but works interactively with the detectors. This allows to inform the detectors of the context and therefore to reduce their search spaces. On the example, the parser asks to the "Buoy-Marker" detector "is there something between 201 and 212?". This allows to reduce the time interval the detector will process.

The internal representation of the model in the parser is an AND/OR graph. This representation is called the implicit graph. Our work extends the ideas of Mahanti (Mahanti et al., 2003) for the parsing. A unit identifying a pattern gives an *AND* node and one identifying an alternative gives an *OR* node. In the implicit graph, nodes represent patterns or alternatives but not instances. Figure 4 gives an example of an implicit graph for the example model. The implicit graph is used to generate an explicit graph. In this last graph, nodes represent instances.

The parsing operation results in a set of graphs. Each graph is a solution. The figure 5 shows an example of graph output by the parser. The nodes represent occurrences either externally detected or internally inferred. The arcs correspond to constituency or dependency relations of the model.

In a solution graph, each node has attributes. As the model can be under-constrained, there may be more than one solution. In particular, the resolution can find more than one acceptable value for attributes.

The parser is currently top-down. It builds the solution graphs starting from a set of given roots. This set can be, for example, a set of pre-detected lexical unit occurrences resulting of a first pass of lexical recognition. It is how the parser process dependency-based models. It then builds trees top-down from each root and merges the trees when possible. It is therefore obvious than solution graphs can have multiple connected components. This occurs, for example, when a signer is interrupted by a question, answers quickly and then continues his/her speech. In the case of constituency-based models, the top-down parsing requires to introduce a detectable root. It is the function of the "Signing" unit in figure 4 which is detected with an activity detector.

In the models we developed, the set of attributes contains *time-start* and *time-end*. Their values make it easy to transform a solution graph into an annotation.

4. Results

The parser has been evaluated for constituency-based and dependency-based structures: the first on real annotations, the second on synthetic data. The results of the parser can be directly observed, quantitatively and qualitatively. The evaluation of the formalism itself is harder to produce. We propose an interpretation of the parser's results to understand what they say about the formalism.

The parser has been run on several occurrences of the constituency-based structures. The external detectors were simulated with a manual annotation of the detectable units. But the small number of occurrences does not allow a quantitative evaluation. In particular, the evaluation corpus contains only one occurrence of a combination of the structures.

We still produce a qualitative analysis of the results. The parser outputs numerous solutions: many false-positives and partial solutions. A simple ranking by the size of the solutions is efficient against the partial solutions.

The false-positives can be classified in two categories: wrong hierarchical order and bad modeling of the lower levels of the syntax (discussed above, in section 2.2.). The first could be addressed with recursive constraints on the compositions. For example constraints like "the locution constituting a question cannot contain a question". Such a feature could be interesting for experiments on models. But in a context of semi-automatic annotation, we rather think that this type of false-positives must be resolved by a human expert. A system requiring this type of intervention of the operator is still of good help: it reduces the work in the task of selecting the right hierarchical organization. This uses the expertise of the operator for high-level problems.

The second type of false-positives comes from the difficulty we met in modeling the syntactic structures of low-level. It is the reason why we developed the dependency part of our formalism.

To evaluate the parser on dependency grammars, we have built a synthetic corpus. The idea behind this is to test the parser against bigger inputs. To generate this corpus, we used the model presented in the section 2.3.

Our generator starts with the random generation of dependency grammars. It then generates random phrases following the grammars. In the absence of measures on annotations, the models were parametrized arbitrarily. The corpus has 5000 grammars with 1 phrase each. All grammars have 20 categories. Every category has 3 to 4 rules each. Rules for non-manual categories have exactly one dependent. For manual categories, sizes have a uniform distribution on $[0, 4]$.

The results of the parsing on the synthetic corpus are visible in figure 6. The results are classed by phrase size. We have an average of 1 to 4 false-positives per phrase. It gives a precision of 52% to 5%. It is hard to draw conclusion from this result as it depends on the parameters chosen at the grammar generation. The recall of 83% to 23% is much more interesting. It validates the computability of the parsing.

5. Conclusion

The formalism of this article showed its ability to represent structures based on constituency as well as dependency relations. It has been done without assumptions on the sequentiality of lexical units nor on the predominance of the manual gestures. Instead, it uses constraints to describe invariants on the composition of the structures and on their temporal organization. We showed that these descriptions

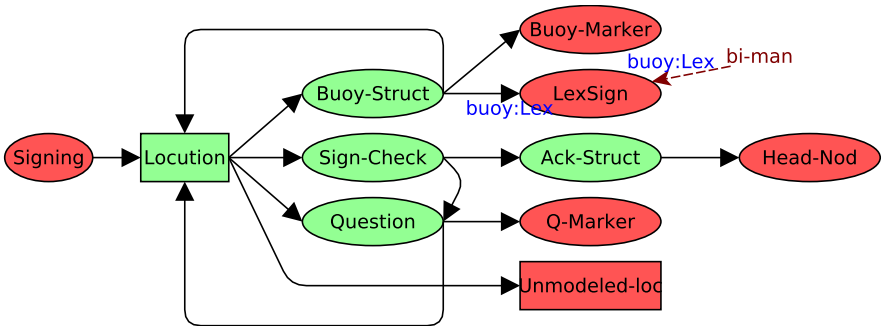


Figure 4: Schematic view of the implicit graph associated to the example model

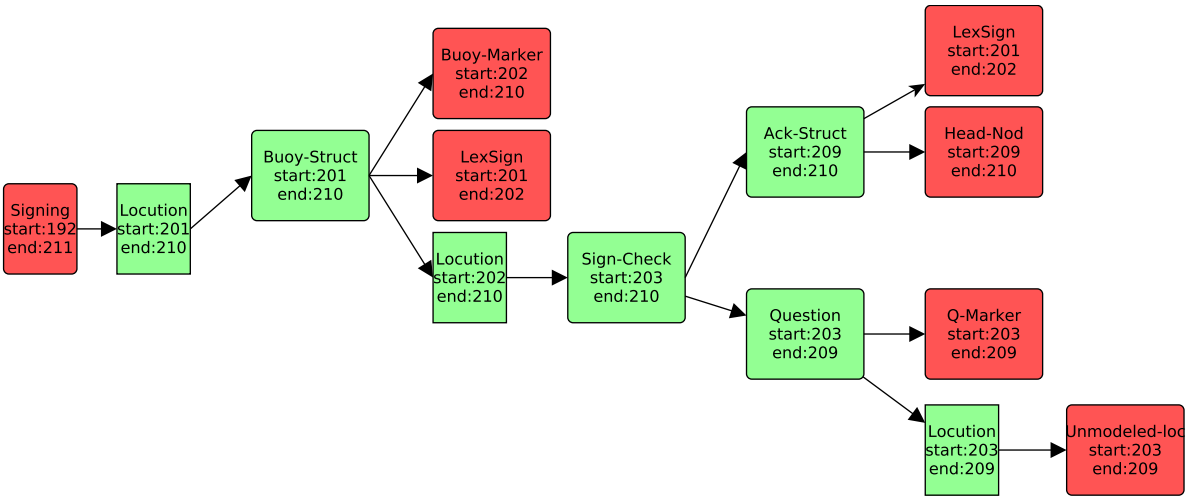


Figure 5: Example of solution graph

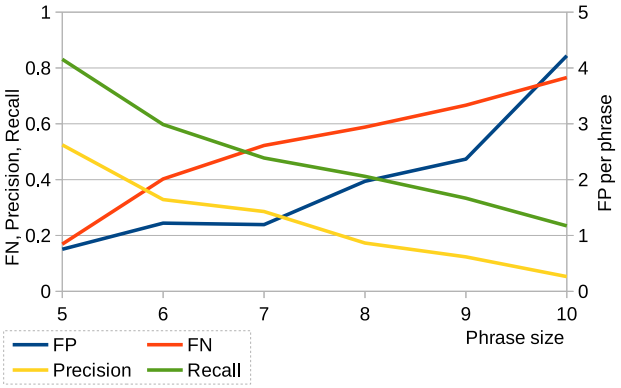


Figure 6: Evaluation on dependency grammars

allow the detection of the structures. The dependency parsing shows promising results as a coarse model. This should ease the use of constituency-based structures by disassociating them from the complete model requirement. However, the articulation between the two paradigms in one model remains to be developed. For now, the solution is to have two separated models, one per paradigm. The dependency-based model is used when a non-modeled pattern is reached. At this time, the human operator decides if the pattern is present and what solution of the dependency parsing will act as the occurrence of the non-modeled pattern.

This work, in its current state, is restricted by some limitations of the generative grammars. But it already avoids the problem of designing a model with a unique root for dependency grammars. This is critical in our context of semi-automatic annotation, as our goal is to enable the detection of structure occurrences, not to produce an interpretable syntactic tree. Unfortunately, the parser is still top-down, and consequently, the constituency-based grammars still need a root. There are plans to modify the current parser to drop the top-down mechanism. This will enable the parser to accept non-rooted models.

To go further in the direction of automatic annotation, several points need to be worked on. First, one will have to build (manually or automatically) a dependency grammar compliant with a real SL. The formalism and the parser can manage models of dependency grammars much more complex than one presented above.

The formalism and the parser do not represent uncertainty. But there are good candidates to introduce uncertainty representation in the existing parser such as fuzzy-CSPs. This extension will certainly improve greatly the results but will also have a computational cost.

6. References

Denis Bouchard and Colette Dubuisson. 1995. Grammar, order & position of wh-signs in quebec sign language. *Sign Language Studies*, 87(1):99–139.

Denis Bouchard. 1996. Sign languages & language universals: The status of order & position in grammar. *Sign Language Studies*, 91(1):101–160.

Noam Chomsky. 1957. *Syntactic structures*. Mouton&Co, La Haye.

Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231.

Arturo Curiel and Christophe Collet. 2013. Sign language lexical recognition with propositional dynamic logic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, page 328–333.

Christian Cuxac. 2000. *La langue des signes française (LSF): les voies de l'iconocité*. Ophrys.

Colette Dubuisson, Lynda Lelièvre, and Christopher Miller. 1999. *Grammaire descriptive de la LSQ*. Université du Québec à Montréal.

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. 2010. DICTA-SIGN: sign language recognition, generation and modelling with application in deaf communication. *International workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (LREC), Valleta, Malte*, pages 80–83.

Michael Filhol. 2009. A descriptive model of signs for sign language processing. *Sign Language & Linguistics*, 12(1):93–100.

Michael Filhol. 2012. Combining two synchronisation methods in a linguistic model to describe sign language. In Eleni Efthimiou, Georgios Kouroupetroglou, and Stavroula-Evita Fotinea, editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, number 7206 in Lecture Notes in Computer Science, pages 194–203. Springer Berlin Heidelberg, January.

Matilde Gonzalez and Christophe Collet. 2011. Signs segmentation using dynamics and hand configuration for semi-automatic annotation of sign language corpora. *Gesture in Embodied Communication and Humain-Computer Interaction*, pages 100–103, May.

Scott K. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press, March.

Ambuj Mahanti, Supriyo Ghose, and Samir K. Sadhukhan. 2003. A framework for searching AND/OR graphs with cycles. *arXiv preprint cs/0305001*.

Carol Neidle, Joan Nash, Nicholas Michael, and Dimitris Metaxas. 2009. A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus. In *Proceedings of the Language and Logic Workshop, Formal Approaches to Sign Languages, European Summer School in Logic, Language, and Information (ESSLLI 2009)*, Bordeaux, France.

M. P. Paulraj, Sazali Yaacob, Hazry Desa, C. R. Hema, and Wan Ab Majid. 2008. Extraction of head and hand gesture features for recognition of sign language. In *Proc. International Conference on Electronic Design ICED 2008*, pages 1–6.

Myriam Vermeerbergen, Lorraine Leeson, and Onno Alex Crasborn. 2007. *Simultaneity in Signed Languages: Form and Function*. John Benjamins Publishing, January.

Hee-Deok Yang and Seong-Whan Lee. 2011. Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. In *2011 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 4, pages 1726–1731.

Non-manual features: the right to indifference

Michael Filhol, Mohamed Nassime Hadjadj, Annick Choisier

LIMSI–CNRS
B.P. 133, 91403 Orsay cedex, France
Email: {michael.filhol,hadjadj,annick.choisier}@limsi.fr

Abstract

This paper discusses the way Sign Language can be described with a global account of the visual channel, not separating manual articulators in any way. In a first section section it shows that non-manuals are often either ignored in favour of manual focus, or included but given roles that are mostly different from the mainly hand-assigned lexical role. A second section describes the AZee model as a tool to describe Sign Language productions without assuming any separation, neither between articulators nor between grammatical roles. We conclude by giving a full AZee description for one of the several examples populating the paper.

Keywords: Sign Language modelling, non-manual features, synchronisation, AZee

1. Unjustified hand focus

1.1. Why ignore articulators?

Pretty much since the beginning of its description, whether naïve or scientific, SL has been "a way to speak with the hands". Initiated with Bébian (1825), established by Stokoe (1960) and completed by Battison (1978), the idea of formal Sign phonology through the description of manual parameters is still the most widely accepted way of describing signs. The number of technical projects involving SL, whether for its synthesis with signing avatars or its recognition with all sorts of devices (video tracking, Kinect, gloves, ring/bracelet sensors), unquestionably regard manual activity as the centre of all signed productions and the key to any underlying structure.

The parametric model eventually integrated an additional "facial expression" parameter, justified even by minimally contrasting lexical pairs such as "skin" vs. "racist" in LSF. But one must admit that it is usually discarded from lexicon descriptions, and occurrences of facial expression change over a lexical unit is often labelled grammatical or prosodic, i.e. almost off the limits of linguistic description, while a manual change in location, orientation or movement will likely be syntactically analysed.

For example, we have recently published the result of an LSF corpus study on event precedence and duration (Filhol, 2013). In this study, all sequences of two events separated by a period longer than 10 days involved the form photographed in fig. 1 and described as (r3) in the cited paper. It is close to the sign glossed "until now" in LSF picture dictionaries (fig. 2), at least in meaning but enough in form also to be used in annotation tasks. However, all occurrences in our study differ in the same way to what the drawing in fig. 2 suggests: the movement is of the linear type (not accelerated/ballistic), the fingers wiggle, the head rotates to the active hand's side, the eyes blink just before the gaze turns to the active side, the torso leans to the opposite side, and the synchronisation of all these features is consistent, etc. Why does no lexical description include those features on the same level as the manual gesture? Parametric descriptions do not even allow the torso tilt, but what makes the feature less lexical then the manual part, whereas we observe the former on every occurrence of the latter?

Figure 1: Snapshot of a form used for periods lasting over 10 days

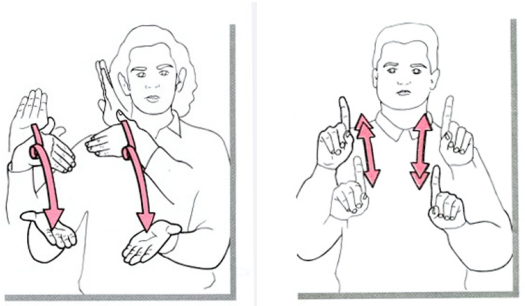


Figure 2: LSF picture dictionary images for "until now" (left) and "party"/"national day" (right)

Sometimes observable movements can indeed be side effects of other relevant body gestures, but this does not generalise to, say, the eye blink or the head rotation in our example. Conversely, parametric descriptions give a hand orientation in the LSF sign for "party"/"national day" like

palms face to face or away from body, whereas all variations occur, the only constraints invariably observed being the articulatory limits at the wrist. Also, should the coarse "facial expression" parameter be used to account for a facial detail such as an eye blink?
With no intention of denying the obvious articulation of hands in signed discourse, we do think the imbalance in focus between their part and that of other gestures should be questioned, manual preference not being justified and leading to flawed observations.

1.2. Why assign roles to articulators?

Not all published work fully discards non-manuals. Different studies exist, and a lot of them conclude assigning syntactic roles to observed articulators: manual placements on the left-right axis for absolute time anchors vs. on the sagittal axis for relative, eye gaze to switch between frozen and depiction mode (Cuxac, 2000), eyebrows combined with head tilts serve as interrogative markers (Baker-Shenk, 1985; Hickok et al., 1996), shoulder line rotation for reported turn taking (Moody et al., 1986), wide eye opening for the adverbial function of quantity (Vergé, 2001), etc.
In our study cited further up, all expressions of event duration, if exceeding 10 days, were found (r4) to involve the same form as fig. 1, including all articulations described but with an additional non-manual feature of semi-closed eyelids, which we note "el:semi-cl". Using the cited rule numbering system where (r2) is the chronological sequence, it is that only feature that differentiates the signed sequence (*event1*, fig. 1 with *duration*, *event2*) between the two meanings below:

- *event1* and *event2* are separated by the given *duration*: r2(*event1*, r3(*duration*), *event2*);
- *event1* is followed by *event2* lasting the given *duration*: r2(*event1*, r4(*duration*, *event2*)).

When the time period is under 10 days, the differences between event separation time (r1) and event duration time (r5) are:

- a change in manual activity—rule (r1) making use of a dictionary sign glossed "immediately followed by, your turn, consequence" whereas (r5) uses one glossed "duration, to last";
- (r1) uses eye gaze whereas (r5) does not;
- (r1) brings the chin forward whereas (r5) brings it up a little;
- (r5) uses the el:semi-cl feature whereas (r1) does not.

The change in manual activity in the case of shorter periods will unquestionably lead the traditional approach to call a lexical difference, optionally commenting on the non-manual features. But the case of longer periods less trivially allows overlooking the non-manual feature. Is the whole form (fig. 1+el:semi-cl) a different lexical item to fig. 1 alone, and to be glossed something like "during"? Or should we assign el:semi-cl the grammatical role of, say, denoting simultaneity of the period and the event to be signed afterwards? Indeed both forms of duration vs. separation use the same feature. Then what about the other non-manuals involved in the latter?

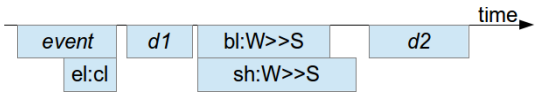


Figure 3: Form diagram for periods between two time boundaries, excl. enunciation time

Figure 4: Snapshot of a form used for periods between two time boundaries, excl. enunciation time

What is more, we have furthered our study since the cited publication, analysing the case of event durations defined by two time boundaries, as would be in English with the expression "from 1905 to 1914". We have found the two additional properties below:

- if the duration is disconnected from the present (enunciation) time and however long it lasts, the form is invariably that sketched in figure 3 (see picture in fig. 4), where:
  - *event*, *d1* and *d2* are the arguments denoting the event and the start and end dates of the period, respectively;
  - "el:cl" stands for an eye blink;
  - "sh:W>>S" is the 'J'-shaped strong hand lateral movement from weak to strong side;
  - "bl:W>>S" is the body movement leaning from weak to strong side, simultaneous to that of the strong hand;
- if the starting boundary is the enunciation time, e.g. "until Tuesday", the form used is that of figure 5 (see picture in fig. 6), where:
  - *event* and *until* are the two arguments, one being placed in either of two time positions;
  - "eg:s-sp" stands for an eye gaze directed to the signing space where the hands are placed;
  - "sh" and "wh" respectively stand for the strong and weak hands;
  - "hd:rot-dwn" is a small head rotation bringing the chin down.

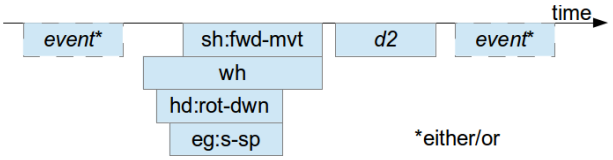


Figure 5: Form diagram for periods between enunciation time and a given time boundary

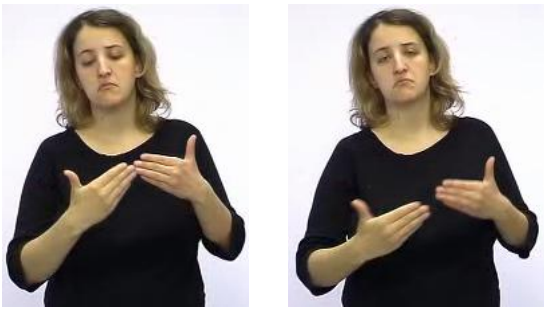


Figure 6: Snapshot of a form used for periods between enunciation time and a given time boundary

The form of fig. 4 is generally glossed "until", following LSF dictionaries' figure 7. Though, no occurrence actually carried the pictured form: all hand movements were performed sidewards, and every occurrence had a body movement along the manual one[1]. Typically, the change of movement is explained with some form of agreement in location, but in almost no case here could we really come to accept any relevant start and end signing space points for the movements. Besides, to our knowledge, no notice was ever taken of the body movement in such case. It might be argued to be the result of a co-articulation or a phonological process ruling it out as unintentional, but we can only admit that generally speaking very few lateral movements of the hand enrol the body in this way. Incidentally, we note that none of the last two forms use the el:semi-cl feature...

Figure 7: LSF picture dictionary image for "until"

[1]A few examples conforming to fig. 7 were found, but all of them were followed by path end points, not time boundaries. We find that result itself incidentally interesting: are both forms different lexical entries?

What we seem to observe is a propensity to explain manual variations as syntactically driven modification to lexical units, and non-manual additions as non-linguistic, optional or pragmatic. Yet looking at corpus videos with the global approach defended in the previous section, it appears likely that a number of articulators participate in most grammatical functions jointly, and there are non-manual features found inseperable from dictionary units. So, neither denying the existence of lexical units in SL nor that hands play a part therein, we do think there is an unjustified tendency to partition the body into different grammatical roles, the predominant assumption and most deeply rooted idea being a lexical track assigned to the hands.

Beyond the manual channel is of course the non-manual channel, now fairly known to participate in the signed message. Now the point of this paper is to propose that beyond the "manual + non-manual" view, there is yet a visual channel as a whole, with no separation between articulators. Non-manual articulators are articulators in the full sense and should probably not all be grouped under a label and defined by what they are not (manual). Just like ignored minorities claim a right to difference and, once visibility is earned, claim the right to indifference to feel fully included in the system.

2. AZee

AZee is a model to describe and synchronise articulatory forms, built with the philosophy above to synthesise signed productions with a virtual signer, or signing avatar. It comes in the wake of Zebedee, a model proposed a few years prior to this work. Initially made for lexical description, Zebedee:

- allowed writing reusable lexical forms including the invariant forms and the contextual dependencies;
- was based on a synchronisation scheme inspired by Liddell & Johnson's description system of posture–transition alternation (Johnson and Liddell, 2011), developed along the horizontal axis in figure 8;
- made exclusive use of necessary and sufficient articulatory constraints, i.e. no Stokoe-like parametric value was mandatory, only the required articulations were to be specified—vertical axis in the figure.

Especially with the last property above, Zebedee did away with fixed parameters and allowed a flexible articulatory description. However, we have seen that when studying all articulators and all grammatical functions, many features do not perfectly align in body postures but consistently precede or follow, say, a manual movement. Zebedee remained limited in that respect because its focus was still on lexical description, therefore on stabilised, hence mostly time-aligned movements.

To address this problem and gain more expressive power in articulator synchronisation and non-lexical description, the AZee extension was proposed, to:

- enable generic functional rules (whether or not lexical) and their associated forms, including invariant and context-dependent specification;

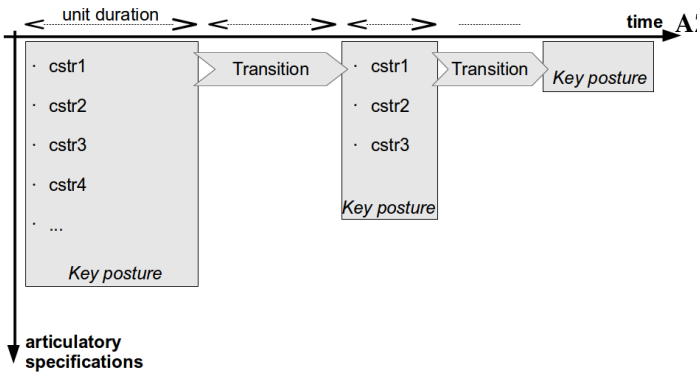


Figure 8: Zebedee

- specify any articulation (whether or not manual) at any time relative to another, for general specification on the time axis.

The basic instruments of the model are a set of native types and a set of typed operators and constants to build expressions normally resulting in XML specifications of animations to synthesise with a software avatar engine. The full set of types is : NUM, BOOL, VECT, POINT, LIST, SIDE, BONE, CSTR, SCORE, AZOP. All are described below.

NUM Numerical values, such as 6.2 or -9.

BOOL Truth values, either TRUE or FALSE.

POINT Points of the signing space. As continued from Zebedee, the signing space in AZee is regarded as a geometric Euclidean space, in which geometric objects can be built as needed and body articulators constrained with.

VECT Vectors of the signing space.

LIST Lists of AZee expressions.

SIDE Left vs. right.

BONE Articulators animated by joint rotations, e.g. the left forearm or the head.

CSTR Constraints that may apply at a point in time, of three main types: bone orientation and placement (forward/inverse kinematics), morphs (for non-skeleton articulators like facial muscles), and eyegaze direction.

SCORE Animation specifications, normally the result of an expression to be used as synthesis input. The only type to cover time, CSTR being articulatory but instantaneous. An XML description excerpt is given in figure 9. It basically specifies a list time-stamped keyframes in a first section, and a list of articulations and morph values to be reached at given keyframes, or held between given keyframes. The basic idea is that any articulator not given a morph value or a joint rotation may be interpolated to reach its next state, or simply take a rest or comfort-selected position.

AZOP Equivalent to functions in functional programming languages. They are to be applied to named argument expressions and result in new expressions. They are most useful to write production rules with non-frozen signed output. For instance, while a shoulder shrug gesture or some non-modifiable sign may be frozen thus described as a SCORE directly, most grammatical rules will be AZOPs with named arguments—such as *duration* in most rules discussed in this paper—and a SCORE output, whose expression depends on the arguments.

```
<Score>
  <KeyFrames>
    <KeyFrame id="1000" time="0"/>
    <KeyFrame id="1001" time="0.0001"/>
    <KeyFrame id="1002" time="1.0001"/>
    <KeyFrame id="1003" time="3.0000999999999998"/>
  </KeyFrames>
  <ScoreSpec>
    <Hold start="1000" end="1001">
      <Rots>
        <Rot joint="left_clavicle" x="-0.198961" y="-
        <Rot joint="right_clavicle" x="0.0075564" y="
        <Rot joint="left_shoulder" x="-0.216117" y="-
        <Rot joint="right_elbow" x="-0.0310633" y="0.
        <Rot joint="right_shoulder" x="0.540892" y="(
        <Rot joint="left_elbow" x="0.00735438" y="-0.
      </Rots>
    </Hold>
    <Hold start="1002" end="1003">
      <Rots>
        <Rot joint="left_clavicle" x="-0.373514" y="-
        <Rot joint="right_clavicle" x="-0.132402" y="
        <Rot joint="left_shoulder" x="0.192286" y="-(
        <Rot joint="right_elbow" x="-0.928085" y="0.1
        <Rot joint="right_shoulder" x="0.227843" y="(
        <Rot joint="left_elbow" x="0.772393" y="-0.2(
      </Rots>
    </Hold>
  </ScoreSpec>
</Score>
```

Figure 9: An AZee output of type SCORE

Here is a selection of AZee operators of various argument and result types, which should give an idea of a few things possible with AZee.

`plus`: numerical sum
Type: NUM, NUM → NUM

`scalevect`: vector scaling
Type: NUM, VECT → VECT

`orient`: orientation constraint
Type: str, BONE, str, VECT → CSTR
Articulatory constraint to orient skeleton bones in the signing space. The first argument is either 'DIR' or 'NRM' depending on whether the bone axis to be oriented is the direction bone (to make it point in a direction) or the normal bone (to lie it in a plane). The second is usually 'along' to align the vector in the given vector direction, but '//' is possible to allow opposite direction.

`place`: placement constraint
Type: site, POINT → CSTR
Articulatory constraint placing a body site at a point in space. The first parameter is a POINT expression, but is not evaluated to 3d coordinates of the point. It must be a body site expression, i.e. one referring to a point on the skin, to be placed at the point given by the second parameter.

`morph`: morph constraint
Type: str, NUM → CSTR
Articulatory constraint to control non-skeletal articulators such as facial muscles. Morphs have ID names, and can be combined with weights. The first argument is the morph ID to be used; the second is its $[0, 1]$ weight.

`key`: hold constraints
Type: NUM, CSTR → SCORE
This operation creates the most basic score. A "key(D, C)" expression returns a score of duration D, made of two animation keyframes between which the enclosed constraint specs C will be held. D can be zero, and C can hold any set of constraints: morphs, orientation constraints, placement constraints...

`sync`: synchronise scores
Type: name, SCORE, list of (name, SCORE, synctype) → SCORE
This operator is the major addition to the Zebedee model, used to synchronise a list of scores. Each score has a name, referred to by the other scores to specify the way they synchronise with the others. A name can be any identifier string; a synctype is a string from the list below, followed by the appropriate boundaries or durations:

- 'start-at', 'end-at': score is untouched and merged starting or ending at a given time position;
- 'start/end', 'start/duration': added score is stretched or compressed to fit the specification;
- 'start/kfalign': score geometry is abandoned and keyframes are aligned with those of the current score...

`azop`: create an AZee operator
Type: list of (str, AZexpr), AZexpr → AZOP
The result is an azop that can be applied to a context of named argument expressions, which will produce a result typed according to the last AZexpr given. This last expression generally contains references to the argument names, as would any parametrised function in a programming language. Alternatively, the 'nodefault' string can be given if no default expression makes sense; the argument then becomes mandatory when applying the azop.

`apply`: apply an AZOP to a context
Type: AZOP, list of (str, AZexpr) → returned by azop
The first argument is the azop to be applied. An azop comes with a list of optional or mandatory named arguments, which together form a context for the azop. The return value and type are given by the azop specification. If the azop is a production rule, it will result in a SCORE.

For example, the expression below describes the azop that models the rule sketched in figure 5, with the event signed first. Indentation denotes a parameter under its operator.

```
 1. azop
 2.   'event'  % argument dependency
 3.   'nodefault'
 4.   'until'  % dependency with default
 5.   empty

 6.   sync     % synchronising 6 boxes
 7.     'WH'   %% weak hand box
 8.     key
 9.       1
10.       place
11.         site
12.           'L_KN1'
13.           w
14.           1
15.         [point expression]
16.       [more constraints: hand cfg...]

17.     'EVT'  %% event box
18.     ref
19.       'event'
20.     'end-at'
21.     'WH:0:-.3'

22.     'DATE' %% time boundary box
23.     ref
24.       'until'
25.     'start-at'
26.     'WH:-1:+.3'

27.     'HEAD' %% head drop box
28.     [describe head drop]
29.     'start/end'
30.     'WH:0:+.1'
31.     'WH:-1:-.4'

32.     'GAZE' %% eye gaze box
33.     look
34.       site
35.         'PA'
36.         w
37.     'start/end'
38.     'HEAD:0:+.1'
39.     'HEAD:-1:0'

40.     'SH'   %% strong hand box
41.     [strong hand movement]
42.     'start/end'
43.     'WH:0:+.3'
44.     'WH:-1:0'
```

Lines 2–5 are declarations of the azop's arguments or contextual dependencies, including their names and default expression if absent on azop application, e.g. on l. 5 where 'until' is given a default empty score value. Lines 7, 17,

22, 27, 32 and 40 each names a part of the full signing activity, all to be synchronised by the `sync` operation. The word "box" here is a reference to the rectangles in the illustrations given in figures 3 and 5. Lines 20, 25, 29, 37 and 42 are sync types, i.e. specify the way in which the containing box is to be synchronised with the previous ones. All '*box*:*kf*:*off*' formatted strings are relative time specifications, creating a new keyframe for insertion if none is present at the specified time stamp. In such string, *kf* is the keyframe number of the identified *box*, from which to *off*set the time stamp. The same way values are indexed in Python lists, keyframe numbers are numbered 0 and up from the first to the last, and -1 and down from the last to the first. Line 39 refers to the final keyframe of the score contained in box `HEAD`; line 43 specifies a positive offset of .3 from the beginning of box `WH`.

This azop can be saved under the reference `"Event will last from now until"` and stored as a production rule capable of turning any (*my_event*, *my_date*) pair of scores into a resulting score, combining all boxed features and meaning that *my_event* will last from now until *my_date*. The expression for it is a simple application of the azop with both of its arguments set:

```
apply
  ref
    '"Event will last from now until"'
  'event'
  [my_event score here]
  'until'
  [my_date score here]
```

The interesting and new thing about this model is that the `sync` operation works with any set of scores and any contained articulation specification, except for anatomically impossible constraints. Nothing has enforced us to animate the hands, and no lexical base stream was needed for description. Evaluating this expression produces an XML specification of joint and morph articulations, as presented in figure 9, to be animated directly. Overall, this means we produce animations directly from semantically relevant rule entries and their contextual arguments.

3. Conclusion

This paper has discussed the fact that non-manual articulators were often either overlooked or segregated from manual activity in signing. Firstly, we have not only proposed that they be considered along with manual articulators, but even that all articulators be equal candidates for carrying meaning in SL productions. Secondly, we have made a case against SL articulator role assignment (i.e. projecting grammatical or syntactic functions to specific articulators), and against the assumption that hands would exclusively carry the lexical role. We propose that instead, all articulators be considered together at every moment, and we have shown that with this approach, articulators often seem to behave jointly for the linguistic functions that surfaced.

Then, to describe the observed signed activity with this recommended philosophy, we have presented the AZee model, extension of its ancestor Zebedee. It is capable of describing SL production rules as well as SL productions. That is, by parametrising description elements, AZee can describe generic and context-sensitive rules associating the signed forms to an established SL function, be it lexical or virtually anything else.

One purpose of AZee is to provide a grammar description model covering all SL features, but the aim of our work is ultimately to synthesise the formal descriptions it enables with virtual signers. The first prototype was built and presented recently (Braffort et al., 2013) through a website interface, and will be improved as we go along searching for new production rules.

4. Acknowledgement

We wish to thank WebSourd® for allowing us to study and cite their video material in figures 4 and 6; see `www.websourd.org`.

5. References

C. Baker-Shenk. 1985. The facial behavior of deaf signers: evidence of a complex language. *Am Ann Deaf*, 130(4):297–304.

R. M. Battison. 1978. *Lexical borrowings in American Sign Language*. Linstok Press.

R.-A. A. Bébian. 1825. *Mimographie, ou essai décriture mimique, propre à régulariser le langage des sourds-muets*. L. Colas, Paris.

A. Braffort, M. Filhol, L. Bolot, M. Delorme, C. Verrecchia, and A. Choisier. 2013. Kazoo: A sign language generation platform based on production rules. In *Sign Language Translation and Avatar Technology (SLTAT)*.

C. Cuxac. 2000. *Langue des signes française, les voies de l'iconicité*. Ophrys.

M. Filhol. 2012. Combining two synchronisation methods in a linguistic model to describe sign language. *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication, Springer LNCS/LNAI*, 7206.

M. Filhol. 2013. A rule triggering system for automatic text-to-sign-translation. In *Sign Language translation and avatar technology (SLTAT)*, Chicago, USA.

G. Hickok, U. Bellugi, and E. S. Klima. 1996. The neurobiology of sign language and its implications for the neural basis of language. *Nature*, 381(6584):699–702.

M. Huenerfauth. 2006. *Generating American Sign Language classifier predicates for English-to-ASL machine translation*. Ph.D. thesis, University of Pennsylvania.

Robert E. Johnson and Scott K. Liddell. 2011. A segmental framework for representing signs phonetically. *Sign Language Studies*, 11(3).

B. Moody, D. Hof, and S. Dumartin. 1986. *La Langue des Signes (tome I) – Histoire et grammaire*. Ellipses (reedited: IVT 1997).

W. Stokoe. 1960. Sign language structure: an outline of the visual communication system of the american deaf studies. *Linguistics, occasional papers*, 8.

F. Vergé. 2001. *Le regard en langue des signes française*. Ph.D. thesis, Université Le Mirail, Toulouse, France.

When nonmanuals meet semantics and syntax: a practical guide for the segmentation of sign language discourse

Sílvia Gabarró-López, Laurence Meurant
FRS – FNRS and University of Namur
Correspondence to: 61, rue de Bruxelles, B-5000 Namur, Belgium
E-mail: silvia.gabarro@unamur.be, laurence.meurant@unamur.be

Abstract

This paper aims to contribute to the segmentation of sign language (SL) discourses by providing an operational synthesis of the criteria that signers use to segment a SL discourse. Such procedure was required when it came to analyse the role of buoys as discourse markers (DMs), which is part of a PhD on DMs in French Belgian SL (LSFB). All buoy markers found in the data had to be differentiated in terms of scope: some markers (like most list buoy markers) seemed to be long range markers, whereas others (like most fragment buoy markers) seemed to have a local scope only. Our practical guide results from a hierarchized and operationalized synthesis of the criteria, which explain the segmentation judgments of deaf (native and non-native) and hearing (non-native) signers of LSFB who were asked to segment a small-scale (1h) corpus. These criteria are a combination of non-manual, semantic and syntactic cues. Our contribution aims to be shared, tested on other SLs and hopefully improved to provide SL researchers who conduct discourse studies with some efficient and easy-to-use guidelines, and avoid them extensive (and time-consuming) annotation of the manual and non-manual cues that are related to the marking of boundaries in SLs.

Keywords: segmentation, discourse unit, head nod, eye blink, head movement, eye gaze, pause, sign hold, role shift, palm-up

1. Introduction

Several studies on different sign languages (SLs) have faced the necessary but tricky question of segmenting signed discourses (Crasborn, 2008; Ormel & Crasborn, 2012). When segmentation is tackled with the sentence as standard unit, the researcher faces the problems of the syntactic delimitation of predicates in SLs and the determination of the syntactic status of simultaneous constructions that are typical to SLs (Crasborn, 2008). Both problems are not solved to date. In a number of studies (Crasborn, 2007; Fenlon et al., 2007; Hansen & Heβmann, 2007; Herrmann, 2009; Hochgesang, 2009; Jantunen, 2007; Nicodemus, 2006; 2009), segmentation has been approached from a prosodic perspective, namely by considering that prosodic cues reflect the syntactic organisation to some extent. From these studies, we know that various manual (e.g. palm-up signs, sign holds) and non-manual cues (e.g. eye blinks, head nods) contribute to the marking of "intonational phrases" or, more generally, of "boundaries" (Fenlon, 2010) in SLs. None of these cues functions as dominant cue by itself; on the contrary, boundaries are frequently marked by a layering of several prosodic cues.

The emergence of large-scale SL corpora and the discourse studies they make possible imply a new (practical) perspective on SL discourse segmentation. In our case, the study of the role of buoys as discourse markers led us to compare the scope of the different buoys markers observed in our data. Some markers (like most list buoy markers) seemed to be long range markers, whereas others (like most fragment buoy markers) seemed to have a local scope only. We observed that such scope differences get a more enlightening interpretation when they are interpreted in terms of "discourse units" rather than in terms of number of signs. Nevertheless, our concern was how to delimit such discourse units in a consistent (and shared between researchers) way since we did not have any tool or guidelines, which allowed us to do so.

The purpose of this work is to solve the above mentioned lack of guidelines for discourse segmentation by extracting a synthesis of the criteria that seem to influence the segmentation of three deaf (two native and one non-native) and two hearing (non-native) LSFB signers. Such synthesis will be organized into a set of guidelines that describe a minimalist, hierarchical and operative set of criteria that allows the standardisation of discourse segmentation among researchers of different SLs, among different SL corpora and within the same SL corpus.

This contribution is divided into four parts: section 2 explains the methodology we used to carry out our study, section 3 gives an account of the quantitative results of this pilot study and tackles one specific cue (eye blinks layered with head nods), section 4 explains the principles that led us to the elaboration of the segmentation protocol and proposes a guideline composed of four steps in order to segment a SL discourse into units, and section 5 contains the summary and conclusions of our research.

2. Methodology

We used a one-hour corpus of one signer (Gabarró-López & Meurant, 2013) made up of two argumentative (A1 and A2), two explicative (E1 and E2), two metalinguistic (M1 and M2) and two narrative (N1 and N2) discourses. Each group was balanced in terms of time. We mixed spontaneous and prepared productions as well as monologues and dialogues, so that the sample contained very different data with the most possible speech contexts.

In order to practically define discourse units, we designed a two-stage process that we named “copy test” and “cut test”. The first stage (“copy test”) consisted in taking a three-minute sample of each genre and asking the three deaf people to repeat the content and the signs of the clips to an experimenter who did not see the video. To do so, they first watched the three minutes of one video and afterwards they had to watch it again and stop the video whenever they thought it would be convenient for them. They repeated each segment to the experimenter who was sitting beside them and who was in charge of coding their fragments in ELAN. This procedure was repeated for the other three videos. It aimed to bring the segmenters to cut the discourses into semantically coherent units. The second stage (“cut test”) consisted in cutting the whole corpus into discourse units. The instructions given to both hearing and deaf annotators were that they had to watch the video and segment whenever they thought it was possible to cut the discourse. Each video was segmented using ELAN by a minimum of two people and by a maximum of four. Moreover, among these four people, three participated in the “copy test” as well.

Once both tests were finished, the tier “Common_units” and “Common_cues” were created. The first aimed at showing the number of annotators who had segmented in a particular place in the “copy test”, whereas the second aimed at gathering all the boundaries where at least two segmenters had coincided in the “cut text” so that we could create the list of cues appearing at that particular boundary.

3. Results

For the sake of clarity, our results are divided into three subsections: the first and the second one contain quantitative data and are related to the “copy test” and the “cut test” respectively, and the third contains qualitative data that tackles the case of a particular cue, i.e. the eye blinks layered with head nods.

3.1 The “copy test”

In this subsection, we will present three different sets of data, which concern the “copy test”: (i) the inter-segmenter agreement, (ii) the frequency of appearance of manual and non-manual cues at common boundaries, and (iii) the distribution and weight of boundary cues.

3.1.1. Inter-segmenter agreement

For the “copy test”, we found a total of 190 boundaries spotted by the participants of the “copy test” being 95 at the beginning (b) of a segment and 95 at the end (e). Both letters (b and e) are followed by the number of segmenters who had agreed on a particular boundary. The following table shows a summary of the data.

Common beginnings or ends	Number of boundaries	Percentage
e1 (idem for b3)	55	57.89%
e3 (idem for b1)	31	32.63%
e2 (idem for b2)	9	9.47%
total (including b)	95 (190)	100%

Table 1: Annotator agreement on the (begin and end) boundaries in the “copy test”

These data show that one boundary is commonly noticed by three segmenters out of three, so a third of the boundaries are undeniable. Most of the boundaries (more than one out of two) were only spotted by an annotator (not always the same one), which means that beyond undeniable boundaries (32.63%) and shared boundaries (9.47%), there is a high number of possible boundaries that varies from one segmenter to the other. Such divergence may probably be related to the capacity of memorising.

Moreover, the comparison between the “copy test” and the “cut test” allows us to refine the analysis of the "e1" boundaries of the "copy test". Indeed, 60% of the "e1" boundaries (33 out of 55) correspond to a boundary, which was at least noticed by two segmenters in the "cut test". This refines the picture of the inter-segmenter agreement in the "copy test". These figures confirm that these boundaries had to be considered as coherent from a discourse perspective and linguistically founded.

3.1.2. Manual and non-manual cues at discourse units’ boundaries

Once the “copy test” had taken place in the twelve-minute sample of the corpus, we crossed the results of the boundaries that had been spotted by any of the three deaf segmenters with the manual and non-manual cues appearing at every boundary that we had coded in the “Common_cues” tier. Table 2 shows every cue that was boundary marking, the number of times that it occurred and the percentage that it represents. The sum of percentages is higher than 100% because the cues of the list are sometimes layered since one boundary is often marked by several combined cues.

The criteria highlighted in grey made up the top seven cues noticed for segmentation and their percentage of appearance is over 10%. Pauses are by far the cue, which coincides more often with the segmentation resulting from the “copy test” (64 occurrences, i.e. at 67% of the boundaries spotted). This is not surprising, since pauses are organised in a systematic way that reliably indicates intonational phrase boundaries (Fenlon, 2010). Our definition of pause for this work coincides with this author: they are periods of no signing at all that can be divided into weak pauses (hands still raised but relaxed) and strong pauses (hands are dropped to the signer’s lap or clasped together).

Eye blinks co-occurring with head nods seem to be the most common and recurrent non-manual boundary marker that segmenters look at with 38 occurrences (40%)[1]. Sign holds (the final handshape of a sign is held in final position for a longer duration) are also an easy-to-notice cue that comes right afterwards with 23 occurrences (24%). Changes in head position layered with eye gaze also appeared as cues in 19 boundaries (20%), whereas eye blinks occurred at 17 boundaries (18%).

Cue	Number of appearances	Percentage
Pause (1)[2]	64	67%
Eye blink layered with head nod (3)	38	40%
Sign hold (2)	23	24%
Change in head position layered with a change in eye gaze (4)	19	20%
Eye blink (8)	17	18%
Role shift (5)	14	15%
Palm-up (9)	11	12%
Head nod (10)	5	5%
Bracketing repetition (6)	4	4%
Head movement (11)	4	4%
Change in eyebrow position (13)	3	3%
Buoy (14)	3	3%
Rhetorical question (7)	2	2%
Change in eye gaze (12)	1	1%

Table 2: Frequency of appearance of the different cues at the 95 common boundaries of the "copy test"

Even if role shift is in the sixth position with 14 occurrences (15%), it is commonplace in narratives and very often the boundary of a discourse unit was found there. On the contrary, palm-ups could be found in all discourses (monologue and dialogue, prepared and spontaneous) but their presence at a boundary is not that common (11 occurrences, i.e. 12%).

3.1.3. Distribution and weight of boundary cues

The data in the previous sub-subsection illustrates the cues used in the "copy test" that coincide with a discourse unit boundary, regardless of whether it was one segmenter, two or the three of them who spotted that boundary. The aim here is to give an account of the cues noticed by the three segmenters at the same time, by the two and by only one.

The three segmenters (e3) coincided in 31 boundaries, 30 were featured by the pause and one was featured by a sign hold. Therefore, the pause is a key cue to mark discourse units' boundaries (not very surprising as we said in 3.1.2) and the sign hold may have the same effect (we have sometimes found cases of 5-seconds holds). In very few cases we had boundaries marked by only two segmenters (e2). Once again the pause was par excellence the most common cue appearing at 8 boundaries out of 9, whereas the role shift was present in the remaining one.

As regards the boundaries noticed by only one segmenter (e1), we observed that 33 boundaries out of a total of 55 in the "copy test" (i.e. 60%) are also boundaries spotted by at least two segmenters in the "cut test". The pause is still the dominant cue with 18 occurrences, whereas the role shift accounts for 12. In conclusion, 28 boundaries out of 33 contain one of these two cues, whereas the remaining 5 are a combination of cues (3+9, 8+4, 8+8+2, 8+4), which means that a blinking has always occurred.

3.2 The "cut test"

In this subsection, we will present two different sets of data that relate to the "cut test": the inter-segmenter agreement and the frequency of appearance of the different manual and non-manual cues. The "cut test" was conducted on a one-hour corpus (including the 12 minutes of the "copy test") and contains four different situations whose discourses were at least segmented by two people each.

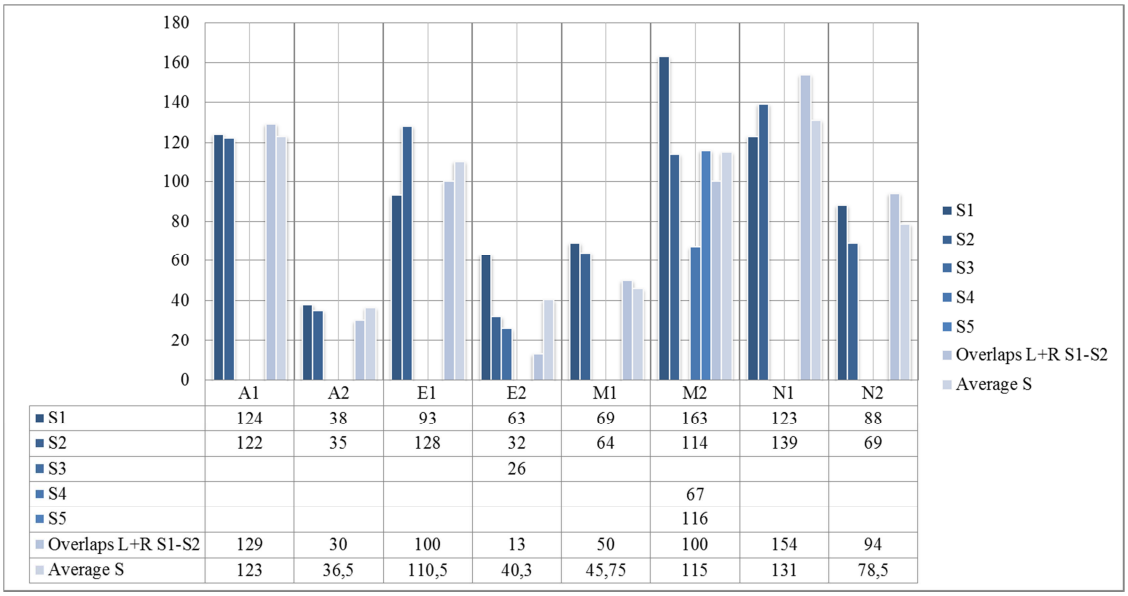
3.2.1. Inter-segmenter agreement

To begin with, we can see that the number of segments in a particular video varies sometimes greatly from one discourse to another due to the different length of each video and to the different situation in which the signer is found, i.e. monologue and dialogue. The agreement between segmenters tends to be high, at least between two segmenters participating in the same annotation file. Figure 1 illustrates the number of segments per segmenter (S) in each discourse, the right and left overlaps taking place within S1 and S2 segments (Overlaps L+R) and the average number of segments resulting from S1 and S2 segmentations.

The segmentations of six discourses out of a total of eight show a high degree of similarity from one segmenter to another. Numbers are very similar in A1, A2 and M1. E1, N1 and N2 show a slightly lower rate of agreement compared to the previous discourses but which, in any case, remains high.

[1] This statement about the semantically-guided boundaries the segmenters spotted is in line with the conlcusions of Herrmann (2010) about the consistency and the frequency of use of eye blinks to mark prosodic boundaries.

[2] The numbers following each cue are the codes that we used in order to annotate them in the "Common_cues" tier. This list of codes and cues comes from our first hypotheses on what cues (be them manual, non-manual, phonologic, syntactic or semantic) seemed to have more influence to spot a boundary.



	A1	A2	E1	E2	M1	M2	N1	N2
S1	124	38	93	63	69	163	123	88
S2	122	35	128	32	64	114	139	69
S3				26				
S4						67		
S5						116		
Overlaps L+R S1-S2	129	30	100	13	50	100	154	94
Average S	123	36,5	110,5	40,3	45,75	115	131	78,5

Figure 1: Inter-segmenter comparison in the "cut test"

E2 and M2 were segmented by more than two people. In M2, we can see that the numbers vary but at least two segmenters, S2 and S5, have very similar figures, and they are very close in the overlaps' sum as well as in the average. In E2, the results we got are weaker than those for the other videos because S1 segments almost double the number of S2 segments. S2 segments embrace most of S1 segments (there are 53 surrounding). Nevertheless, if we compare the segmentation performed by S2 with that of S3, we get more consistent results since S3 had 26 segments with 17 left overlap and 16 right overlap (i.e. 33 L+R) and the average of segments is 29.

3.2.2. Manual and non-manual cues at common boundaries

The "cut test" gave as a result 591 segments where at least two segmenters had coincided. Table 3 illustrates the name and the code of each cue, the number of appearances of each one and the percentage. As in Table 2, the sum of percentages is higher than 100% because the cues of the list are often layered in a single boundary.

If we compare this table with the previous one, we can see that results are not divergent. On the one hand, the same seven cues are found at the top of both lists with two almost anecdotal inversions: the change in head position layered with a change in eye gaze is in the third position and the sign hold in the fourth in the "cut test" list whereas it was the other way round for the "copy test", and the same happens with role shift which is now in the fifth position and the eye blinks in the eighth for the "cut test". On the other hand, the percentages are similar from one experience to the other, which means that regardless of the instruction that is given and how it is carried out, the same cues appear to be influent when it comes to segment the discourse into units.

In addition, Table 3 has a supplementary cue: the repetition of a sign (AA or AAA) that we only found in M2. We think that it is due to the nature of the video: it is a non-prepared dialogue on metalinguistic issues. Even if the number of boundaries where it occurs is not representative, the sample we took for the "copy test" does not include repetitions of a sign so we do not know whether a segmenter would have spontaneously marked a boundary there or not.

Cue	Number of appearances	Percentage
Pause (1)	304	51.4%
Eye blink layered with head nod (3)	266	45%
Change in head position layered with a change in eye gaze (4)	187	31.6%
Sign hold (2)	142	24%
Role shift (5)	137	23.2%
Eye blink (8)	81	13.7%
Palm-up (9)	77	13%
Head movement (11)	43	7.3%
Head nod (10)	27	4.6%
Change in eyebrow position (13)	21	3.6%
Bracketing repetition (6)	18	3%
Rhetorical question (7)	17	2.9%
Change in eye gaze (12)	13	2.2%
Buoy (14)	12	2%
Repetition of a sign (AA or AAA) (15)	2	0.3%

Table 3: Frequency of appearance of the different cues within 591 segments arising from the "cut test"

3.3 Eye blinks layered with head nods

Eye blinks layered with head nods (cue 3) is one of the most commonly spotted cues at discourse unit boundaries

being present in 45% of the boundaries. However, it is a special and sometimes tricky cue that deserves a specific subsection.

Unlike all the other six cues that could be found at the top of Table 2 and Table 3 (pause, change in head position layered with a change in eye gaze, sign hold, role shift, eye blink and palm-up), eye blinks combined with head nods can also act as linkers between two syntactic components; the first component is dependent on the second one, and thus do not correspond to a discourse unit. This means that while we can say that some other cues are conclusive to mark the end of a segment, we have to be careful with cue 3 (c3) because if we always segment there, we can lose the true syntactic construction of the discourse unit as well as its meaning. The three examples below illustrate this phenomenon of eye blinks combined with head nods.

(1) COMMUNICATION-SUPPORT-WORKERS
ce-up c3
SIGN-WRONG OUT PEOPLE SEE GOOD IT GOOD
"Even if communication support workers do not sign well, outside people see it and think they do well"

(2) DATE MEETING DATE CONFERENCE DATE
ce-up c3
TRAINING SEMINAR INTERPRETER THERE-IS-NOT NOT FIND PALM-UP REPLACE COMMUNICATION-SUPPORT-WORKER TAKE SAY NO
"When there is a meeting, a conference or a seminar and there is no interpreter there because none was found, and it is replaced by a communication support worker, say no!"

(3) YEAR UP-TO-NOW DEAF GROWING-GROUP
c3
COLLEAGUES STRUGGLE WANT INTERPRETER HIGH-LEVEL
"For years now, we (a growing group of deaf colleagues) have struggled to get high-level interpreters"

In the first two examples, the eye blink layered with a head nod that occurs in the middle of the utterances is the link between the two parts of a temporal syntactic structure, so no segmentation must be made there. Nevertheless, these cases where c3 is not a boundary can be easily isolated because (i) they come close after a boundary, (ii) there is no other associated cue, and (iii) the chin and the eyebrows go up (ce-up) in the first part of the segment before the eye blink layered with a head nod takes place.

The third example is different from the other three in articulatory and semantic terms. Here c3 marks the end of a kind of parenthetical comment that makes explicit the agent of the utterance, i.e. "*we (a growing group of deaf colleagues)*". Once again, the two first criteria that we mentioned above (cue 3 is near a boundary and not combined with another cue) are valid to distinguish whether it is a discourse unit boundary or not.

Anyhow, when an eye blink layered with a head nod is not associated with other cues, the segmenter will have to verify the possible role of cue 3 as a syntactic linker, especially if such cue is close to a discourse unit boundary.

4. A proposal for SL discourse segmentation

As we said at the beginning of this contribution, our purpose is to create a set of guidelines, which allow the standardisation of discourse segmentation among researchers of different SLs, among different SL corpora and within the same SL corpus. The tool that we are proposing aims to facilitate inter and intra corpus/ora comparisons in the field of discourse analysis and thus to facilitate the elaboration of studies on the position of an element as regards segment boundaries and the development of automatic language processing tools, to name a few of its potential usages.

4.1 The principles of the guidelines

To conceive these guidelines for discourse segmentation, we decided to base our research on the spontaneous segmentation carried out by three deaf signers (two natives and one non-native) and two hearing non-native signers (see previous sections). Such procedure was systematized, the criteria taken into account for the segmentation was minimized and the criteria that could be easily spottable when watching a video were favoured, so priority was given to phonological criteria, i.e. to the "visible markers" (Fenlon, 2010). Since our goal was to avoid the time-consuming annotation of manuals and non-manuals as well as long lists of cues to look at, we limited as much as possible the number of elements to take into account for the segmentation. Last but not least, we wanted to propose a tool that avoided wrong segmentations or, in other words, we did not want to create a too powerful and rigid procedure that would allow the segmentation in the right places but also in the wrong ones.

4.2 The guidelines for discourse segmentation

The intuitions we had after the first segmentations of video were that we would compulsory need a combination of at least four cues – a pause, a sign hold, an eye blink layered with a head nod or a change in head position layered with a change in eye gaze – in order to segment without mistakes. Surprisingly, the results of both tests showed that we only need a set of two cues to process an almost complete segmentation that is consistent with the linguistic intuitions of the signers: cue 1 (pauses) and cue 3 (eye blinks layered with head nods). Then, three additional cues (5, 8 and 4) allow the segmentation to be refined.

To get optimal results from our segmentation protocol, the segmenter needs to watch the video thrice, if it is the first time that he is confronted with the discourse. The first

time he will only watch the video, the second he will segment it into discourse units (steps I, II and III) and the third he will verify that the segmentations are situated in the right places (step IV). However, the last two viewings will suffice if the segmenter already knows the video or has already worked with it.

The four steps for the segmentation into discourse units are the following:

I. As a general rule and for all kinds of discourse, segment at every pause (i.e. periods of no signing no matter whether hands are still raised, dropped to the signer's lap or clasped together) and at every sign hold.
II. For narrative discourses, which usually involve characters and dialogues, segment at the end of every constructed dialogue and role shift.
III. Segment systematically at every eye blink layered with a head nod (cue 3) (or at every combination of a blink (cue 8) in the close context of a change in eye gaze and head position (cue 4)).
IV. Remove all the eye blinks layered with head nods acting as discourse unit linkers (see for example the three criteria given in section 3.3).

The identification of manual and non-manual cues meets the tracking of semantic units of role taking and some syntactic relationships.

5. Summary and conclusions

This paper has contributed to the topic of segmentation of sign language (SL) discourses by creating a practical and easy-to-use layout for discourse segmentation which avoids time-consuming annotation of every nonmanual. To do so, we have tried to (i) understand the hierarchy of criteria that lead native and non-native signers towards the identification of segment within a discourse, and (ii) see how we could organise in an operational and minimalist way the intuitions of signers. Our objective was not to predict where a signer would segment spontaneously, but to standardize the segmentation among researchers working in the field of discourse analysis and providing them with a systematization of the linguistic intuitions of signers.

We designed two tests in order to elicit the spontaneous segmentations. The first one ("copy test") involved the three deaf segmenters, who were asked to watch four video samples of a one hour corpus and stop them whenever they found it necessary in order to repeat in detail the same signs with the same meaning to another researcher who was copying their segments into an ELAN file. The second one ("cut test") consisted in having the five segmenters (at least two per file) viewing and segmenting the whole corpus containing different discourses directly in ELAN.

The results show a high consistency between both tests. Seven manual and non-manual cues are the most commonly used by segmenters to spot segment boundaries: pauses, eye blinks layered with head nods, changes in head position layered with changes in eye gaze, sign holds, role shifts, eye blinks and palm-ups. The results also show a high rate of inter-segmenter agreement. We could then consider the spotted boundaries as coherent from a discourse perspective and linguistically founded. The comparison between the "copy test" and the "cut test" proved that more than half of the boundaries spotted in the "copy test" corresponded to a boundary which was at least noticed by two segmenters in the "cut test", which means that these boundaries had to be considered linguistically coherent.

We have also tackled the particular case of the eye blinks layered with head nods, which sometimes may have a linker role rather than a boundary marking cue role. As a final point, we have presented the principles, which guided us towards the creation of this segmentation protocol and the steps which compose it.

Since this is a pilot study, we are well aware of its shortcomings, the first one being that the videos are featured by one signer only, hence we want to test our protocol with a larger sample of the LSFB Corpus which contains a wider sample of discourses and different signers. Such testing would allow us to get even more solid results and would also prove whether these guidelines are suitable for segmenting the discourse of any signer. Finally, we would also like other SL researchers to test these guidelines with their data on other SLs and give us feedback on their experience and the possible issues to implement.

6. Acknowledgements

We would like to thank Christophe de Clerk, Susana Sánchez, Aurélie Sinte and Bruno Sonnemans for their collaboration as well as Gemma Barberà for her valuable comments. Our research is funded by a F.R.S-FNRS Research Fellow Grant and the F.R.S-FNRS Incentive Grant for Scientfic Research n° F.4505.12.

7. References

Crasborn, O. 2007. How to recognize a sentence when you see one. *Sign Language & Linguistics* 10: 103–111.

Crasborn, O. (2008) Open Access to Sign Language Corpora. *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, Crasborn, Hanke, Efthimiou, Zwitserlood & Thoutenhoofd, eds. ELDA, Paris. pp 33-38

Fenlon, J. 2010. *Seeing sentence boundaries: the production and perception of visual markers signalling boundaries in signed languages*. Doctoral thesis submitted at the University College London.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. 2007. Seeing sentence boundaries. *Sign Language and Linguistics* 10(2): 177–200.

Gabarró-López, S., and Meurant, L. 2013. The Use of Buoys Across Genres in French Belgian Sign Language. *Proceedings of COLDOC 2013: La question des genres à l'écrit et à l'oral,* November 13 – 14, 2013, Paris, France (in press).

Hansen, M., and Heβmann, J. 2007. Matching propositional content and formal makers. Sentence boundaries in DSG text. *Sign Language and Linguistics* 10(2).

Herrmann, A. 2009. Prosody in German Sign Language. *Workshop on Prosody and Meaning,* 17–18 September 2009, Frankfurt am Main, Germany.

Herrmann, A. 2010. The interaction of eye blinks and other prosodic cues in German Sign Language. *Sign Language and Linguistics* 13(1), 3-39.

Hochgesang, J. A. 2009. Is there a 'sentence' in ASL? Insight on segmenting signed language data. Talk presented at *Sign Language Corpora: Linguistic Issues Workshop*, 24 July, London, UK.

Jantunen, T. 2007. The equative sentence in Finnish Sign Language. *Sign Language and Linguistic* 10(2): 113–143.

Nicodemus, B. 2006. *Prosody and utterance boundaries in ASL interpretation*. Paper presented at the DGfS [Deutsche Gesellschaft für Sprachwissenschaft] 2006 workshop "How to recognize a sentence when you see one: methodological and linguistic issues in the creation of sign language corpora", 23–24 February, Bielefeld, Germany.

Nicodemus, B. 2009. *Prosodic markers and utterance boundaries in American Sign Language interpretation*. Gallaudet University Press, Washington D.C. 20002.

Ormel, E., and Crasborn, O. 2012. Prosodic correlates of sentences in signed languages: A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies* 12(2): 109-145

Last train to "Rebaudengo Fossano": the case of some names in avatar translation

Carlo Geraci[+], Alessandro Mazzei[*]
[+]Institut Jean-Nicod, CNRS, Rue d'Ulm 29, 75005 Paris
[*]Università degli Studi di Torino, Corso Svizzera 185, 10149 Torino, Italy
E-mail: carlo.geraci76@gmail.com mazzei@di.unito.it

Abstract

In this study, we present an unorthodox case study where cross-linguistic and cross modal information is provided by a "non-manual" channel during the process of automatic translation from spoken into sign language (SL) via virtual actors (avatars). Specifically, we blended written forms (crucially, not subtitles) into the sign stream in order to import the names of less-known train stations into Italian Sign Language (LIS). This written Italian-LIS blending is a more effective compromise for Deaf passengers than fully native solutions like fingerspelling or using the local less-known SL names. We report here on part of an ongoing project, LIS4ALL, aiming at producing a prototype avatar signing train station announcements. The final product will be exhibited at the train station of Torino Porta Nuova in Turin, Italy.

Keywords: Sign Language, Automatic translation, avatar

1. Background

Avatar technology is becoming more and more popular as a tool to implement automatic translation into sign language. Current projects investigate relatively small domains in which avatars may perform decently, like post office announcements (Cox et al., 2002), weather forecasting (Verlinden et al., 2002), the jurisprudence of prayer (Almasoud and Al-Khalifa, 2011), driver's license renewal (San-Segundo et al., 2012), and on train announcements (e.g. Braffort et al., 2010, Ebling and Glauert, 2013).

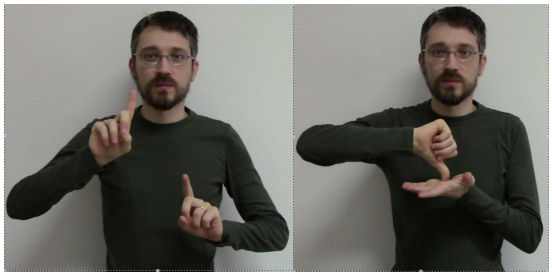
LIS4ALL is a project of automatic translation into LIS where we faced the domain of public transportation announcements. Specifically, we are developing a system of automatic translations of train station announcements from spoken Italian into LIS. The project is the prosecution of ATLAS, a project of automatic translation into LIS in the domain of weather forecasting (http://www.atlas.polito.it/index.php/en). We are using the same symbolic (rule-based) translation architecture to process the italian input and generate the final LIS string. In particular, we are enlarging the types of syntactic constructions that the avatar can translate and we are also enlarging the electronic lexicon built for ATLAS (around 1500 signs) by adding new signs specific of the train station domain. Indeed this latter was one of the most challenging aspect of the project especially once the domain of train stations is addressed. Prima facie this issue would look like a special case of proper names, something that should be easily addressed by generating specific signs (basically one for every station). However, the solution is not as simple as it seems. Indeed, several problematic aspects are hidden once looking at the linguistic situation of names in LIS (and more generally in SL).

1.1 Lexical issues

The linguistic situation of names is quite heterogeneous in LIS and can be summarized as follows:

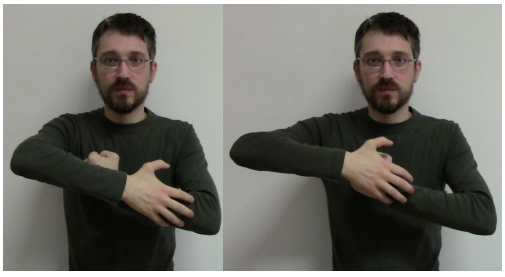
1. Sign names fully acknowledged by the Italian Deaf communities.
2. Sign names only acknowledged by (part of) the local Deaf community.
3. There is no sign name even within the local community.

The first option illustrates the case of most main stations in big cities. Normally, the name of the station is semantically transparent, as in (1a) or it involves the name of some prominent character of the Italian history, as in the case of "Milano Porta Garibaldi" (Garibaldi was the hero of the Italian unification).



(1) MILANO CENTRALE

Unfortunately, however, most of the trains go to and stop at anonymous locations. In some cases, local dialects have a specific sign for those stations (normally, the name of the town where the train stops) as in (2).



(2) CASTELVETRANO

Finally, there are Italian names for which not even the local Deaf community has already developed a local sign name. In those cases, human signers adopt the last resorts at their disposal, namely either they fingerspell the name, or they labialize it, as in the case of "Rebaudengo Fossano", a small village outside Turin.
Fingerspelling is the typical way in which borrowings from spoken languages are realized (Brentari, 2000). However, this practice is not fully adopted by the Italian Deaf communities yet. Indeed, old signers may not know the manual alphabet and in some cases they even refuse to use it, rather preferring labializing the forms in spoken Italian (Volterra, 1987 and Caselli et al., 1996).
Once we leave the domain of human signers and enter the world of avatar signers, additional issues are raised which are specifically connected to the fingerspelling and labializing strategies. Clearly, labialization is a solution that cannot be usefully pursued for practical reasons: The avatar technology is designed to be portable on different devices including smartphones. Within this framework, lipreading would be almost impossible for most users of the service. Furthermore, working in the domain of public transportation announcements, the timing issue is not trivial. Announcements are normally broadcasted and fingerspelling would introduce additional delay to the sign production, which normally is more time consuming than speech.

2. A non-manual practical solution

After having preliminarily consulted some members of the local Deaf Association of the city where the automatic translation system will be first released (ENS Torino), a twofold solution is going to be adopted:

1. Sign names fully acknowledged by the Italian Deaf communities will be maintained by the signing avatar.
2. Blended written Italian-LIS sign forms will be used.

While names of main stations in big cities are preserved in their original LIS forms, as in (3), a new strategy is developed for less-familiar stations and gaps in the vocabulary. The avatar will play a classifier sign indicating a wide board while the name of the station will appear in written Italian "centered on the board", as shown in (4).

(2) MILANO CENTRALE

This technical solution blends a manual sign (a generic classifier) with a non-manual component. However, rather than using the standard non-manual channels (facial expressions or body postures), this solution adopts a tool which is not internal to sign language, namely the written form of the dominant language. From the communicative perspective, this solution is much more performative than standard fingerspelling for at least three reasons:

1. It allows a faster assessment of the lexical item since the written input is produced simultaneously and not letter by letter
2. It does not overload the processing of the entire sentence
3. It is accessible to all signers, even those with lower levels of literacy.

From the timing perspective, blended forms are much quicker to perform than fingerspelling making the entire announcement more alignable with its spoken counterpart. An issue to be developed further is how long the blended form must last on the screen. We are planning to use knowledge from reading times in Deaf subjects with low literacy to determine it. At the moment we do not exclude the possibility that longer names will display longer than shorter ones.

(3) REBAUDENGO FOSSANO

3. Technical issues

We are developing our idea for station names inside the ATLAS architecture (Mazzei et al. 2013). The ATLAS project concerned the translation from Italian to LIS in the specific application domain of the weather forecasts. The ATLAS system is a knowledge-based and restricted-interlingua translation system, since it uses extra-linguistic information and deals with only two languages.
The system is a pipeline composed by five distinct modules (Figure 1). The modules are: (1) a dependency parser for Italian; (2) an ontology based semantic interpreter; (3) a generator; (4) a spatial planner; (5) an avatar that performs the synthesis of the sequence of

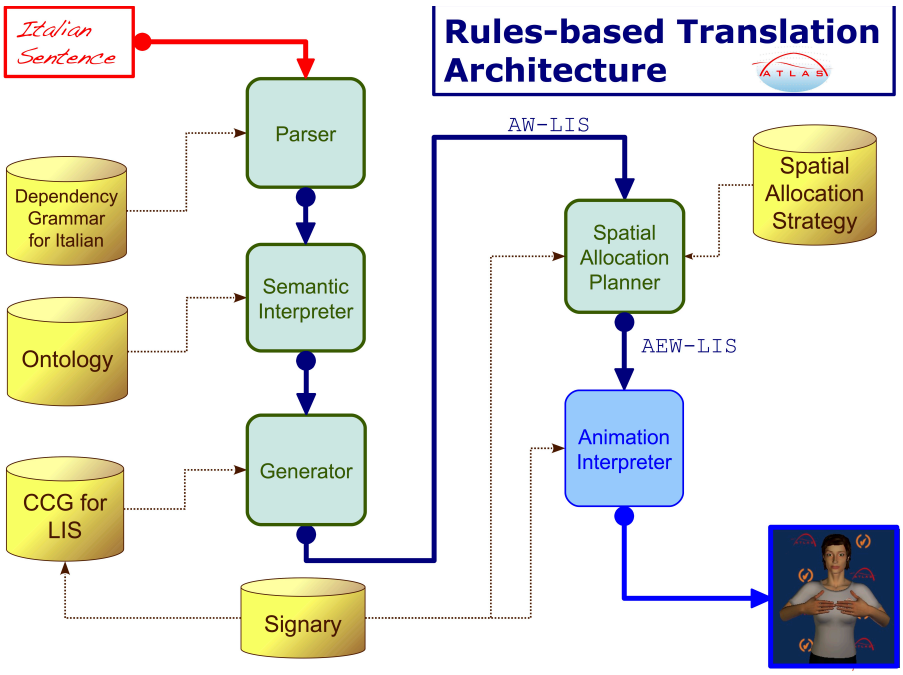


Figure 1: ATLAS architecture

signs, i.e. the final LIS sentence.

n order to integrate our solution in the ATLAS architecture, we need to modify the generator and the avatar. The ATLAS generator is composed by two submodules: the SentenceDesigner microplanner and the OpenCCG realizer (Mazzei 2012). The SentenceDesigner is an expert system that decides about the syntactic organization and which signs to use in the generation. In contrast, the realizer decides about the signs order and their inflections. So, we need to implement a double access procedure to the signing lexicon in SentenceDesigner. In a first attempt, SentenceDesigner will search in the lexicon for a direct translation of an Italian station name into LIS (see above "Milano centrale"). If at least one translation is found, then the avatar follows the standard ATLAS communication pipeline and performs the (sequence of) sign(s). In contrast, if this procedure does not produce results, for instance when there is a lexical gap in the LIS dictionary for the station name, SentenceDesigner commands the avatar to produce the Italian-LIS blending for that specific station name in real time. Moreover we need to augment the avatar to allow for the production of a real time Italian-LIS blending from a string (up to 40 characters). Finally, we need to augment the communication protocol between SentenceDesigner and the avatar, by adding a new tag to the AEWLIS (ATLAS Extended Written LIS), i.e. to the XML language in use for the communication between the generator and the avatar.

4. Social issues

Last but not least, we are also concerned with the impact of our choices for the broad Deaf communities. On the one hand, the use of written forms blended along with the sign stream is a technical solution to a practical problem. On the other hand, for the Deaf communities the risk exists that a wrong message is sent that sign languages are not fully adequate to all communicative situations. We are planning to assess these aspects with an on-line questionnaire in which we ask the Italian Deaf communities i) which form they prefer for both famous and less-known destinations: Sign name, Fingerspelling or written blending; and ii) whether they feel the blending solution as dangerous for their sign language.

5. Conclusions

One of the most challenging aspects of avatar translation from spoken into SL is how to implement NMM, which are normally exploited by signers during the sign stream. This is true both for lexical NMMs and those with phrasal scope (Van Zijl and Combrink, 2006). While this domain opens several research questions, most of which without a clear solution (Ong and Ranganath, 2005), we showed that an additional non-manual option is made available by current technologies, which avatars may resort to when the contextual situation requires it. Written text blending is an economic solution to a practical problem posed by the timing of public transportation announcements.

6. Acknowledgments

This work has been partially supported by the project LIS4ALL, partially funded by Regione Piemonte, Innovation Hub for ICT, 2011-2014, POR-FESR 07-13.
Part of the research leading to these results also received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement N°324115–FRONTSEM (PI: Schlenker). Research was partially conducted at Institut d'Etudes Cognitives (ENS), which is supported by grants ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0087 IEC. This work is dedicated to Leonardo Lesmo who substiantially contributed to its realization.

7. References

Almasoud, A. M. and Al-Khalifa, H. S. (2011). A proposed semantic machine translation system for translating arabic text to arabic sign language. In Proceedings of the Second Kuwait Conference on e-Services and e-Systems, KCESS '11, New York, NY, USA. ACM, pp. 23:1–23:6.

Braffort, A. et al. (2010). Sign language corpora for analysis, processing and evaluation. In Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Brentari, Diane (ed.) (2000). Foreign Vocabulary in Sign Languages. Mahwah, NJ: Lawrence Erlbaum Associates.

Caselli et al. (1996). Linguaggio e sordità. Il Mulino, Bologna.

Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., and Abbott, S. (2002). Tessa, a system to aid communication with deaf people. In Proceedings of the fifth international ACM conference on Assistive technologies. ACM, pp. 205–212.

Ebling, S. and Glauert, J. (2013). Exploiting the full potential of JASigning to build an avatar signing train announcements. In: Third International Symposium on Sign Language Translation and Avatar Technology, Chicago, IL, USA, 18 October 2013 - 19 October 2013.

Mazzei, A. (2012). Sign Language Generation with Expert Systems and CCG. In Proceedings of the 7th International Natural Language Generation Conference, Starved Rock State Park Utica, IL USA. Association for Computational Linguistics, pp. 105–109.

Mazzei, A., Lesmo, L., Battaglino, C., Vendrame, M., and Bucciarelli, M. (2013). Deep natural language processing for italian sign language translation. In Proc. of XIII Conference of the Italian Association for Artificial Intelligence, volume 8249 of LNCS. Springer), pp. 193–204.

Ong, S. C. W. and Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Trans. Pattern Anal. Mach. Intell., 27(6):873–891.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., and D'Haro, L. F. (2012). Design, development and field evaluation of a spanish into sign language translation system. Pattern Anal. Appl., 15(2):203–224.

Van Zijl, L. and Combrink, A. (2006). The south african sign language machine translation project: Issues on non-manual sign generation. In Proceedings of the 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, SAICSIT '06, Republic of South Africa. South African Institute for Computer Scientists and Information Technologists, pp. 127–134.

Annotation of Mouth Activities with iLex

Thomas Hanke
Institute of German Sign Language and Communication of the Deaf,
University of Hamburg
thomas.hanke@sign-lang.uni-hamburg.de

Abstract

This paper describes the support for mouth activity annotation provided by the iLex annotation workbench on a holistic level connected to the lexical database, on a feature level, as well as in the context of semi-automatic annotation.

Keywords: Annotation, mouthing, mouth gesture, lexical database

1. Introduction

In a purely bottom-up approach an annotation practice used for mouth activities would try to describe the phenomena and leave it to a second step to classify (e.g. between mouthing and mouth gestures) and relate (e.g. to spoken language words) (cf. Keller, 2001). For practical reasons, however, the first step is often skipped, and separate coding systems are applied to what is categorised either as mouthing derived from spoken language or mouth gesture where there is no obvious connection between the meaning expressed and any spoken language words expressing that same meaning. This happens not only for time (=budget) reasons, but also because it is difficult for coders to describe mouth visemes precisely if the sign/mouth combo already suggests what is to be seen on the mouth. While there are established coding procedures to avoid influence as far as possible (like only showing the signer's face, provided video quality is good enough), they make the approach very time-consuming, even if not counting quality assurance measures like inter-transcriber agreement. Some projects undertaken at the IDGS in Hamburg therefore leave it with a spoken-language-driven approach: The mouth activity is classified as either mouth gesture or mouthing, and in the latter case the German word is noted down that a competent DGS signer "reads" from the lips, i.e. that word from the set of words to be expected with the co-temporal sign in its context that matches the observation. Standard orthography is used unless there is a substantial deviation. For mouth gestures, holistic labels are used. These two extremes span a whole spectrum of coding approaches that can be used for mouth activities. We present different aspects of how iLex, the Hamburg sign language annotation workbench, supports the whole range of solutions from more time-series-like systems to those evaluating co-occurrence and semantic relatedness, from novice-friendly decision trees to expert-only modes to support semi-automatic annotation.

2. iLex Background

Unlike other transcription environments, iLex does not follow a document-centric approach, but keeps all annotation in a relational database. Consequently, tags are not simply text, but are structured database entities themselves, such as tokens describing an instance of a type. This allows the user to immediately access other tokens of the same type as (phonetic and context) data, as a video snippet, or an avatar performance. The complete integration of a lexical database into the annotation process in our view is crucial when transcribing a language not having an established written form.[1]

Mouth activities, are not part of the token records, but are annotated as text tags on a separate tier.[2] Being text tags, mouthings are not considered as instantiations of spoken language lexemes, the tag is a mere form description. However, this does not mean that mouth activity annotation does not profit from the integrated approach:

3. Mouthing in the Lexical Database

In the iLex lexical database, types have a field to store a default mouth activity typically co-occurring with the sign. In some cases, certain mouth gestures are an integral part of the sign, these would be stored here. In the case of lexicalised form-meaning combinations[3], one or more mouthings can be stored here that typically occur in this context.

As these mouthings are good candidates for the mouth tier tags overlapping with a certain token, iLex provides easy access to them via a context menu to create the mouth tag.

The iLex database can be set up to provide extra suggestions here, e.g. all mouthings that the informant currently being transcribed has already used in combination with the token's type.

Only if the observed mouth activity does not match with any of the suggestions, the user needs to open a specialised editor in order to describe the observation (cf. section 4).

[1] A more detailed description of the iLex workbench can be found in Hanke, 2002, Hanke/Storz, 2008 and Hanke et al., 2010.

[2] The reason for this is evident: Mouthings can stretch over more than one sign (token).

[3] For details on the type hierarchy implemented in iLex and how it is explored for modelling the sign lexicon, cf. Konrad et al., 2012.

At the same time, mouthings are an essential bit of information in the lemma revision process: When a few tokens for a lexicalised form-meaning combination co-occur with mouthings that derive from spoken language words not semantically related, this might be an indication that they actually belong to another type, even if they share the same (manual) form.

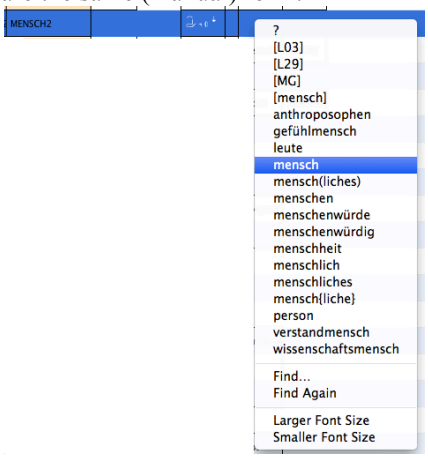


Figure 1: Context menu for mouthing to be cotemporal with the sign MENSCH2

4. Mouth Editor

iLex currently supports three conventions how to store mouthings as text: Orthography, IPA, and SAMPA.

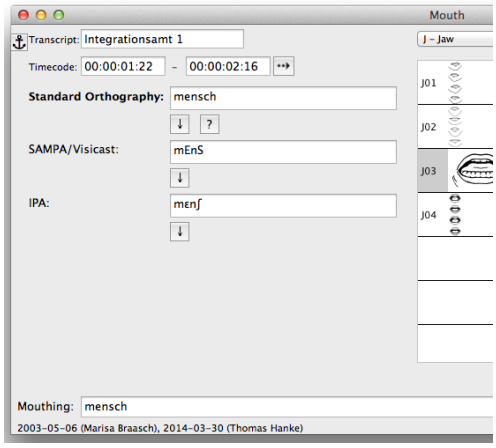


Figure 2: Left side of mouth activity editor: Mouthing

As visemes are equivalence classes of visually indistinguishable phonemes, any member of the class can represent the viseme, allowing visemes to be encoded by a subset of IPA. Whether one always uses the same IPA letter for one class or keeps with the original phoneme, is irrelevant for describing the observation, but certainly makes a difference when testing two annotations for equalness. SAMPA (cf. Gibbon et al., 1997) was suggested in the context of the ViSiCAST and eSIGN projects (cf. Hanke, 2004) to describe visemes as SAMPA text is (was) easier to handle (being ASCII text) than IPA. However, for the purpose of viseme labelling, SAMPA can simply be considered a coding variation of IPA.

As said in the introduction, using spoken language orthography seems weird to describe visemes, but has its advantages, not limited to the transcribers' convenience. The pronunciation data in iLex allow the program to derive the viseme sequence from the orthography entered. For German, iLex also manages to derive the viseme gestalt for abbreviated mouthings from the abbreviated orthography as well as to compute the viseme gestalt for compounds.

iLex allows the user to annotate mouth gestures on separate tiers or in line with mouthings. In the latter case which seems preferable to us, some distinguishable code set is needed to tell mouth gesture codes apart from mouthing. For this reason, we use the convention to include mouth gesture codes in square brackets.

A specialised editor for using a mouth gesture code set introduced in the ViSiCAST project (Hanke et al., 2001) is implemented in iLex. As these codes are rather arbitrary, is is most important that the system supports the user by showing a textual and video description for the code selected.

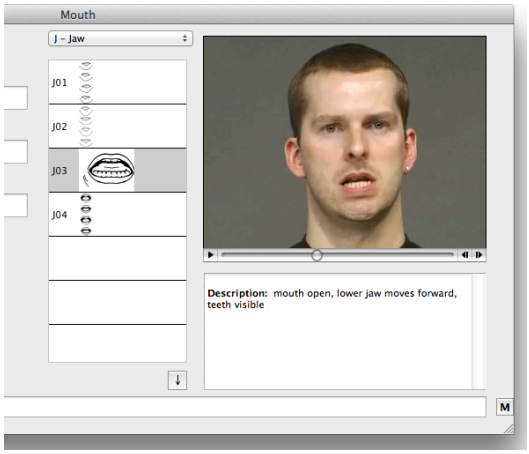


Figure 3: Right side of mouth activity editor: Mouth gestures

Nevertheless, as some mouth gestures occur very rarely, iLex also offers an experimental "expert system" to determine the right code: Following the ideas of Sutton-Spence/Day (2001), the user has to answer a number of relatively easy questions on his/her observation, and system then provides the code.

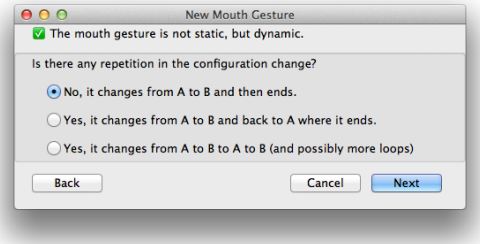


Figure 4: "Expert system" for entering mouth gestures by answering a series of questions

5. Tag Alignment

Unless one is interested in the exact timing of mouth activities, iLex allows the user to set up the mouth tier to depend on the token tier in order to save time: Tag boundaries are then shared between these tiers, but mouth tags can still span several token tags.

For DGS, we observe some signers who (sequentially) combine mouthing and mouth gesture within one manual

sign. In the case of separate tiers for mouthings and mouth gestures, this means that a sub-sign granularity is necessary, i.e. the tiers have to be set up not to depend on each other. For one common tier for mouthings and mouth gestures, codes can simply be concatenated into one tag spanning the whole sign duration.

6. Compatibility with the eSIGN Approach

The main components of the eSIGN software are an avatar system that is able to sign from phonetic data (cf. Elliott et al., 2004) and an editor that allows scripting of such avatar performances (cf. Hanke, 2004). In order to avoid re-writing the necessary phonetic information, the editor works with a local database or links into the iLex database. However, from within iLex it is also possible to save a transcript as an eSIGN document. Obviously, for this to work the transript needs to contain all necessary phonetic descriptions. With respect to mouthings and mouth gestures, this means that the data is coded in one of the aforementioned systems. If orthography is used, the conversion relies on available pronunciation data. If another coding system is used for mouth gestures, the user can still provide a mapping onto the eSIGN formats for the conversion to work.
For the iLex user, this approach has the advantage that an anonymised version of a sign performance can be created with minimal effort.

7. Feature-Level Annotation

For detailed phonetic analysis iLex provides another mechanism than simple textual tags: Binary features. By assigning a closed vocabulary to a binary features tier, iLex prompts the user with a list of all the features (the elements of the vocabulary) in order to check those that apply for the tagged time stretch. This approach still works with rather large sets of features when it is no longer feasible to reserve one tier per feature.

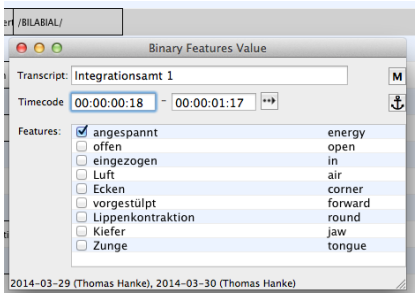


Figure 5: The feature angespannt=tense alone from the full set displays in the transcript as /BILABIAL/ from Bergman/Wallin's reduced set

Here we show an example using the feature set from Bergman/Wallin, 2001.
The display in the transcript can just be the selected features or any function thereof. In the example, the display is automatically computed from the input features by means of a user-provided mapping table, in this case implementing the Bergman/Wallin reduced feature set.

8. Towards Semi-Automatic Annotation

While lipreading is known to be a hard problem both for humans and automatic systems, it is a lot easier to identify the mouthing given the identity of the sign coarticulated as that sign narrows down the search space to only a couple of probable mouthings. We currently experiment with feature vectors obtained from short-range 3D sensors imported into iLex transcripts in order to first determine whether there is mouth activity during a sign, and if so, which of the candidate mouthings best fits with the feature vectors observed. Even when applying some thresholding, this approach increases the risk that unusual sign/mouthing combinations remain undetected. It therefore remains to be seen if this automation is a time saver when a certain annotation quality is to be guaranteed.

9. Acknowledgements

This publication has been produced in the context of the joint research funding (DGS Corpus) of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.
The research leading to these results has also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135 (Dicta-Sign).

10. References

Bergman, B., Wallin, L. (2001) A preliminary analysis of visual mouth segments in Swedish Sign Language. In Boyes Braem, P. & Sutton-Spence, R. (eds.): *The hands are the head of the mouth*. Hamburg: Signum, pp. 51-68.

Elliott, R., Glauert, J., Jennings, V., Kennaway, R. (2004). An overview of the SiGML notation and SiGMLsigning software system. In: O. Streiter, O. & Vettori, C. (eds.): *From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication. Proceedings of the Workshop on Representing and Processing of Sign Languages.* 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal. Paris: ELRA, pp. 98–104.

Gibbon, D., Moore, R., Winski, R. (eds., 1997): *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: de Gruyter.

Hanke, T. (2002). iLex. A tool for Sign Language Lexicography and Corpus Analysis. In González Rodriguez, M. & Paz Suarez Araujo, C. (eds.): *Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain*; Vol. III. Paris: ELRA, pp. 923–926.

Hanke, T. (2004). Lexical Sign Language Resources: Synergies between Empirical Work and Automatic Language Generation. Paper presented at the Fourth

International Conference on Language Resources and Evaluation, LREC 2004, Lisbon, Portugal.

Hanke, T., Langer, G., Metzger, C. (2001). Encoding non-manual aspects of sign language. In Hanke, T. (ed.), ViSiCAST Report D5.1: Interface Definitions.

Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In Crasborn, O., Efthimiou, E., Hanke, T., Thoutenhoofd, E.D, Zwitserlood, I. (eds.): *Construction and Exploitation of Sign Language Corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages.* 6th International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Maroc. Paris: ELRA, pp. 64–67.

Hanke, T., Storz, J., Wagner, S. (2010). iLex – Handling multi-camera recordings. In Dreuw, P., Efthimiou, E., Hanke, T., Johnston, T., Martínez Ruiz, G., Schembri, A. (eds.): *Corpora and Sign Language Technologies. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages.* 7th International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta. Paris: ELRA, pp. 110–111.

Keller, J. (2001). Multimodal representations and the linguistic status of mouthings in German Sign Language (DGS). In Boyes Braem, P. & Sutton-Spence, R. (eds.): *The hands are the head of the mouth*. Hamburg: Signum, pp. 191–230.

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., Regen, A. (2012). From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In Crasborn, O. Efthimiou, E., Fotinea, S.-E., Hanke, T., Kristoffersen, J., Mesch, J. (eds.): *Interaction between Corpus and Lexicon. Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages.* 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey. Paris: ELRA, pp. 87–94.

Sutton-Spence, R., Day, L. (2001). Mouthings and mouth gestures in British Sign Language (BSL). In Boyes Braem, P. & Sutton-Spence, R. (eds.): *The hands are the head of the mouth*. Hamburg: Signum, pp. 69–85.

Release of Experimental Stimuli and Questions for Evaluating Facial Expressions in Animations of American Sign Language

Matt Huenerfauth
The City University of New York (CUNY)
Computer Science Department, Queens College
65-30 Kissena Blvd, Flushing, NY 11367 USA
E-mail: matt@cs.qc.cuny.edu

Hernisa Kacorri
The City University of New York (CUNY)
Computer Science Program, The Graduate Center
365 Fifth Ave, New York, NY 10016 USA
E-mail: hkacorri@gc.cuny.edu

Abstract

We have developed a collection of stimuli (with accompanying comprehension questions and subjective-evaluation questions) that can be used to evaluate the perception and understanding of facial expressions in ASL animations or videos. The stimuli have been designed as part of our laboratory's on-going research on synthesizing ASL facial expressions such as Topic, Negation, Yes/No Questions, WH-questions, and RH-questions. This paper announces the release of this resource, describes the collection and its creation, and provides sufficient details to enable researchers determine if it would benefit their work. Using this collection of stimuli and questions, we are seeking to evaluate computational models of ASL animations with linguistically meaningful facial expressions, which have accessibility applications for deaf users.

Keywords: American Sign Language, facial expression, non-manual signals, stimuli.

1. Introduction

Synthesis of American Sign Language (ASL) animations can provide benefits for deaf and hard-of-hearing people with lower levels of written language literacy (Huenerfauth, 2004a). This is underscored by the literacy rates of deaf adults in the United States on standardized testing (Traxler, 2000) and the large number of ASL users (over 500,000) in the United States (Mitchell et al., 2006). In prior experimental studies, we determined that the use of emotional and linguistically meaningful facial expressions in ASL animations significantly increased viewers' comprehension and perceived quality of animations (Huenerfauth, Lu, and Rosenberg, 2011). To produce an animation with natural facial expressions, a skilled animator and ASL signer could carefully control the face of the avatar on a fine-grained timeline, but such an approach is time-consuming and depends too much on the skills of the animator. Thus, a more automated solution is needed to minimize the required input in order to produce an animation; this minimal input script would include only the sequence of glosses, the type of facial expression needed, and the starting and ending glosses in the sentence when it should occur.

Many prior sign language animation systems lack sophisticated models in support of non-manuals, which are necessary to automatically synthesize clear and understandable facial expressions. There has been recent work by several groups to improve the state-of-the-art of facial expressions and non-manual signals for sign language animation, e.g.: Wolfe et al. (2011) and Schnepp et al. (2012) used linguistic findings to drive eyebrow movement in animations of interrogative (WH-word) questions with or without co-occurrence of affect. Schmidt et al. (2013) used clustering techniques to obtain lexical facial expressions. Gibet et al. (2011) used machine-learning methods to map facial motion-capture data to animation blend-shapes.

This paper presents a collection of stimuli to evaluate the perception and understanding of facial expressions in ASL animations. Section 2 describes the research project for which the stimuli were developed; section 3 provides basic information about the stimuli and briefly explains the linguistics of the facial expressions within each. Section 4 gives additional detail about how the stimuli and questions were engineered to measure the perception and comprehension of facial expressions. Section 5 describes how facial movements in the stimuli videos were identified and recorded, and section 6 describes prior studies that used some of these stimuli. Section 7 contains information about how to obtain the collection.

2. Our Research on ASL Animation

The goal of our ongoing research is to improve technologies for generating ASL animations through the inclusion of linguistically meaningful ASL facial expressions. We seek to develop computational models to generate facial expressions that convey grammatical syntax information such as topic, negation, rhetorical questions, WH-word questions, and yes/no questions (Kacorri, 2013). It is necessary to model how elements of the face move during ASL facial expressions, how these movements are timed in relation to the manual signs, and how these facial movements co-occur or segue into one another. In pursuit of this goal, our lab has begun to analyze linguistically annotated ASL videos (Liu et al., 2013) and automatically tracked facial landmarks in these videos (Yu et al., 2013) so that we may create signer-independent models that can generate grammatically correct ASL animations with facial expressions.

To evaluate our animation models, native ASL signers typically view our animations and answer subjective Likert-scale and comprehension questions (Huenerfauth, 2004b; Huenerfauth et al., 2007; Huenerfauth, 2008). Inventing stimuli and comprehension questions that effectively measure whether participants understand the information conveyed specifically by the model-driven face can be challenging. Several facial expressions affect

the meaning of ASL sentences in subtle ways (Kacorri, Lu, and Huenerfauth, 2013b) and often signers may not consciously notice a facial expression during an ASL passage (Huenerfauth, Lu, and Rosenberg, 2011; Kacorri, Lu, and Huenerfauth, 2013b).

During our multi-year project, we have experimented with different forms of stimuli design strategies to elicit ASL passages and comprehension questions that can measure whether the viewer has understood linguistic facial expressions correctly (Kacorri, Lu, and Huenerfauth, 2013b). After three years of user studies on ASL facial expressions that convey grammatical syntax information (Huenerfauth, Lu, and Rosenberg, 2011; Kacorri, Lu, and Huenerfauth, 2013a; Kacorri, Lu, and Huenerfauth, 2013b; Kacorri, Harper, and Huenerfauth, 2013), we have designed a collection of scripted ASL multi-sentence single-signer passages and corresponding comprehension questions that probe whether human participants watching these stimuli have understood the information that should have been conveyed specifically by the facial expressions. We are now sharing with the research community the set of stimuli and questions we have developed in support of our research on non-manual linguistic phenomena.

3. Overview of the Collection

This paper is the first announcement of the release of this stimuli collection, which includes: 48 ASL passages performed by a native signer; 192 comprehension questions (4 questions for each passage, each question performed by 2 native signers, male and female); a set of Likert-scale subject questions about the grammatical correctness, ease of understanding, and naturalness of movement of the passages; and a set of Likert-scale questions asking whether participants noticed specific categories of facial expressions. The collection consists of video recordings of a native ASL signer, ASL transcriptions of each passage, English translation of the ASL passages and comprehension questions as plaintext files, and two sets of questionnaires with the Likert-scale questions. The English translations of the ASL stories includes both the indented meaning when the ASL facial expression is performed correctly and a second ambiguous meaning when the facial expression is not correctly perceived by the person viewing the story.

Each stimulus focuses on a particular facial expression in one of the following categories listed below. Each is illustrated in Figure 1 and informally described below; please consult ASL linguistics references for more detailed explanations, e.g., (Neidle et al., 2000).

- Yes/No Questions: The signer raises his eyebrows while tilting the head forward during a sentence to indicate that it should be interpreted as a question.
- WH-Questions: The signer furrows his eyebrows and tilts his head forward during a sentence that should be interpreted as information-seeking, typically with a "WH" word such as what, who, where, when, how, which, etc.
- RH-Questions: The signer raises his eyebrows and tilts his head backward and to the side to indicate a question that should be interpreted rhetorically.

Figure 1: Still images taken from videos included in the stimuli collection described in this paper, with each image illustrating a moment when a particular facial expressions is occurring: (a) YN-Question, (b) WH-Question, (c) RH-Question, (d) Topic, (e) Negation, and (f) Emotional Affect (an example of anger is shown in this image).

- Topic: The signer raises his eyebrows and tilts his head backward during a phrase at the beginning of a phrase that should be interpreted as a topic.
- Negation: The signer shakes his head left and right during the verb phrase which should be interpreted with a negated meaning, often with the sign NOT.
- Emotional affect: These facial expressions are not linguistically governed, but they include several typical affective facial expressions that can indicate sadness, anger, frustration, etc. during a sentence.

The value of this collection is that the stories and questions were carefully engineered so that the participant must perceive and understand the facial expression in order to answer the comprehension questions correctly. For each stimulus, if the manual portion of the performance were considered alone (without the facial expressions), then there would be an ambiguity or an alternative semantic interpretation possible for the stimulus. Our comprehension questions have been designed to detect when a participant has misunderstood the stimulus due to the facial expression not being successfully perceived or understood. Thus, these stimuli can be used to evaluate the quality of automatic animation-synthesis systems for generating animations of ASL with facial expressions.

Table 1 provides a listing of the number of stimuli in the collection of each type.

Table 1: Collection Overview.

Type of facial expression	Number of stimuli (Average number of glosses per stimulus)	Codenames of these stimuli in the collection
Emotional Affect	8 stimuli (6.88)	E1, E2, E3, E4, E5, E6, E7, E8
WH-word Questions	9 stimuli (13)	W1, W2, W3, W4, W5, W6, W7, W8, W9
Yes/No Question	7 stimuli (9.29)	Y1, Y2, Y3, Y4, Y5, Y6, Y7
Topic	7 stimuli (10)	T1, T2, T3, T4, T5, T6, T7
Rhetorical Question	11 stimuli (11.82)	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10, R11
Negation	6 stimuli (16.5)	N1, N2, N3, N4, N5, N6

4. Design of Stimuli and Questions

Prior to the design of the stimuli, a native ASL signer was given 6 categories of facial expressions and was introduced to premise that the passage must be ambiguous in its meaning if the facial expression were not understood. The native ASL signer invented, performed, and transcribed the ASL passages, and the passages were discussed and edited in collaboration with a team of other native ASL signers at the laboratory. Next, the two ambiguous meanings were translated into English sentences. Consulting the ASL transcription and the two ambiguous English translations, a second native ASL signer performed the ASL passages for the video recordings in our collection. Finally, linguistic researchers at our laboratory engineered the comprehension questions for each story such that they would receive different answers, depending on the perception and understanding of the facial expression. The collection includes a sample HTML form where the 4 comprehension questions are embedded in video format and the answers are collected on a 7-point Likert scale from "definitely no" to "definitely yes."

While researchers can access the full collection of stimuli and questions, this section explains a specific example of each category of stimuli to illustrate how each stimulus can have alternative interpretations, if the facial expression were not correctly understood.

4.1 Example: Topic

The following sentence is an example of a stimulus with a Topic facial expression (which should occur during the gloss "SWEET FOOD"): NEW RESTAURANT INCLUDE PASTA PIZZA SWEET FOOD MY SISTER COOK EXPERT. When the Topic face is perceived, then the stimulus has the approximate meaning: "The new restaurant has pasta and pizza. As for sweet foods (pastries), my sister is an expert chef." We have intentionally designed the stimulus so that it is performed at a human conversational speed without any long pauses during the signing that would emphasize the sentence boundary before "SWEET." This has been done so that the meaning of the stimulus is strongly affected by whether the viewer perceives the Topic facial expression. When the Topic face is not perceived, then the sentence boundary may be less clear (especially when the sentence is performed by an animated avatar that typically lacks the subtle acceleration and timing of a human signer). In such a case, the viewer may interpret "SWEET FOOD" as being the third item in the list of foods available at the restaurant; thereby the stimulus has the meaning: "The new restaurant has pasta, pizza, and sweet foods (pastries). My sister is an expert chef." One of the comprehension questions for this stimulus is: Does the new restaurant have sweet foods? The answer depends on whether the Topic facial expression was perceived and understood.

4.2 Example: WH-Word Questions

The following sentence is an example of a stimulus with a WH-Question facial expression (which should occur during the glosses "HER BIRTHDAY PARTY WHEN"): THAT MARY HER BIRTHDAY PARTY WHEN MARY DRUNK. When the WH-Question face is perceived, then the stimulus has the approximate meaning: "When is Mary's birthday party? Mary is drunk." When the WH-Question face is not perceived, then it may be less clear to the viewer where the sentence boundary is located. In such a case, the viewer may interpret "WHEN MARY DRUNK" as a question (albeit in English-like word order); thereby the stimulus would have the meaning: "It is Mary's birthday party. When did Mary got drunk?" One of the comprehension questions

for this stimulus is: Does Charlie know when the party is? (The signer appearing in the video is introduces as "Charlie" at the beginning of the study.) The participant is more likely to answer "no" to this question if the WH-Question facial expression was correctly perceived.

4.3 Example: Rhetorical Questions

The following sentence is an example of a stimulus with a RH-Question facial expression (which should occur during the glosses "WHY"): ALEX NOW GO-GO PARTIES WHY FINISH DIVORCE. When the RH-Question face is perceived, then the stimulus has the approximate meaning: "Alex is now often going to parties because he is divorced." When the RH-Question face is not perceived, then the sentence boundary may be less clear. In such a case, the viewer may interpret "WHY FINISH DIVORCE" as a question; thereby the stimulus has the meaning: "Alex is now often going to parties. Why did he get divorced?" One of the comprehension questions for this stimulus is: Does Charlie know why Alex started going to parties? The answer depends on whether the RH-Question facial expression was perceived and understood.

4.4 Example: Yes/No Questions

The following sentence is an example of a stimulus with a Yes/No Question facial expression (which should occur during the glosses "ALL FOOD CHEAP POINT"): BOB'S DINER THAT YOUR SISTER HER FAVORITE RESTAURANT ALL FOOD CHEAP POINT. When the YN-Question face is perceived, then the stimulus has the approximate meaning: "Bob's Diner is your sister's favourite restaurant. Is all the food cheap?" When the YN-Question face is not perceived, then the final sentence could appear to be a declarative statement. Thus, the stimulus has the meaning: "Bob's Diner is your sister's favourite restaurant. All the food is cheap." One of the comprehension questions for this stimulus is: Does Charlie know if the restaurant is expensive? If the YN-Question facial expression was correctly perceived and understood, then the participant is more likely to answer no to this question.

4.5 Example: Negation

The following sentence is an example of a stimulus with a Negation facial expression (which should occur during the glosses "HAVE SCIENCE CLASS"): ALEX TEND TAKE-UP MATH CLASS. NOW SEMESTER, SCHOOL HAVE SCIENCE CLASS. ALEX TAKE-UP TWO CLASS." When the Negation face is perceived, then the stimulus has the approximate meaning: "Alex usually takes math classes. This semester, the school doesn't have any science classes. Alex is taking two classes. " When the Negation face is not perceived, then the meaning of the middle sentence is inverted: "This semester, the school has science classes." One of the comprehension questions for this stimulus is: Does the school have science classes this semester? The answer depends on whether the Negation facial expression was perceived and understood.

4.6 Example: Emotional Affect

The following sentence is an example of a stimulus with an emotional affect facial expression (this example includes an angry facial expression during the entire sentence): LAST FRIDAY, MY BROTHER TAKE MY CAR. DRIVE SCHOOL. When the emotional affect facial expression is perceived, then the stimulus has the approximate meaning: "Last Friday, my brother took my car to drive to school." (The sentence has the subtext that the signer is upset about this.) When the emotional affect face is not perceived, then this subtext is not conveyed. One of the comprehension questions for this stimulus is: Is Charlie angry at his brother? The answer depends on whether the emotional facial expression was perceived and understood.

4.7 Likert-scale Questions

In addition to the four comprehension questions that are designed specifically for each stimulus, this collection also includes a set of Likert scale questions that can be used to measure participants' subjective evaluation of each. The set of Likert scale questions is identical for all of the stimuli, and it includes three subjective evaluation questions and four questions measuring whether participants' noticed a particular facial expression.

- "Good ASL grammar?": A subjective evaluation question of how grammatically correct was the presented signing with answers on a 1-to-10 Likert scale where 1 indicates bad and 10 perfect.
- "Easy to understand?": A subjective evaluation question on comprehensibility of the signed message with answers on a 1-to-10 scale where 1 indicates confusing and 10 clear.
- "Natural?": A subjective evaluation question on how naturally moving the signer appeared with answers on a 1-to-10 scale where 1 indicates that the signer moves like a robot and 10 that the signer moves like a person.
- "Did you notice a … facial expression?": Four questions in relation to how much participants noticed an emotional, negative, interrogative, or topic facial expression during the story with answers on a 1-to-10 scale from "yes" to "no".

The collection includes an HTML questionnaire with these Likert-scale questions and the options for the answers as radio buttons.

5. Facial Feature Extraction on Recordings

We used automatic face tracking software (Visage Technologies, 2014) to analyze the video recordings of the 48 ASL passages and produce files that contain information about the head pose and facial features of the human signer for each frame of the video. The tracking results, part of the collection, are shared as comma-separated values (CSV) files. Head pose data is given as translation from the camera in the 3 dimensions (x, y, z) and as head rotation (pitch, yaw, roll). The obtained facial features follow the MPEG-4 facial action parameters (Tekalp, 1999) for each frame of the video. For example, the eyebrow position in every frame is defined by 8 facial action parameters (FAP30-FAP37) as the vertical and horizontal displacement of the left and right eyebrow from a neutral pose of the signer's face. This information could be used by future researchers to

animate the face of a virtual human character (Pandzic and Forchheimer, 2003) performing these stimuli passages. Such a character could be displayed as a baseline for comparison in an experimental evaluation study.

For optimal results, the Visage software was used in offline mode. The quality of the results is bounded by the performance of the software on the video recordings and the initial manual process of mask fitting to the face as shown in Fig. 2. For example, the tracker may lose the face if the head movement is too fast or if large parts of the face are covered, e.g. by the hands. We observed that this is happening for 0%-7.6% (avg. 1.6%) of the story duration in our stimuli collection. In this case, the lost frames are indicated with a tracking status other than "OK" in the comma-separated values file, and all the extracted head and facial features would normally have the value 0 in such cases. We processed the data and filled in the values of the lost frames using spline interpolation (smoothing degree 1) while maintaining the tracking status information. Although interpolation may work well for the facial feature values, it can sometimes be problematic for head rotation, because it is currently represented in the form of Euler angles (pitch, yaw, roll). We advise future researchers to consider first converting the head rotation into another representation (e.g. quaternions) and then to apply interpolation techniques to fill in the rotation values for the lost frames.

Figure 2: Fitted face shape mask in Visage software.

6. Stimuli Quality as Measured by Participants in Previous Studies

This stimuli collection contains passages appropriate for use during a user study evaluating facial expressions in ASL animations. A subset of these passages and comprehension questions has already been used in prior studies at our laboratory (Kacorri, Lu, and Huenerfauth, 2013a; Kacorri, Harper, and Huenerfauth, 2013). The following stimuli in this collection were included in these two prior studies: E1, E2, E4, E5, E6, E8, W2, W3, W4, Y3, Y5, Y6, R3, R5, R7, N1, N2, N3, T3, T4, T5.

The first study consisted of a user study in which native ASL signers viewed human videos (with natural facial expressions) and ASL animations (without any facial expressions) and responded to comprehension questions (Kacorri, Lu, and Huenerfauth, 2013a). In the second study, identical stimuli were shown and similar participants were recruited, but in this study, the participants viewed the animations on a computer screen that was mounted above a desktop eye-tracking system that tracked their gaze location on the stimulus (Kacorri, Harper, and Huenerfauth, 2013). Full details of the studies appear in the original publications. Figure 3 presents the human video and the no-facial-expression animation results from these two studies. Bars are shown separately for each category of stimuli: emotional affect, negation, topic, WH-question, YN-question, and RH-question. Here we see that the stimuli with facial expressions received higher comprehension question scores than the stimuli without facial expressions, which suggests the suitability of these questions for user studies evaluating the perception of facial expressions. In future work, we intend to conduct more rigorous studies of the efficacy of these stimuli and questions, and we intend to examine the quality of the additional stimuli that were not included in these two prior studies. We also welcome feedback and improvements to the stimuli from other researchers who make use of this collection.

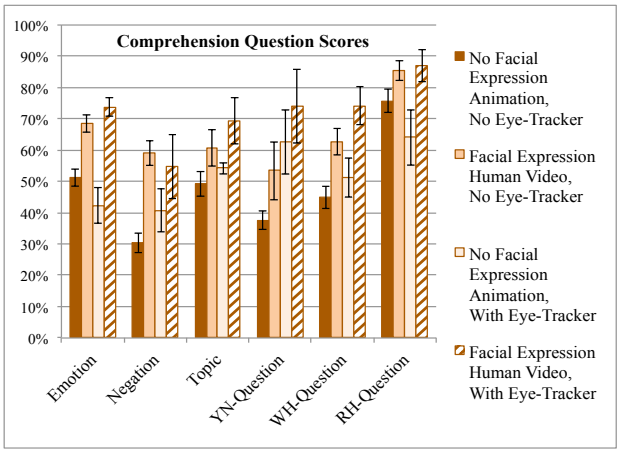


Figure 3: Comprehension question scores from the subset of stimuli in the collection used in prior evaluation studies.

7. Availability of the Collection

As with prior ASL corpora resources released by our laboratory (Lu and Huenerfauth, 2009; Lu and Huenerfauth, 2012), this stimuli collection is available for use by other sign language animation researchers, details appear here: http://latlab.cs.qc.cuny.edu/lrec2014

We invite members of the research community to provide feedback to us about the stimuli in this collection, and we welcome recommendations of additional stimuli designs or edits that would enhance the collection (which we would look forward to incorporating into a future release of this resource). While the current collection of stimuli has not yet been rigorously evaluated, we see a benefit for rapidly releasing this resource to the research community for use and feedback. Ultimately, the field of sign language animation synthesis may benefit from the community identifying a standard set of evaluation stimuli and questions for system evaluation, to better enable comparison of systems and progress in the field.

In future work at our laboratory, we are continuing to investigate the design of animation models for ASL facial expressions, and we are continuing to make use of these stimuli and questions to evaluate the quality of our animation results.

8. Acknowledgements

This material is based upon work supported in part by the US. National Science Foundation under award number 0746556 and 1065009. This work has also been supported by Visage Technologies AB through a free academic license for character animation software. Miriam Morrow, Jonathan Lamberton, and Jennifer Marfino assisted with the design of stimuli, and we are grateful for advice and feedback from Susan Fischer.

9. References

Gibet, S., Courty, N., Duarte, K., and Naour, T. L. 2011. The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. ACM Transactions on Interactive Intelligent Systems (TiiS), 1(1), 6.

Huenerfauth, M. 2004a. American Sign Language Spatial Representations for an Accessible User-Interface. In Proc. of the 3rd International Conference on Universal Access in Human-Computer Interaction. Las Vegas, NV, USA.

Huenerfauth, M. 2004b. Spatial and planning models of ASL classifier predicates for ma- chine translation. In 10th International Conference on Theoretical and Methodological Is-sues in Machine Translation: TMI.

Huenerfauth, M., Zhao, L., Gu, E., and Allbeck, J. 2007. Evaluating American Sign Language Generation through the Participation of Native ASL Signers. In Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2007), Tempe, Arizona, USA.

Huenerfauth, M. 2008. Evaluation of a psycholinguistically motivated timing model for animations of American Sign Language. In Proc. 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008).

Huenerfauth, M., Lu, P., and Rosenberg, A. 2011. Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. In Proc. 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2011). New York: ACM Press.

Kacorri, H. 2013. Models of linguistic facial expressions for American Sign Language animation. ACM SIGACCESS Accessibility and Computing, (105), 19-23.

Kacorri, H., Lu, P., and Huenerfauth, M. 2013a. Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation. ACM Trans. Access. Comput. 5, 2, Article 4.

Kacorri, H., Lu, P., and Huenerfauth, M. 2013b. Evaluating Facial Expressions in American Sign Language Animations for Accessible Online Information. In Proc. International Conference on Universal Access in Human Computer Interaction (UAHCI). Las Vegas, NV, USA.

Kacorri, H., Harper, A., and Huenerfauth, M. 2013. Comparing native signers' perception of American Sign Language animations and videos via eye tracking. In Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (p. 9). ACM.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D.N., and Neidle, C. 2013. Recognizing Eyebrow and Periodic Head Gestures Using CRFs for Non-Manual Grammatical Marker Detection in ASL. In Proceedings of FG 2013: 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai China.

Lu, P., and Huenerfauth, M. 2009. Accessible Motion-Capture Glove Calibration Protocol for Recording Sign Language Data from Deaf Subjects. In Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2009), Pittsburgh, Pennsylvania, USA.

Lu, P., and Huenerfauth, M. 2012. CUNY American Sign Language Motion-Capture Corpus: First Release. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Mitchell, R., Young, T., Bachleda, B., and Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. Sign Lang Studies, 6(3):306-335.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee, R. 2000. *The Syntax of American Sign Language: functional categories and hierarchical structure.* Cambridge, MA: MIT Press.

Pandzic, I. S., and Forchheimer, R. (eds.). 2003. MPEG-4 facial animation: the standard, implementation and applications. Wiley.

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., and Piater, J. 2013. Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models. Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT).

Schnepp, J., Wolfe, R., McDonald, J. C., Jorge Toro. 2012. Combining emotion and facial nonmanual signals in synthesized american sign language. In Proc. 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012), 249-250.

Tekalp, M. 1999. Face and 2-D Mesh Animation in MPEG-4, Tutorial Issue on the MPEG-4 Standard, Image Communication Journal, Elsevier.

Traxler, C. 2000. The Stanford achievement test, 9th edition: national norming and performance standards for deaf and hard-of-hearing students. J Deaf Stud & Deaf Educ, 5:4, pp. 337-348.

Visage Technologies. 2014. Visage Technologies Face Tracking. Retrieved February 4, 2014, from http://www.visagetechnologies.com/products/visagesdk/facetrack/.

Wolfe, R., Cook, P., McDonald, J. C., and Schnepp, J. 2011. Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. Sign Language & Linguistics, 14(1), 179-199.

Yu, X., Huang, J., Zhang, S., Yan, W., and Metaxas, D. N. 2013. Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model. In IEEE International Conference on Computer Vision.

How to Use Depth Sensors in Sign Language Corpus Recordings

Rekha Jayaprakash, Thomas Hanke
Institute of German Sign Language and Communication of the Deaf, University of Hamburg
{rekha.jayaprakash, thomas.hanke}@sign-lang.uni-hamburg.de

Abstract

We describe the experimental setup of positioning two depth sensors in the existing DGS corpus studio configuration. This includes investigation of the challenges of including depth sensors in the setup that already consists of other cameras. We also discuss about how these sensors can be helpful in automatic analysis of non-manuals like facial expression recognition for corpus recordings with our experimental configuration.

Keywords: Kinect, Carmine 1.09, DGS corpus, studio setting, non-manuals recognition

1. Introduction

Recently, combined camera and depth sensor devices caused substantial advances in Computer Vision directly applicable to automatic coding a signer's use of head movement, eye gaze, and to some extent, facial expression. Automatic and even semi-automatic annotation of non-manuals would mean dramatic savings on annotation time and are therefore of high interest for anyone working on sign language corpora.

Optimally, these devices need to be placed directly in front of the signer's face at a rather short distance. While this might be ok for some experimental setups, it is not acceptable in a corpus setting for at least two reasons: (i) The signer looks at the device instead of into the eyes of an interlocutor. (ii) The device is in the field of view of other cameras used to record the signer's manual and non-manual behaviour.

We report on experiments determining the degradation in performance when moving the devices away from their optimal positions in order to achieve a recording setup acceptable in a corpus context. For these experiments, we used two different device types (Kinect and Carmine 1.09) in combination with one mature CV software package specialised on face recognition (Faceshift).

2. Setup

The experiment is located in an existing studio configuration adapted from the DGS corpus recording setup (Hanke et al., 2010). The major change is that the signers are standing instead of sitting. For the first round of experiments, only one signer is prsent, signing into an HD camera at face level. For the time being, we ignore the visibility of the two sensors in the total scene camera perspective, but concentrate on its visibility in the frontal view. (The bird's eye view turned out not to pose a problem.) In this context, the signer is located facing the HD camera a distance of about 2.8 meters distance.

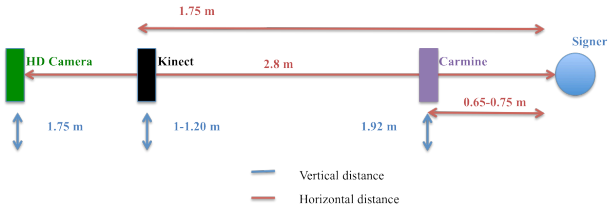


Figure 1: Diagrammatic representation of our final setup

Now we consider this length as our area of interest to position the sensors. In the experiment, a deaf colleague produced random signing, i.e. we did not make use of the monitors to provide elicitation materials which made things easier as we had to accommodate some improvised mounting devices (cf. fig. 2) for this experimental purpose. Finally we arrived at the configuration as shown in fig. 1 & 2.

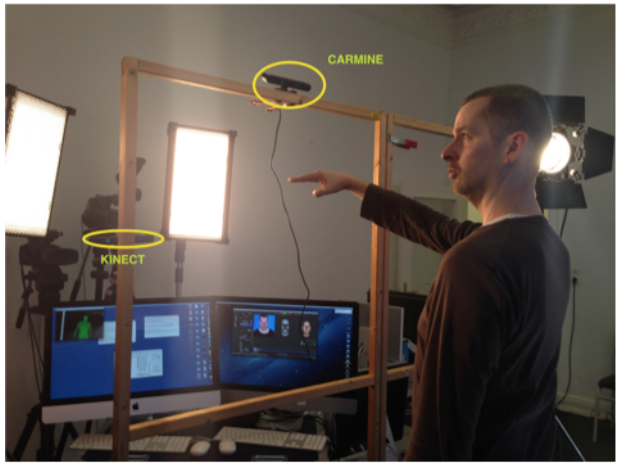


Figure 2: Setup with Carmine and Kinect sensor

3. Depth Sensor Positions

Placing these two depth sensors in the existing DGS corpus studio setup and arriving at the final optimal solution of the current studio setup are described in this section.

The Kinect is well known for its full body tracking capability if operated at a distance of more than half a meter. The Carmine 1.09 is a near mode depth sensor that can sense from less than half a meter, making it a good candidate for facial expression recognition software (e.g Faceshift). There are couple of important constraints to be considered about the performance range of the depth sensors and challenges to be resolved during recording.

Constraints:

1) The Carmine 1.09 can only be placed 0.65 meters maximum away from the face (front-facing) with permissible rotation along horizontal axis.

2) The Kinect should be placed between 1.5 to 1.75 meters away from the signer to get good skeleton

tracking.
Challenges:
1) The signer's eye gaze should not be distracted by the sensor.
2) These sensors can appear in other camera's fields of view.
Now we analyse different positions based on the combination of constraints and challenges within the length of interest i.e. between the signer and front facing HD camera.

3.1 Facial Expression Recognition

Faceshift is a facial motion capture software package that takes input from depth sensors like Carmine. Carmine 1.09 is a near-mode sensor recommended for Faceshift. Prior to the performance test, the system was trained to the signer's face for achieving a calibrated expression model for his/her face. For example, most common facial expressions like neutral, smile, frown etc., are considered as sample data for training and classification. The facial expression recognition highly relies on good training data of each individual signer. Another important point is that we are interested in estimating an optimal orientation of the Carmine device such that tracking and recognition are consistent and independent of different signers' physiognomies.

We tested performance of this setup varying the parameters distance (ranging from 0.35 to 0.8 meters) and rotation. As we can see from fig. 4, good lighting and the face close enough to the sensor result in a good accuracy of expression recognition, even with some rotation. By analyzing several test data we observed that the optimal distance is 0.70m. After resting the base of Carmine on a mounting surface (in our case a wooden frame and a stand), we rotated the head manually in 'yaw' direction in order to find best orientation. (The reason why there seems to be different lighting in the samples is sensor rotation.) The vertical field of view of the Carmine device is 45 degrees. It required a lot of trial and error experiments to adjust the sensor head rotation, based on the following rule: (1) Forehead and chin area should be visible prominently in the field of view (see fig. 3) as they are relevant clues in tracking the signer's face.

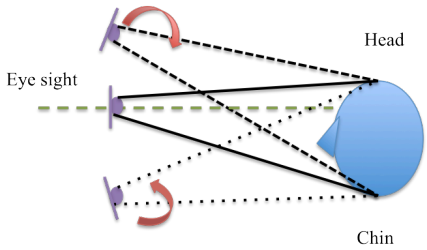


Figure 3: Sensor head rotation in 'yaw' direction

Once we had the optimal distance as well as rules for determining the best possible orientation and position of the sensor for one particular signer, the next step was to find a solution for achieving an optimal orientation of the sensor which is acceptable for signers of varying height. Of course one could adjust the height of the sensor but the setup should be tolerant enough not to require time-intensive calibration In order not to touch the sensor at all in our experiments, we simply asked shorter signers to stand on some pedestal-like boxes.
We also varied the sensor horizontally to the left and right of the signer's face within the range of 30 cm as shown in fig. 4. Within this range, the movement does not have an impact on face tracking (given a distance of 70cm).

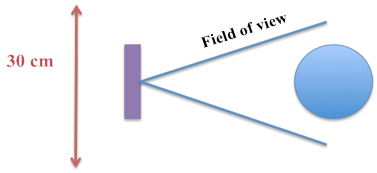


Figure 4: Sensor horizontal sliding

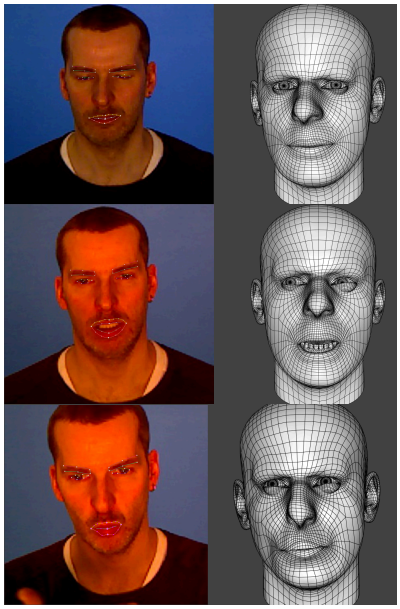


Figure 5: Sample data (1-3) showing accuracy of expression recognition based on table 1

Sample	Distance in meters	Performance
1	0.80	Unreliable
2	0.75	Fewer false positives
3	0.70	Better tracking

Table 1: Comparison of performance with varying distance

Figures 5, 6 and table 1 explain the dependency between the distance from the sensor, the orientation and the recognition quality. Considering this fact, we observed that the optimal height of mounting the Carmine is above the signer' head level with reasonable rotation, the upper option in fig. 3.

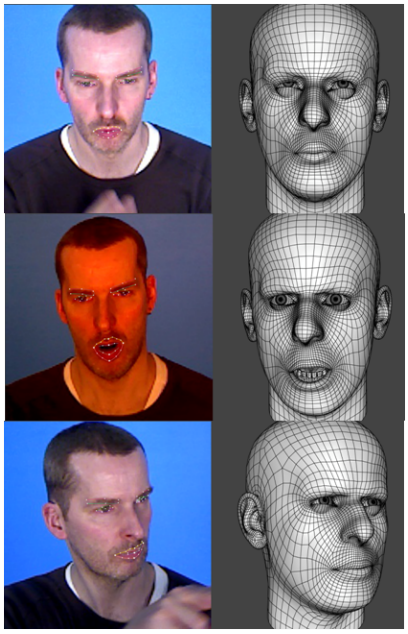


Figure 6: Sample (1-3) data showing sensor rotation

Sample	Rotation range in degrees	Performance
1	25 -30	Unreliable eye tracking
2	20-25	Good
3	10-20	Better tracking

Table 2: Comparison of tracking performance with varying rotation

Once the optimal rotation is set, the system gets trained for that particular signer's face and the recording begins. After initiating the recording, the Carmine sensor should not be rotated as that will result in inconsistent tracking for that signer. Prediction of rotation variation of the sensor head is not necessary. We have the option of extracting the head rotation in the 'yaw' direction. From the normal case (see fig. 3), i.e when the sensor is frontal to the signer's face, the rotation values can be even higher than this. But in our case where the sensor is not exactly facing the signer's face, but slightly from above. So we have restrictions to have minimum rotation as shown in Table 2. Another important issue is inconsistency in the lip movement recognition, which occurs due to head movement and tracking failures after occlusion of the lips. This issue was rectified to some extent in the refinement process. Fig. 5 shows the lip expression recognition of sample (2) and (3) from fig. 7. There is a possibility of analysing 48 built-in facial expressions in Faceshift.

After achieving a satisfying outcome from performance tests and height adjustment of the Carmine sensor, we filmed some sample data to check the visibility of the sensor in the front-facing HD camera. What we observed is that the Carmine sensor remains invisible in the field of view when the camera focuses on the signer's signing space below the head as shown in fig. 6. However, if the camera is set to also capture signs above the head, a little part of the sensor mounting became visible.

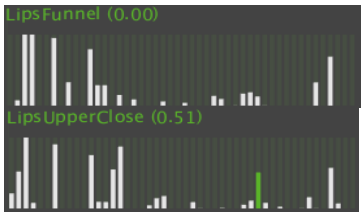


Figure 7: Lip expression accuracy (in green) of sample (2) and (3) from fig. 6

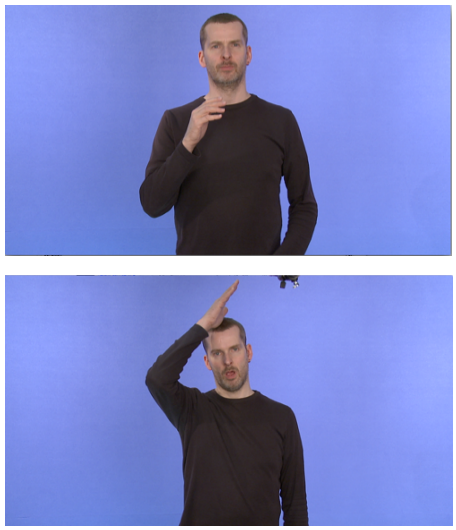


Figure 8: Visibility of Carmine from front facing HD camera when signing occurs below head and above head

Feedback from our transcription team suggests that a bit of appearance of Carmine will not disturb their further work with the film. For production videos, the mounting could later be removed automatically from the movie footage as long as there is no overlap of the signer's hands and the mounting device.

3.2 Kinect Positioning for Skeleton Tracking

The Kinect (xbox) is placed in front of the frontal HD camera as shown in fig. 1 and at a distance of 1.75 meters away from the signer and 1.0 meter above the ground. This is the final position where we could get satisfying results.

Before deciding the best position for the Kinect, we tried to explore the various pitfalls with different heights of placing the sensor as given in table 3. Since the motion is mainly happening in the upper part of the body (Torso), there were heights at which the tracking started collapsing by dropping too large an amount of frames initially. This is crucial because initial frame drops cannot be afforded in our case (for sign language corpus analysis later).

We show a couple of test cases to prove the dependency between distance nearer to the signer and skeleton tracking performance. We placed the Kinect sensor at:

a) Test case 1: a distance of 2.10 meters and a height of 1.70 meters to make sure it fits as close as possible to front facing HD camera. Tracking failed due to calibration failure.

b) Test case 2: a distance of 1.50 meters and a height of 1.40 meters. Tracking starts only after dropping frames, but it is unreliable. As you can see from fig. 9 (2) the green color of the tracker indicates second

user being detected in the scene, which is not true in our case.

Figure 9: Different test cases (1-3) of Kinect positions showing the colour frames and tracking performance

c) Test case 3: a distance of 1.75 meters and height of 1.00 meters. Tracking and calibration are good.

As shown in table 3, we also tested the tracking performance with different heights. The possibility of moving the Kinect away from the signer was restricted due to the space constraint in the current studio setup. In a more regular setup, there will be enough space to test different other positions and heights.

Kinect at a Distance of 1.75 meters away from the signer	Kinect's Height in meters	Tracking performance
	1.60	Calibration failed & no tracking
	1.30	Tracking started only in the middle of the film
	1.00	Tracks well

Table 3: Comparison of Skeleton tracking performance with varying height at fixed distance from the signer

4. Future work

When trying to apply the current approach to the studio setting with two informants (whether seated or standing), The current solution for the Carmine devices can simply be doubled. However, one degree of freedom for positioning the Kinect devices is lost: Following the results obtained so far, the only reasonable position for the Kinect devices is directly above the screens used for elicitation material in order to minimize distraction to the informants. The experiments carried out so far suggest that a setup like that shown in fig. 10 will be possible. Fig. 11 shows possible configurations how to place two Kinect devices (one for each signer) relative to each other in order to minimize the space needed. Another problem to be researched is synchronization issues involved in non-manuals recognition resulting from the use of two different sensors requiring different recording software.

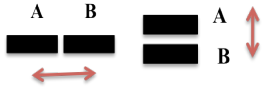


Figure 10: DGS corpus studio setup - Two signers interacting in sitting position, Kinect devices mounted between the two screens

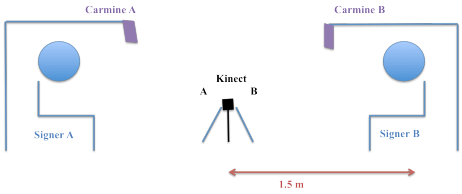


Figure 11: Possible configurations of placing two Kinect devices

5. Technical details

The depth sensors we use are Kinect xbox 360 for body tracking and Carmine 1.09 for facial feature tracking. These two sensors are operated using two different software packages. Data recording from Kinect xbox and Carmine are achieved by OpenNI & OpenCV program and Faceshift software at 640x480p30 respectively (for both depth and RGB channel). The recording with Kinect can be done automatically (continuously) or manually for each user.

6. Conclusion

Although the current studio setup has limited space to accommodate extra sensors (and their stands!), our additional sensors positions do not make the informants feel uncomfortable or the images more difficult to process by human annotators. The positioning of the sensors for the current corpus studio configuration increases our confidence that it will be possible to use these two depth sensors in corpus recordings resulting in valuable automatic annotation of non-manuals. As a by-product, we might be able to annotate emotional facial expressions.

7. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

8. References

Hanke, T., König, L., Wagner S., Matthes S., (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 106–109.

http://www.faceshift.com/product/

http://www.openni.org/

http://www.primesense.com/

Mouth-based non-manual coding schema used in the Auslan corpus: explanation, application and preliminary results

Trevor Johnston, Jane van Roekel
Macquarie University
Sydney, Australia
E-mail: trevor.johnston@mq.edu.au, jane.vanroekel@mq.edu.au

Abstract

We describe a corpus-based study of one type of non-manual in signed languages (SLs) — mouth actions. Our ultimate aim is to examine the distribution and characteristics of mouth actions in Auslan (Australian Sign Language) to gauge the degree of language-specific conventionalization of these forms. We divide mouth gestures into categories broadly based on Crasborn et al. (2008), but modified to accommodate our experiences with the Auslan data. All signs and all mouth actions are examined and the state of the mouth in each sign is assigned to one of three broad categories: (i) mouthings, (ii) mouth gestures, and (iii) no mouth action. Mouth actions that invariably occur while communicating in SLs have posed a number of questions for linguists: which are 'merely borrowings' from the relevant ambient spoken language (SpL)? which are gestural and shared with all of the members of the wider community in which signers find themselves? and which are conventionalized aspects of the grammar of some or all SLs? We believe these schema captures all the relevant information about mouth forms and their use and meaning in context to enable us to describe their function and degree of conventionality.

Keywords: sign language, corpus, ELAN, non-manuals, Auslan, Australian Sign Language

1. Introduction

The mouth is prominent site of non-manual activity and movements of the mouth are an obvious accompaniment to manual signing. The linguistic status of mouth actions in SLs, like other non-manuals, is a question of debate. There are two major types of mouth actions: those that are transparently complete or partial silent articulations of the spoken words of the ambient SpL (mouthings), and those that are not (mouth gestures). An early issue of interest was the amount of mouth actions, especially mouthings, that various SLs typically manifested. An area of debate concerned the status of mouthings: were they an integral part of a SL or were they marginal, stemming from language contact or borrowing? Addressing both these questions involved describing what mouthing and mouth gestures did and thus categorizing them into types.

2. Previous research

Research has shown that mouthings frequently accompany manual signs in many SLs and there is some evidence—though the datasets have never been very large or varied—that the rate varies according to text-type. They occur very frequently with fingerspellings and some signers appear to always mouth when fingerspelling (e.g., Sutton-Spence & Day 2001). Mouthings have been shown to occur more with nouns and plain verbs than with morphologically complex signs such as indicating verbs (also known as agreement and spatial verbs) or with depicting signs (also known as classifier signs). Mouthing has been shown to add meaning to some signs by indicating a more specific reading of a sign, e.g., the Auslan sign spouse with the English mouthings 'wife' or 'husband' (Johnston & Schembri 2007) which may or may not be considered a specification of the form of the sign (e.g., Schermer 2001). A mouthing can even add independent semantic information (e.g., Vogt-Svendsen 2001). Though mouthings are often closely temporally aligned with their co-articulated sign, they may be stretched, reduced, or repeated to maintain an alignment with the duration and rhythm of the manual sign, especially if the sign has itself been modified (Fontana 2008). Finally, the mouthing itself may spread regressively or progressively to adjacent signs (Crasborn, van der Kooij, Waters, Woll, & Mesch 2008).

Mouth gestures are all other communicative mouth actions that are not mouthings and they 'do not derive from spoken language' (Boyes-Braem & Sutton-Spence 2001).

For a study of spreading behaviour in mouthings in three SLs, Crasborn et al. (2008) called mouthings M-type mouth actions; adverbials mouth gestures were called A-type, and echoes became part of a slightly broader E-type category (for 'semantically empty' following Woll (2001)). Enactions were discriminated into two sub-types—those involving only the mouth in which the mouth represents itself doing the action described by the sign, such as bite or laugh or lick (these were call 4-type for 'mouth "for" mouth'): and W-type (for 'whole of face') in which mouth is simply part of a large whole of face expression, e.g., an open mouth with wide eyes for surprise. Crasborn et al. note that W-type mouth gestures are not specifically part of any mouth-based semiotic system because any mouth action form is linked to the whole face and consequently its interpretation is a function of the enaction and not just a function of (the conventional value of) the mouth form (Figure 1).

More recently, some researchers have focussed on the question of the gestural nature of some mouth actions in that they have the same kind of relationship to the conventional manual signs of a SL as do manual gestures to

the conventional spoken words of a SpL (Pizzuto 2003). Fontana (2008), drawing on the work of Kendon (2004, 2008) and McNeill (2000), makes the radical suggestion that all mouth actions can be analysed this way. A similar but more conservative observation is made in Dachkovsky and Sandler (2009) and Sandler (2009). They identify a category of gestural iconic mouth actions to distinguish them from syllabic E-type (or 'lexical') mouth components and from the conventional adverbial and adjectival A-type modifiers already identified by other SL researchers which they also accept (graphically represented in Figure1).

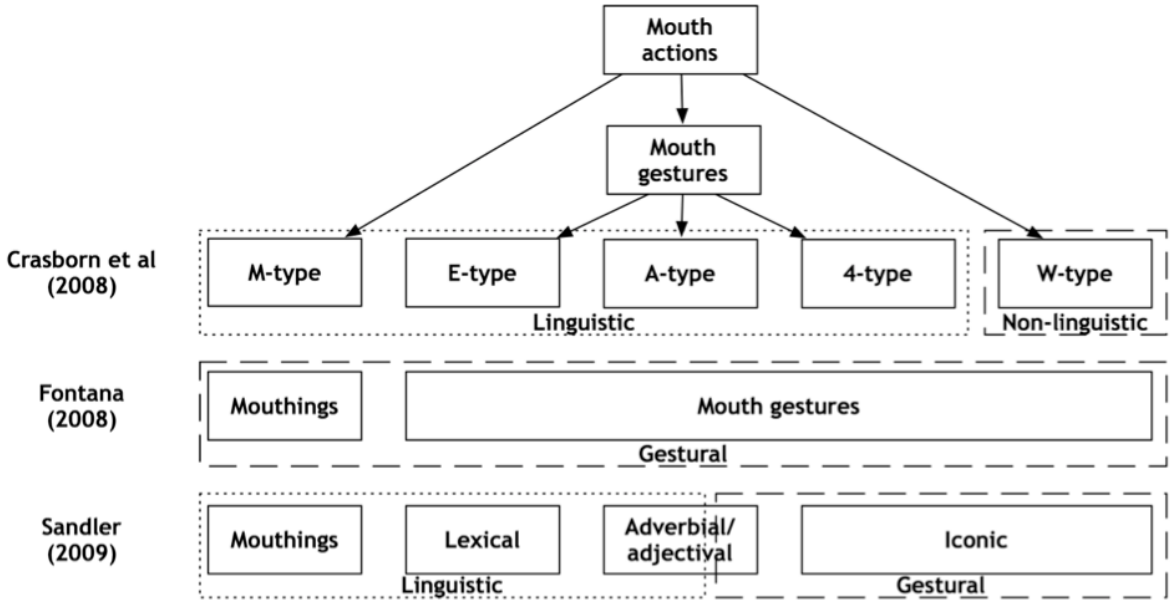


Figure 1 Three potential categorizations of mouth actions

3. This study

3.1 Methodology

Fifty video texts were selected from the Auslan corpus for analysis. All signs and all mouth actions were examined and all mouthings and mouth gestures were identified, categorized and annotated.

The data in this study has been drawn from the Auslan corpus of native or near-native signers (for further details see Johnston & Schembri 2006). For this study, 50 video clips were selected from the corpus, representing 38 individuals, 3 text types (monologue, dialogue, and elicited) during 5 hours and 58 minutes of the corpus, representing 16,920 manual sign tokens. The signed texts ranged from 1:32 to 38:30 minutes in duration. The 50 video clips consisted of 25 monologues (narratives of which there were 25 retellings of two Aesop's fables); 10 dialogic texts (free conversation or responses to a series of interview questions); and 15 sessions of 40 elicited picture descriptions.

3.2 Annotation schema

The 50 texts were chosen from a subset of the 459 texts that had previously been given at minimum a basic annotation (i.e., they had been glossed and translated) using ELAN multi-media annotation software according to guidelines detailed in the Auslan Corpus Annotation Guidelines[1] and Johnston (2010). All signs and all mouth actions were examined and the state of the mouth in each sign was assigned to one of three broad categories: (i) mouthings, (ii) mouth gestures (both of which we have already briefly characterized), and (iii) no mouth action. Mouth gestures were divided into types that were based on Crasborn et al. (2008) but additional sub-groupings were (temporarily) created to accommodate finer distinctions we felt salient in the Auslan data (Figure 2). We are prepared to further adapt or even abandon these categories if needs be after considering the first annotation implementation, aggregation of data, and analysis.

These new sub-categories are: **prosodic** = a tensed posture of the mouth that is held for a period of time, even if relatively briefly, without changing dynamically rather than any specific mouth posture as such; **spontaneous** = involuntary or spontaneous expressions (indexes almost) of the state of the mind of the signer (e.g., amused, confused, concerned); **editorial** = expressions as meta-comments about what the signer is signing that do not intentionally modify the manual signs; **constructed actions** = full enactments that involve all of the face; **congruent** = a default expression that match the semantics of the lexical sign, such as smiling while signing happy; **adverbial expressive** = clearly intend to modify and add meaning to the manual sign(s) but they are not limited to the mouth and they are also strongly enacting (Figure 2).

[1] Downloadable from www.auslan.org.au

Annotations were added to two ID-glossing and two grammatical class tiers (one for each hand of the signer), and four tiers for information on mouth actions. Annotations for mouth gestures were made on the 'mouth gesture form' tier (called MouthGestF), and on the 'mouth gesture meaning' tier (called MouthGestM). The annotations for mouthings were made on the 'mouthing form' tier (called Mouthing) (Figure 3).

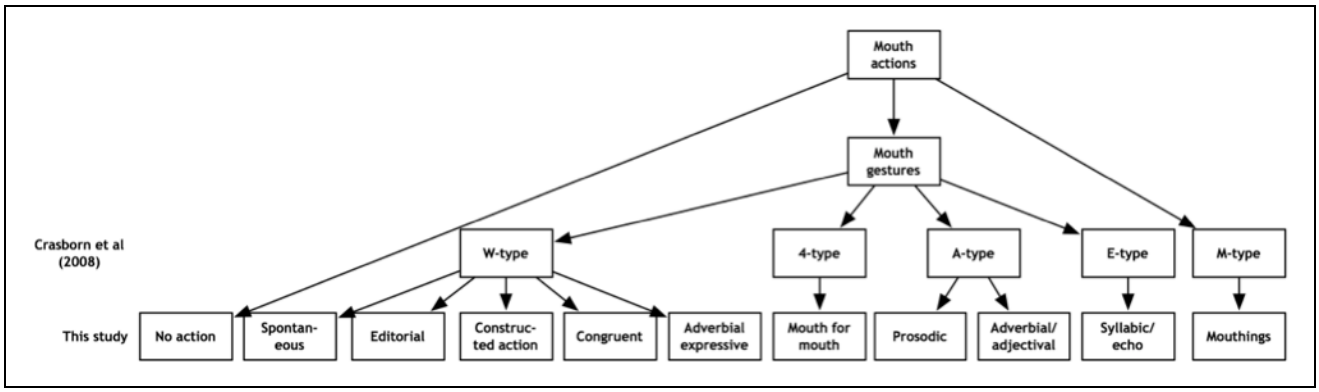


Figure 2 Types of mouth actions annotated in this study

Figure 3 The relevant tiers [file: SSNc2a00:00:29.000]

Mouth action	MouthGestF begins with	MouthGestM tier contains
No mouth action	no annotation	no annotation
M-type (mouthing)	English word (Table 2)	no annotation
Mouth gesture types		
E-type (echo or empty)	syllable:gloss	various meanings as needed
A-type (modifying)		
intonational	gloss (Table 3)	meaning: e.g., *activity, emphasis*
adverbial	gloss (Table 3)	meaning: e.g., *large-amount, careless, unpleasant*
4-type (mouth for mouth)	CongruentMouth Only	*enactment*
W-type (whole-of-face)		
spontaneous	no annotation	no annotation
editorial	comment	no annotation or various meanings as needed
constructed action	ConstructedAction	no annotation or various meanings or descriptions as needed
	ConstructedAction:gloss (Table 3)	the gloss for an A-type mouth gesture
congruent	CongruentWhole Face	*expression, enactment, emphasis*
adverbial expressive	ConstructedAction:adv	*expression*

Table 1 The annotation schema for mouth actions

The annotation schema is summarized in Table 1. It is important to note that each type of mouth gesture has a unique MouthGestF annotation or a unique combination of values on both the MouthGestF and MouthGestM tiers. This enables various constraints to be applied with an ELAN search routine to identify and quantify mouth gestures of a certain type or sub-type only. For example, one may perform an ELAN multi-tier conditional search to extract statistics on, say, intonational mouth gestures because they will be the only ones that have the annotations "activity" or "emphasis" on the MouthGestM tier associated with any of the form glosses listed in Table 2 on the MouthGestF tier. For example, if "emphasis" occurs on a MouthGestM tier which occurs with a "CongruentWholeFace" annotation on the MouthGestF tier, it is an instance of a congruent W-type mouth gesture, not

an intonational one. Each combination of values in Table 1 is unique to a sub-category of mouth gestures.

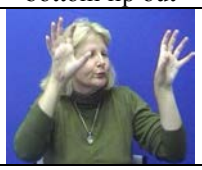
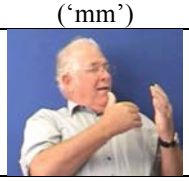
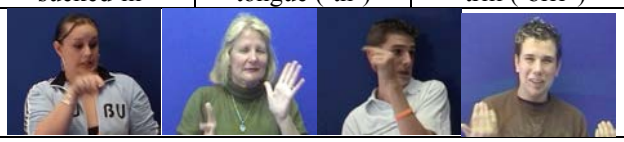
		
blow	bottom lip out	down
		
lip-curl	lips-out	lips-pressed (‘mm’)
		
open	puff	slightly-open
		
sucked-in	tongue (‘th’)	trill (‘brrr’)
		
wide (‘ee’)		

Table 2 Examples of glosses for mouth gesture forms

Annotating the alignment of mouth actions with manual signs Where a mouth action clearly spreads across two or more signs, it is marked separately on each one and the subsequent annotations are suffixed with -prog (progressive) or –regress (regressive). Only apparently significant spreading is annotated in this dataset—the exact onset and offset time of mouth gestures is not a focus of this study though it has been in other studies (Sandler 1999; Crasborn et al. 2008)

Annotating mouthings Mouthings are often incomplete or partial. We found that it was important to annotate which part or parts of the associate English word were mouthed (Table 3). This was partly based on the realization that competent but non-fluent or non-native signers can easily mistaken a partial mouthing for a mouth gesture when annotating the data thus resulting in an overall inflation of mouth gesture counts.

Degree of articulation	**Representation**	**Examples**
Complete articulation	complete	race, rabbit, village, far
Initial segment	i(nitial)	v(illage), sa(me), diff(erent), sh(eep)
Medial segment	(me)di(al)	(no)th(ing), (re)mem(ber) , (b)e(st)
Final segment	(fi)nal	(success)ful, (fin)ish, (im)prove. (to)day
Initial & final segment only	in(i)tial	f(ini)sh, d(ea)f, s(uc)cesful
Suppressed unreadable	Suppressed unreadable	(lady), (have)

Table 3 The annotation schema for mouthings

In extracting mouthing counts from the data the parentheses may be removed, or left, depending on the type of analysis desired. For example, for a straightforward count of the distribution of English words mouthed in the corpus—and their association with particular sign tokens—the parentheses would be removed. (This can be removed from the entire corpus by a multi-file (domain) search and replace on the appropriate Mouthing tier; alternatively, one may remove them after the exported annotations are opened in a database program.) For pattern comparison of form and meaning with selected mouth gestures, the material enclosed in parentheses would be deleted. The remaining vowels, consonant clusters or syllables can then be compared to the semantics of the source sign and the semantics of similar-looking mouth gestures that occur with other manual signs.

3.3 Preliminary results

The results presented here are taken from the first iteration of the implementation of this annotation schema. They are not definitive. They are merely indicative of the type of information that can be easily extracted from the corpus given these annotations. The annotations were partly motivated by the type of functions available in ELAN for searching, filtering and exporting annotations within that program.

The key functions used in ELAN included: multi-file multi-tier searches using explicit values or regular expressions; automatically generated statistical profiles across multiple annotation files (domains); multi-file (domain) processing, in particular, EXPORT MULTIPLE FILES AS > ANNOTATION OVERLAPS INFORMATION. The latter were exported as tab delimited files into Excel for further processing.

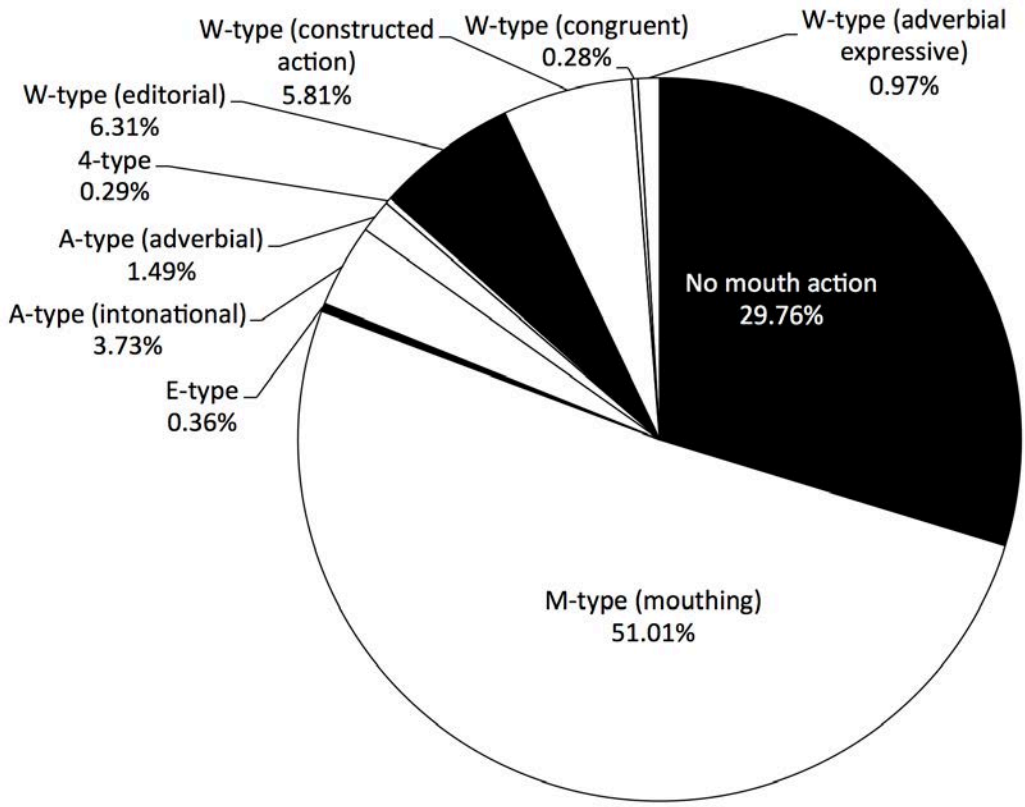


Figure 4 Overall distribution of mouth actions

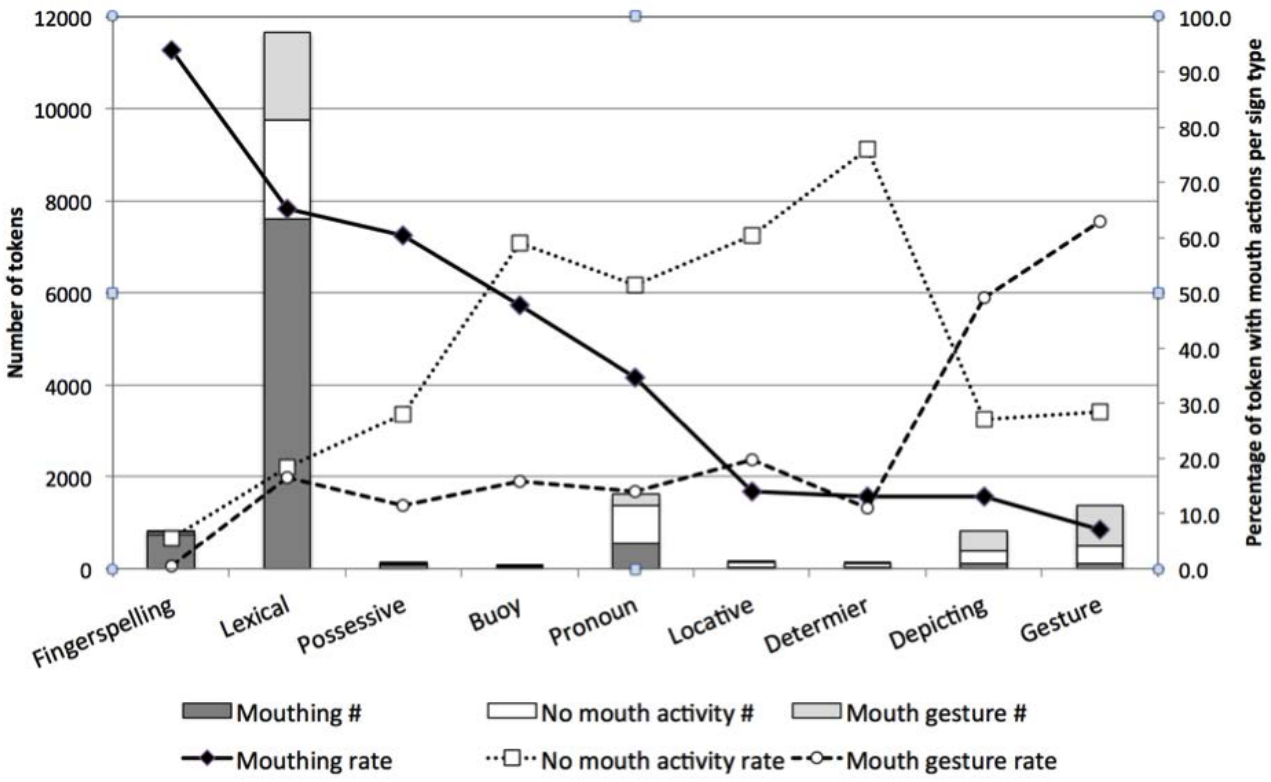


Figure 5 Mouth actions by **sign type**

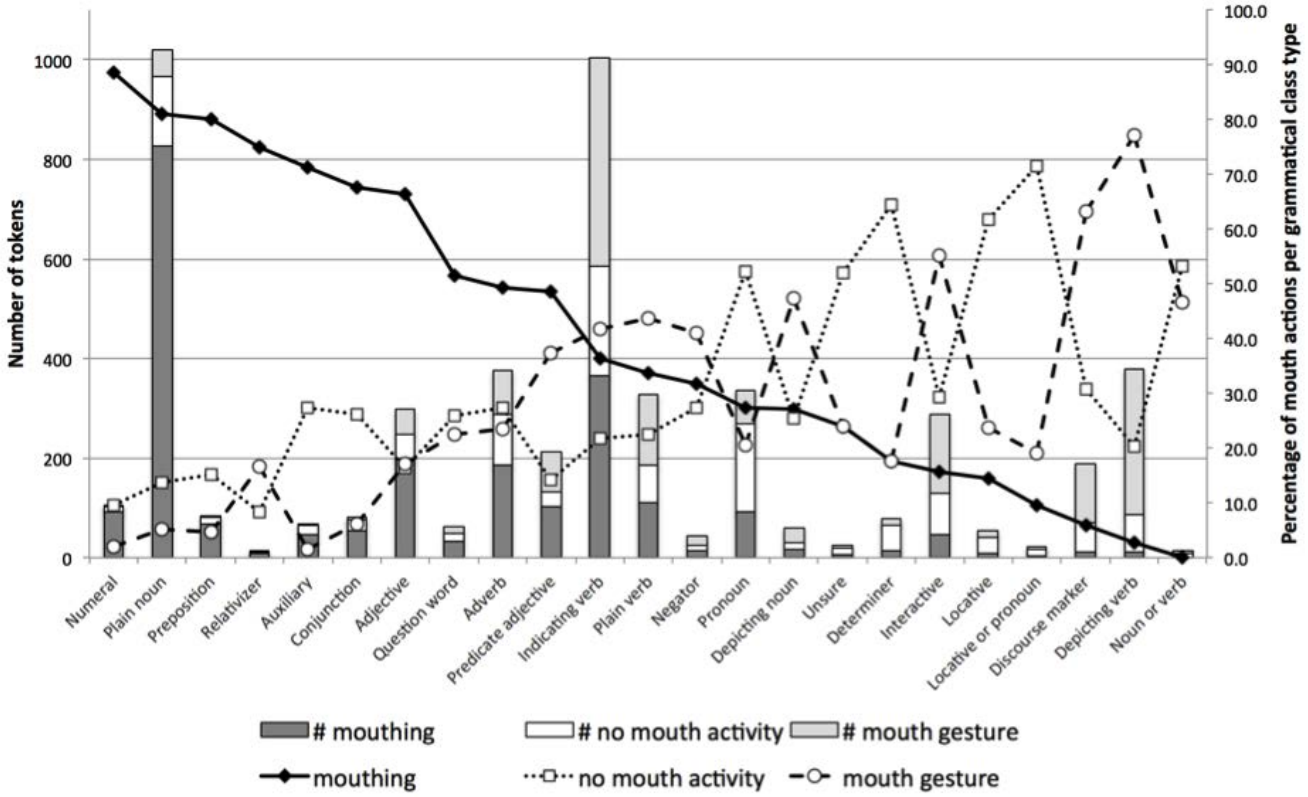


Figure 6 Mouth action by **grammatical class**

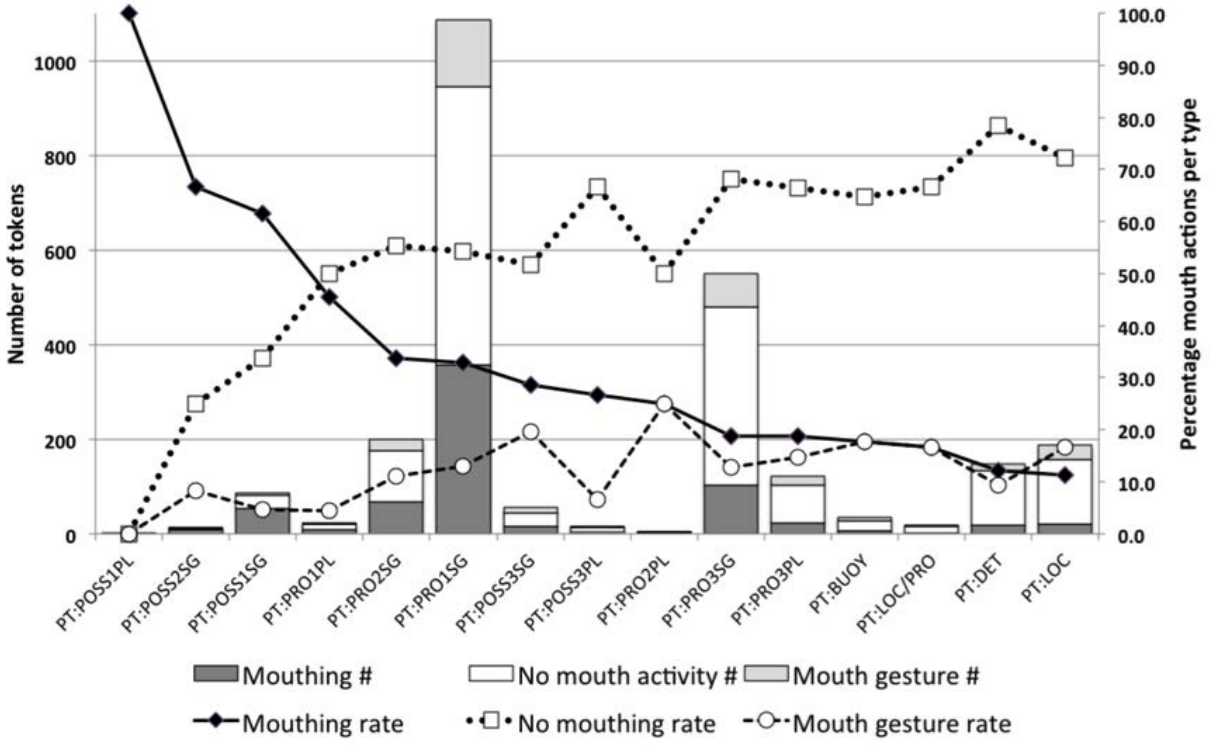


Figure 7 Mouth actions and **pointing sign**

Lexical frequency The rates of mouthing for lexical frequency suggest that lexical frequency, as such, has marginal impact on mouthing rates. However, it did emerge that the highest ranking lexical signs have a smaller the range of English words mouthed with each sign (and this is unsurprisingly often the same word that has been adopted for the ID-gloss in the corpus).

3.3.1 Characteristics of types of mouth actions

E-type (semantically empty mouth gestures) Syllabic mouth gestures were very rare in the data. Only 66 signs

and their mouth actions fell into this category. Several signs occurred with different mouth gestures, often with the same effect (Table 4).

The first observation to make is that at 11 potential forms, the number of different syllabic mouth patterns in this dataset is actually quite small. A second important observation is that syllabic mouth gestures in Auslan do have flexible but consistent meanings across a range of signs, they are not exactly semantically empty as such as suggested in the SL literature.

Mouth gesture form	Mouth gesture meaning	Tokens
PAH	SUDDEN	31
AP	EMPHASIS	12
(L/B)AM	DISAPPEAR	8
WOOF	EMPHASIS	4
PAH-PAH	EMPHASIS	2
POOH	REMOVE	2
(L/B)AM	EMPHASIS	2
POW	EMPHASIS	2
BOOM	EMPHASIS	1
AM	EMPHASIS	1
ALARM	EMPHASIS	1
Total		**66**

Table 4 E-types by meaning and token count

MG form (tokens)	MG meaning gloss	Meanings in more detail
TONGUE (111)	CARELESS	carelessly, easily, without regard, petulantly, with deliberate careless enjoyment, reckless, slipshod, insouciant
LIPS-OUT (16)	EASE	easily, without regard, petulantly, with enjoyment
TRILL (13)	LARGE AMOUNT	large amount, a lot of, unimpeded, energetic, powerful, engine/machine-powered
BLOW (10)	SMOOTH	smooth, unimpeded, quickly, ongoing
TONGUE (7)	UNPLEASANT	unpleasant, distasteful, bad
TRILL (3)	EASE	easily, unimpeded, with enjoyment
LIPS-PRESSED (2)	EASE	easily but deliberately, enjoyable
BOT-TOM-LIP-OUT (2)	CARELESS	careless, easily, without regard, petulantly, with deliberate careless enjoyment, reckless, slipshod, insouciant
Total (165)		

Table 5 Form/meaning pairings for A-type

A-type (adverbials) the majority are actually intonational in character, according to our definitions. In effect, only a very small number of the total number of signs in the dataset represent putative dedicated conventional A-type adverbial mouth gestures. Only a very small set of recurring semantic descriptors needed to capture the apparent contribution of these mouth gestures. The data suggests that the semantic component of the mouth gesture is quite broad (Table 5).

Only two broad meaning labels appeared necessary to capture the effect of **intonational mouth gestures: emphasis and activity**. Emphasis was the broader default reading (71%), and wide was the most preferred mouth gesture with this force (21%); activity accounted for the force of the remaining 29%, of which 48% were achieved with wide. Overall, wide accounted for almost a third of all intonational mouth gestures.

W-type (whole of face) 82% are of the **constructed action** sub-type. There are very few of the other types. Indeed, constructed actions represent 45% of all mouth gesture types (i.e., mouth actions excluding mouthings).

4-type (mouth for mouth) Token count was extremely low (N = 68). The token frequencies are unremarkably linked to the narratives chosen for the re-tells or the elicitation materials, e.g., GRAZE, YELL, CAPTURE, EAT, AMERINDIAN, SPEECH, LAUGH, SHOUT, CHEW, ANGRY, etc.

Time alignment of mouth actions with manual signs (spreading and 'holding') Only mouthings have been thus far annotated and processed for spreading activity. There are approximately 305 spreading mouthings (245 progressive and 60 regressive). Importantly, they are strongly associated with pointing signs (PT) with approximately 50% spread to PTs (both progressively and regressively). There are >100 cases in which mouthing articulation proper spans only one sign but the mouth shape is held progressively (for the duration of the following sign). Once again, approximately 50% involves a following PT sign.

M-type (mouthings) Most, but not all, mouthings were completely and clearly articulated (over 95%). A small number of mouthing tokens (approximately >30) were not accompanied by any manual sign yet the were clearly not redundant. They provided essential disambiguating or logico-cohesive information to the utterance. They were conjunctions, prepositions or adverbial like but, or, for, just, maybe; sentence modifiers like *I-don't-know*, *I-think*; or other interactives like *no*, *yes*, *not-true*.

Degree of articulation	Tokens
Complete articulation	8911
Initial segment	262
Medial segment	13
Final segment	26
Initial & final segment only	23
Suppressed articulation	6
Unreadable	64

Table 6 Type of mouthing

3.3.2. Variation

Individual variation in mouth action rates, text type variation, and sociolinguistic variation are all also very important for understanding the role of mouth actions in SLs. However, these results are not reported here as these deal with wider questions on the function and interpretation of the role of mouth actions in SLs rather than the annotation and classification of mouth actions and other non-manuals.

4. Acknowledgments

Australian Research Council (ARC) grant #DP1094572. Research assistants and annotators who have contributed to the Auslan corpus in approximate chronological order since 2005 (most recent first): Ben Hatchard, Michael Gray, Gabrielle Hodge, Lindsay Ferrara, Julia Allen, Gerry Shearim, Karin Banna, Louise de Beuzeville, Dani Fried, Della Goswell, Adam Schembri.

5. References

Boyes-Braem, P., & Sutton-Spence, R. (Eds.). (2001). *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg, Germany: Signum Press.

Crasborn, O., van der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics, 11*(1), 45-67.

Dachkovsky, S., & Sandler, W. (2009). Visual Intonation in the Prosody of a Sign Language. *Language & Speech, 52*(2–3), 287-314.

Fontana, S. (2008). Mouth actions as gesture in sign language. *Gesture, 8*(1), 104-123.

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics, 15*(1), 104-129. DOI: 110.1075/ijcl.1015.1071.1005joh.

Johnston, T., & Schembri, A. (2007). *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge: Cambridge University Press.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kendon, A. (2008). Some reflections on the relationship between 'gesture' and 'sign'. *Gesture, 8*(3), 348-366.

McNeill, D. (Ed.). (2000). *Language and gesture*. Cambridge: Cambridge University Press.

Pizzuto, E. (2003). Review of Boyes-Braem, P., & Sutton-Spence, R. (Eds.). (2001). The hands are the head of the mouth: The mouth as articulator in sign languages. Hamburg, Germany: Signum Press. *Sign Language and Linguistics, 6*(2), 300-305.

Sandler, W. (1999). Prosody in Two Natural Language Modalities. *Language and Speech, 42*(2-3), 127-142.

Sandler, W. (2009). Symbiotic symbolization by hand and mouth in sign language. *Semiotica, 174*, 241-275.

Schermer, T. (2001). The Role of Mouthings in Sign Language of the Netherlands: Some Implications for the Proudction f Sign language Dictionaries. In P. Boyes-Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 66-87). Hamburg: Signum Press.

Sutton-Spence, R., & Day, L. (2001). Mouthings and mouth gestures in British Sign Language (BSL). In P. Boyes-Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp. 66-87). Hamburg: Signum Press.

Vogt-Svendsen, M. (2001). A comparison of mouth gestures and mouthings in Norwegian Sign Language (NSL). In P. Boyes-Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages* (pp.??). Hamburg: Signum Press.

Woll, B. (2001). The sign that dares to speak its name: Echo phonology in British Sign Language (BSL). In P. Boyes-Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg: Signum Press.

Weakly Supervised Automatic Transcription of Mouthings for Gloss-Based Sign Language Corpora

Oscar Koller[1,2], Hermann Ney[1] and Richard Bowden[2]

[1] Human Language Technology and Pattern Recognition - RWTH Aachen University, Germany
[2] Centre for Vision Speech and Signal Processing - University of Surrey, Guildford, UK
{koller,ney}@cs.rwth-aachen.de, r.bowden@surrey.ac.uk

Abstract

In this work we propose a method to automatically annotate mouthings in sign language corpora, requiring no more than a simple gloss annotation and a source of weak supervision, such as automatic speech transcripts. For a long time, research on automatic recognition of sign language has focused on the manual components. However, a full understanding of sign language is not possible without exploring its remaining parameters. Mouthings provide important information to disambiguate homophones with respect to the manuals. Nevertheless most corpora for pattern recognition purposes are lacking any mouthing annotations. To our knowledge no previous work exists that automatically annotates mouthings in the context of sign language. Our method produces a frame error rate of 39% for a single signer on the alignment task.

Keywords: Sign Language, Mouthing, Lip Reading, Unsupervised Automatic Annotation

1. Introduction

Sign languages consist of several information streams that convey meaning. Historically, research on automatic recognition of sign language has focused on the manual components of the signs, such as the hand shape, its orientation, position and movement (Starner et al., 1998; Vogler and Metaxas, 2004; Zaki and Shaheen, 2011). These manual parameters are widely considered to contain a large part of the information in sign language. However, it is clear that a full understanding of sign language, particularly with respect to idioms, grammatical structures and also semantics, is not possible without further exploring the remaining information channels, namely facial expressions (mouthing, eye gaze) and upper body posture (head nods/shakes and shoulder orientation). Mouthing can be observed in many European sign languages. Nevertheless, its linguistic status is still debated (Sandler, 2006). However, there is a lot of evidence that mouthings can discriminate homophones with respect to the manual parameters and thus constitute an important feature for automatic recognition of sign language, which has not been exploited in current approaches. This is due to the fact that sign language corpora intended for pattern recognition and machine learning usually do not have any mouthing annotations.

This work aims to automatically annotate mouthings for gloss-based sign language corpora when annotations are not available. The employed corpus is recorded from broadcast news and constitutes a translation from German speech to sign language performed by hearing interpreters. We use the automatic transcriptions of the speech and exploit this as weak supervision through the fact that mouthings in sign language often correspond to parts of orally pronounced words.

In Section 2. related work in viseme recognition and linguistics is shown. In Section 3. we present the corpus and the manual annotation used for evaluation. Section 4. presents the approach. Finally, results are given in Section 5. and Section 6. draws conclusions with future work.

2. Related Work

Two types of mouth actions can be observed in sign languages: mouthings and mouth gestures. While mouthings are silently pronounced parts of spoken words that originate from speech contact, mouth gestures constitute patterns unrelated to spoken language. Mouthings occur often with nouns and with morphologically simple signs (Crasborn et al., 2008). Furthermore, they are often related to lexical items (Sutton-Spence, 2007), while mouth gestures have a morphological role (Horst Ebbinghaus and Jens Hessmann, 2001). The status of mouthings in sign language is highly debated in the linguistic community. Some researchers understand it as part of sign language, while others see it as separate entity. Refer to (Crasborn et al., 2008) for details on this debate. However, in German Sign Language (DGS) mouthings play an important role. DGS contains many signs with identical manual parameters that have related meanings and seem to be only disambiguated by combination with different, though semantically related, mouthings (Horst Ebbinghaus and Jens Hessmann, 1994; Kutscher, 2010). In terms of synthesis, (Kipp et al., 2011) have analysed the perception of sign language avatar systems and found that the absence of mouthings strongly disturbs the Deaf evaluators. Movement of cheeks and lips, but also teeth and tongue were determined crucial for understanding certain mouthings.

In this paper, we deal with signing of sign language interpreters. The question arises, if their mouthings differ from native Deaf mouthings. However, not much literature has systematically researched this question. (Weisenberg, 2009) found that sign language interpreters adjust their mouthing with respect to their target audience. However the study only evaluates four interpreters and the influence of Deaf family members is neglected. In a study comparing three native and two non-native signer, (Lisa Monschein, 2011) reports that the non-native (hearing) signers do not use more mouthings than the native Deaf signers.

Visemes, the visual representations of phonemes in the mouth area, were first mentioned by (Fisher, 1968). Nowa-

days lipreading and viseme recognition is a well established, yet challenging research field in the context of audio-visual speech recognition. The first system was reported by (Petajan, 1984) who distinguished letters from the alphabet and numbers from zero to nine and achieved 20% error rate on that task. Since then, the field has advanced in terms of recognition vocabulary, features and modelling approaches. (Ong and Bowden, 2011) achieved an error rate of 13.2% using sequential patterns for lipreading. A good overview is given in (Potamianos et al., 2003). Previous applications of viseme recognition specifically to automatic sign language recognition are very rare. The state of mouth openness has been used to distinguish signing from silence (Pfister et al., 2013). However, little work has been done in training viseme models in an unsupervised or weakly supervised fashion. Most deal with the problem of clustering visemes in order to find an optimal phoneme to viseme mapping (Luca Cappelletta and Naomi Harte, 2012) and to our knowledge no previous application of dedicated viseme recognition to sign language recognition exists.

3. Corpora

The proposed approach uses the publicly available RWTH-PHOENIX-Weather corpus, containing continuous signing in DGS of 7 hearing interpreters. The corpus consists of 190 TV broadcasts (weather forecast) recorded on public TV. It provides a total of 2137 manual sentence segmentations and 14717 gloss annotations. Glosses constitute an economical way of annotating sign language corpora. They represent an approximate semantic description of a sign, usually annotated w.r.t. the manual components. The same gloss 'MOUNTAIN' denotes the sign alps but also any other mountain, as they share the same hand configuration and differ only in mouthing. Moreover, the RWTH-PHOENIX-Weather corpus contains 22604 automatically transcribed and manually corrected German speech word transcriptions. The boundaries of the signing sentences are matched to the speech sentences. It is worth noting that the sentence structures for spoken German and DGS do not correlate. This is a translation rather than a transcript. Furthermore, it has to be noted that the corpus contains signing of professional hearing interpreters. Some have Deaf family members and grew up with sign language as mother tongue, others did not. As the interpreters translate live, they face very tight time constraints. Due to the direct interpretation task, it can be expected that the interpreter's mouthings are partly closer to speech, than they usually would be. Nevertheless, this remains to be proven.

To evaluate this work, we annotated 3 sentences per signer on the frame level with viseme labels totalling 2082 labelled frames. The annotation was performed three times by a competent non-native signer. While annotating, the annotator had access to the video sequence of signing interpreters showing their whole body (not just the mouth), the gloss annotations and the German speech transcriptions. In each of the three annotation iterations, the frame labels varied slightly due to the the complexity and ambiguity of labelling visemes. See (Yuxuan Lan et al., 2012) for a human evaluation. We consider each annotation to be valid, yielding 1.6 labels per frame (see Table 4).

4. Weakly Supervised Mouthing Alignment

The approach exploits the fact that mouthings are related to spoken language and its words, for which automatic spoken language transcripts are part of the RWTH-PHOENIX-Weather corpus. However, the relation between speech and mouthings is loose and holds for some signs only.

Visual features of the mouth region are extracted. These consist of ten continuous distance measurements around the signers mouth and the average colour intensity of three areas inside the mouth (to capture tongue and teeth presence), as shown in Fig 1. The distance measurements are based on salient point locations on the interpreter's face tracked using the deformable model registration method known as Active-Appearance-Models (AAMs). For details refer to (Schmidt et al., 2013).

The features are clustered using Gaussian clustering and Expectation Maximization (EM) while constraining the sequence of features to the sequence of automatically transcribed German words in a Hidden-Markov-Model (HMM) framework. Thus, we consider the weakly supervised viseme training to be a search problem of finding the sequence of visemes $v_1^Z := v_1, \ldots, v_Z$ belonging to a sequence of mouthings (or silently pronounced partial words) $m_1^N := m_1, \ldots, m_N$, where the sequence of features $x_1^T := x_1, \ldots, x_T$ best matches the viseme models. We maximise the posterior probability $p(v_1^N|x_1^T)$ over all possible viseme sequences for the given sequence of glosses.

$$x_1^T \rightarrow \hat{v}_1^Z(x_1^T) = \arg\max_{v_1^Z} \left\{ p(m_1^N) p(x_1^T|v_1^Z) \right\}, \quad (1)$$

where $p(m_1^N)$ denotes the pronunciation probability for a chosen mouthing. We model each viseme by a 3 state HMM and a garbage model having a single state. The emission probability of a HMM state is represented by a single Gaussian density with a diagonal covariance matrix. The HMM states have a strict left to right structure. Global transition probabilities are used for the visemes. The garbage or 'no-mouthing' model has independent transition probabilities. We initialise the viseme models by linearly partitioning the data.

The given word sequence that stems from the Automatic Speech Recognition (ASR) transcripts is reordered to better match the syntax present in DGS. This is done by aligning the manual gloss annotations and the speech transcripts with the GIZA++ toolkit (Och and Ney, 2003) commonly used in statistical machine translation to align source and target language. Furthermore, a lexicon is built that includes a finite set of possible pronunciations for each German word. This lexicon consists of different phoneme sequences for each word and an entry for 'no-mouthing'. However, the mouthings produced by signers often do not constitute fully pronounced words, but rather discriminative bits of words. Thus, for each full pronunciation we add multiple shorter pronunciations to our lexicon ψ by truncating the word w which consists of a sequence of phonemes $s_1^N = s_1, \ldots, s_N$, such that $\psi = \left\{ w' : s_1^{N-\phi} | \phi \in \{0, \ldots, \phi_{trunc}\} \wedge N - \phi \geq \phi_{min} \right\}$, where we empirically set $\phi_{trunc} = 10$ and $\phi_{min} = 3$.

Finally, to account for the difference in articulatory phonemes and visual visemes, we need to map phonemes

	∑	A	E	F	I	L	O	Q	P	S	U	T	gb	ratio
Signer 1	275	13	25	8	27	8	30	28	19	19	18	54	143	1.43
Signer 2	266	25	40	24	18	8	27	29	18	25	16	58	147	1.64
Signer 3	318	35	18	23	51	15	39	70	34	21	16	83	185	1.86
Signer 4	236	43	35	38	27	8	12	33	14	20	15	46	63	1.50
Signer 5	320	36	32	23	56	8	19	48	22	14	39	44	103	1.39
Signer 6	366	65	39	38	28	6	44	43	28	12	36	98	191	1.72
Signer 7	301	28	21	23	56	18	40	42	32	2	14	79	136	1.63
∑	2082	11.8	10.1	8.5	12.7	3.4	10.1	14.1	8.0	5.3	7.4	22.2	46.5	1.60
ratio	1.60	1.76	1.80	1.78	1.99	2.04	1.79	1.90	1.75	1.88	1.77	1.90	1.43	

Table 1: Frame annotation statistics for each of the employed 11 visemes (Eeva A. Elliott, 2013) on the RWTH-PHOENIX-Weather corpus. The last line shows relative annotation per viseme in [%]. 'gb' denotes frames labelled as non-mouthings/garbage. 'ratio' refers to the average labels per frame, which reflect the uncertainty of the annotator.

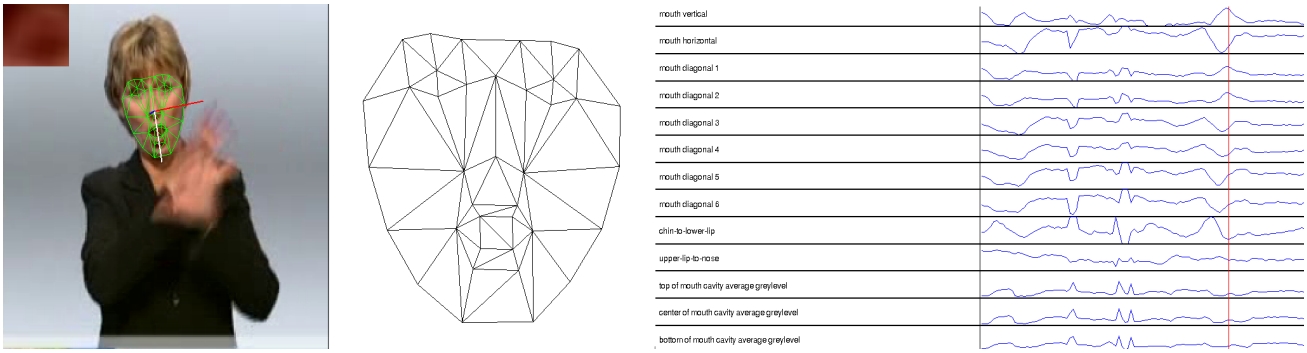


Figure 1: Feature extraction, left: fitted AAM grid and inner mouth cavity patch, center: rotated and normalised AAM grid, right: high-level feature values over time

to visemes. Two different mappings are compared in this work. A mapping to 16 visemes by (Weiss and Aschenberner, 2005) (compare left side of Table 2) and a mapping to 12 visemes by (Eeva A. Elliott, 2013) (see right part of Table 2). Furthermore, we propose a mapping ourselves that considers a many-to-many relationship depending on the context of a viseme, i.e. the preceding and succeeding viseme. The mapping consists of 29 visemes and one 'no-mouthing' entry and is displayed in Table 3. It has been created using a phonetic decision trees (Beulen, 1999). All visemes are clustered based on their feature representation, while considering visual properties (roundness or openness).

5. Results

In the scope of this paper we provide a solution to automatically annotate mouthings in sign language corpora with not more than gloss annotations and speech transcripts as source of weak supervision given. With this in mind, we perform a forced alignment on the RWTH-PHOENIX-Weather data using different phoneme-to-viseme mappings to assess how suitable each is for the task of modelling sign language mouthings. See Figures 3 and 4 for qualitative examples of some alignments on our data.

We can determine the alignment error per frame based on the 2082 manually annotated frames (see Section 3.) for each of the seven signers. We compare the case of not using any viseme mapping and modelling 40 phonemes of the spoken language instead ('Phonemes'), a viseme mapping with 16 visemes by (Weiss and Aschenberner, 2005)

(Weiss and Aschenberner, 2005)		(Eeva A. Elliott, 2013)	
Visemes	Phonemes	Visemes	Phonemes
A	a a~ a:	A	a a~ a:
C	j C	E	e: E E:
E	i: I e: E: E	F	f v
F	f v	I	i: I j
M	m	L	l
N	n l	O	2: 9 o: O
O	o: O	P	b m p
P	p b	Q	6 C g h k N @ R x
Q	@ 6	S	S tS
R	h r x N	T	d n s ts t z
S	s z	U	u: U y: Y
T	t d k g	A I	aI
U	u: U	A U	aU
Y	y: Y 2: 9	O I	OY
Z	S tS	P F	pf
A E	aI		
A U	aU		
O E	OY		
P F	pf		

Table 2: Tested phoneme to viseme mappings in SAMPA.

('Weiss'), a mapping with 12 units by (Eeva A. Elliott, 2013) ('Elliott') and our proposed many-to-many viseme mapping with 30 context dependent visemes ('Proposed'). Results are given in Table 4, with the frame error rate per signer and averaged across the 7 signers given. It has to be noted that depending on the number of visemes, a cer-

	Visemes	SAMPA Phonemes	Context	
			Left	Right
Open/Round	A_1	a a~ a:	#	l
	A_2	a a~ a: aI		
	aU	aU		
	L_1	l		#
	L_2	l		
Semi-Open/Round	S_1	S	#	
	S_2	S tS		
	O	OY 2: O o:		
	U	U u: y:		
	Y	Y	t k	# s 6
	@	@		
Tense	E	e: E i I		
	F	f v pf		not #
Semi-Closed	LT_1	d l t	y:	#
	LT_2	b d l t	y:	not #
	LT_3	R l t ts	f	e: E i I
	CON_1	R d g h k l n t ts z	#	e: E i I
Strong Context	CON_2	6 C R b d f g k m n p s t ts v x z		#
	CON_3	6 N R f g k l m n s t v x z	U u: aU	
	CON_4	6 ts k n t		f v
	$CON_{5\text{-}11}$	different consonants+context		
Closed	M_1	b m p	#	
	M_2	b m p		

Table 3: Proposed many-to-many phoneme to viseme mapping depending on context. '#' refers to word boundaries.

tain error rate can be achieved by guessing a frame's label. In order to appropriately compare the mappings with different numbers of viseme models we define another error rate ('compensated ER') that removes all correct classifications achieved by chance. On average, over all signers 'Elliott' outperforms 'Weiss', which outperforms 'Phonemes' (56.84% to 60.21% to 74.16% respectively). Our proposed mapping lags 3% behind with 59.49%. However, if we consider the 'compensated ER' our proposed mapping outperforms all others by between 4% and 17%. Apart from the averaged results, we note that the alignment error rates differ among all signers. This can be explained by the fact that each signer's mouthing differs slightly. It manifests itself in different sets of preferred visemes by each signer, whereas not all visemes can be equally well modelled. Table 5 shows the alignment statistics of the whole data set using the 'Elliott' viseme mapping. Relative frame alignments per viseme are reported for all 180000 frames present in the data set. This allows us to observe the signers' mouthing preferences. As such, Signer 1 pronounces 'A' and 'O' more frequently than average. Our models represent these two visemes particularly well, which might explain why the viseme alignments for Signer 1 perform better than on other signers. The last line in Table 5 shows empirically determined occurrence frequencies reported in (Eeva A. Elliott, 2013) for reference. We see that 'T' and 'Q' are as reported (as well as in our paper) the two most frequently occurring visemes. The same similarity holds for the least frequently occurring viseme 'S'. On average our method aligns 44% of all frames to no-mouthings, which have been excluded

	'Phonemes'	'Weiss'	'Elliott'	'Proposed'
Signer 1	74.45	39.66	39.09	49.57
Signer 2	77.25	59.87	57.96	63.8
Signer 3	82.80	76.82	69.14	68.39
Signer 4	61.17	54.29	40.66	40.74
Signer 5	73.83	58.43	55.68	56.6
Signer 6	71.51	63.05	62.12	68.65
Signer 7	74.17	62.96	60.26	60.51
Total	74.16	60.21	56.84	59.49
#visemes	40	16	12	30
chance ER	96.12	90.36	87.5	94.18
compensated ER	78.03	69.85	82.87	65.31

Table 4: Frame error rates (ER) per signer in [%] for no viseme mapping ('phonemes'), a mapping by (Weiss and Aschenberner, 2005) ('Weiss'), (Eeva A. Elliott, 2013) ('Elliott') and our proposed mapping ('Proposed'). Lower is better.

in the linguistic reference. All viseme alignments seem to roughly correspond to the linguistic reference, however, we note that viseme 'Q' is only aligned 15.50%, whereas Elliott reports over 25%.

Figure 2 shows the top 15 glosses with the most frequently aligned mouthings in the corpus. We see that sensible mouthings have been chosen by our proposed weakly supervised alignment scheme. Furthermore it is shown that the approach is able to spot the different mouthings that specify signs with the same manuals and thus the same gloss annotation but with different mouthings. For example, the gloss REGEN (RAIN) has been found to occur with mouthings /R e g/ (rain) and /S aU 6/ (shower). Moreover, it is apparent that the weak supervision allows to spot mouthings that only share a semantic relation to the employed gloss, but actually constitute different words. Such an example is the mouthing /g R a t/ (degree) belonging to the gloss TEMPERATUR (TEMPERATURE), which represents an information stemming from the audio transcripts.

By showing the most commonly aligned mouthings, Figure 2 also contains information about pronunciation reductions. The type of reduction that we allowed (see Section 4.), was truncating the ends of the pronunciations. We see that apparently shorter pronunciations are preferred, as most of the most frequently aligned viseme sequences in the red bars consist of only 3 visemes (e.g./R e g/, /m O 6/, /n a x/). This coincides with the expectation that mouthings in sign language are more context cues than full silently pronounced words. Among the displayed mouthings there is only one that is very unlikely to actually have occurred (AUCH: /d a b/), which most likely constitutes noise injected during the statistical reordering process. In terms of word types, the result follows linguistic findings that mouthings mainly occur with nouns, as 13 of the 15 glosses are nouns.

6. Conclusions

In this paper we show how to automatically annotate mouthings in sign language corpora with no more than gloss annotations needed and speech transcripts as source of weak supervision. We further compare the impact of

	frames	A	E	F	I	L	O	Q	P	S	U	T	gb
Signer 1	49753	6.07	4.74	3.76	5.74	1.97	5.11	9.33	3.09	0.97	2.83	12.38	44.01
Signer 2	7399	6.27	3.24	2.34	3.89	1.50	4.05	7.95	4.58	1.20	5.84	11.83	47.30
Signer 3	27381	3.42	6.32	3.82	6.52	1.50	4.24	8.21	3.76	1.41	3.02	13.02	44.77
Signer 4	33394	4.60	4.91	3.04	3.94	1.34	4.22	6.45	3.08	0.92	2.34	8.64	56.50
Signer 5	41845	5.34	7.99	4.68	7.10	2.54	4.70	10.42	4.57	1.13	4.84	14.72	31.97
Signer 6	9841	4.88	3.94	4.16	4.89	2.93	4.55	9.14	4.89	1.06	8.45	11.77	39.36
Signer 7	19750	4.38	2.62	3.89	4.81	2.14	4.66	7.30	4.85	1.00	3.44	9.40	51.51
$\sum$	189363	5.04	5.39	3.82	5.62	1.97	4.62	8.63	3.85	1.08	3.69	11.96	44.33
$\sum$ no gb	105414	9.05	9.69	6.87	10.10	3.53	8.30	15.50	6.91	1.93	6.63	21.49	-
comparison	-	8.57	5.05	4.59	8.18	4.97	3.83	25.66	6.79	2.60	5.31	24.39	-

Table 5: Frame alignment statistics in [%] for each of the employed 11 visemes on the RWTH-PHOENIX-Weather corpus.'gb' denotes non-mouthings/garbage. The last line shows comparative statistics from (Eeva A. Elliott, 2013).

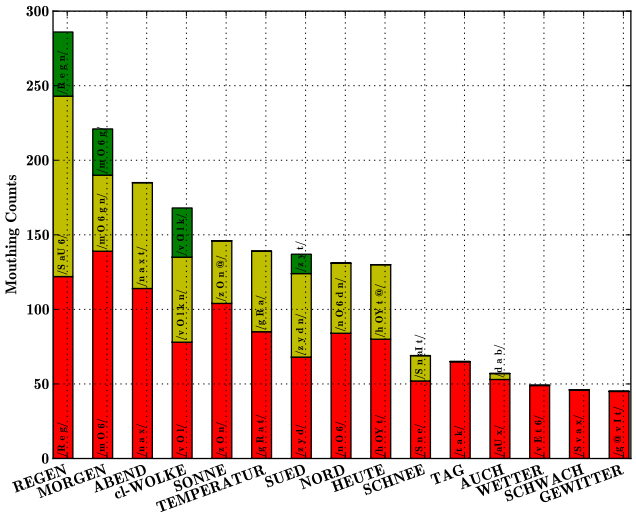


Figure 2: Top 15 glosses with the most frequent occurring mouthings shown in SAMPA annotation on the bars. Any mouthings occurring less than 20% w.r.t. all mouthings of a gloss have been filtered out for better readability.

four different schemes to map phonemes to visemes and find that a many to many mapping that relies on visemic context is best if one takes into account the complexity of the classification.

We achieve a frame error rate of 39.09% in the alignment task for a specific signer and 56.84% averaged over all signers. Furthermore, we show that our proposed method yields alignment statistics comparable to those in the linguistic literature. Finally, the mouthings are shown to further disambiguate gloss transcriptions of a sign. As expected, the mouthings represent reduced forms of German words.

In terms of future work, we plan to apply our method to native Deaf signing to separate influence from the German to DGS interpretation task and to include it into a sign language recognition pipeline. Furthermore, there is a need to find features that better represent tongue and inner mouth and modelling of mouth gestures remains untouched.

7. References

Klaus Beulen. 1999. *Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular*. Mainz.

Onno Crasborn, Els Van Der Kooij, Dafydd Waters, Bencie Woll, and Johanna Mesch. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1).

Eeva A. Elliott. 2013. *Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language*. Thesis, Freie Universität, Berlin, Germany.

Cletus G. Fisher. 1968. Confusions among visually perceived consonants. *Journal of Speech, Language and Hearing Research*, 11(4):796.

Horst Ebbinghaus and Jens Hessmann. 1994. German words in german sign language: Do they tell us something new about sign languages? In Carol Erting, editor, *The Deaf Way: Perspectives from the International Conference on Deaf Culture*. Gallaudet University Press.

Horst Ebbinghaus and Jens Hessmann. 2001. Sign language as multidimensional communication - or: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem and R. L. Sutton-Spence, editors, *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*, pages 133–153. Signum Press, Hamburg.

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of ACM Conference on Computers and Accessibility*, ASSETS '11, page 107–114, New York, NY, USA. ACM.

Silvia Kutscher. 2010. Ikonizität und indexikalität im gebärdensprachlichen lexikon – zur typologie sprachlicher zeichen. *Zeitschrift für Sprachwissenschaft*, 29(1), January.

Lisa Monschein. 2011. Empirical research on mouth patterns considering sociolinguistic factors: A comparison between the use of mouth patterns of deaf l1- and hearing l2-users of german sign language (DGS), August.

Luca Cappelletta and Naomi Harte. 2012. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM*, page 322–329.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Eng-Jon Ong and Richard Bowden. 2011. Learning sequential patterns for lipreading. In *Proceedings of the British Machine Vision Conference*, page 55.1–55.10. BMVA Press.

Eric David Petajan. 1984. *Automatic Lipreading to En-*

Figure 3: Example showing the frame alignment of signer 1 following the phoneme to viseme mapping from (Eeva A. Elliott, 2013). Original gloss annotation: PLUS EINS BIS SIEBEN TEMPERATUR. Audio transcription: Und das ganze dann bei plus ein und sieben Grad.

Figure 4: Example showing the frame alignment of signer 5 following the phoneme to viseme mapping from (Eeva A. Elliott, 2013). Original gloss annotation: MONTAG WAHRSCHEINLICH SONNE cl-WOLKE. Audio transcription: Am Montag mal Sonne, mal Wolken.

hance Speech Recognition (Speech Reading). Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA. AAI8502266.

Tomas Pfister, James Charles, and Andrew Zisserman. 2013. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the British machine vision conference*, U. K. Leeds.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, September.

Wendy Sandler. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press, February.

Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Enhancing gloss-based corpora with facial features using active appearance models. In *International Symposium on Sign Language Translation and Avatar Technology*, volume 2, Chicago, IL, USA.

Thad Starner, Joshua Weaver, and Alex Pentland. 1998. Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375.

Rachel Sutton-Spence. 2007. Mouthings and simultaneity in british sign language. In Myriam Vermeerbergen, Lorraine Leeson, and Onno Alex Crasborn, editors, *Simultaneity in Signed Languages: Form and Function*, page 147. John Benjamins Publishing.

Christian Vogler and Dimitris Metaxas. 2004. Handshapes and movements: Multiple-channel ASL recognition. In *Lecture Notes in Computer Science*, page 247–258. Springer.

Julia Weisenberg. 2009. *Audience effects in American Sign Language interpretation*. Ph.D. thesis, State University of New York at Stony Brook.

Christian Weiss and Bianca Aschenberner. 2005. A german viseme-set for automatic transcription of input text used for audio-visual speech synthesis. In *Proc. Interspeech*, pages 2945–2948, Lisbon, Portugal.

Yuxuan Lan, Richard Harvey, and Barry-John Theobald. 2012. Insights into machine lip reading. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4825–4828, March.

Mahmoud M. Zaki and Samir I. Shaheen. 2011. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572–577.

Discourse-Based Annotation of Relative Clause Constructions in Turkish Sign Language (TID): A Case Study

Okan Kubus
Universität Hamburg
Hamburg, Germany
E-mail: okankubus@gmail.com

Abstract

The functions of relative clause constructions (RCC) should be ideally analyzed at the discourse level, since the occurrence of RCCs can be explained by looking at interlocutors' use of grammatical and intonational means (cf. Fox and Thompson, 1990). To date, RCCs in sign language have been analyzed at the syntactic level with a special focus on cross-linguistic comparisons (see e.g. Pfau and Steinbach, 2005; Branchini and Donati, 2009). However, to our knowledge, there is no systematic corpus-based analysis of RCCs in sign languages so far. Since the elements of RCCs are mostly non-manual markers, it is often unclear how to capture and tag these elements together with the functions of RCCs. This question is discussed in light of corpus-based data from Turkish Sign Language. Following Biber et al. (2007), the corpus-based analysis of RCCs in TID follows the "top-down" approach. In spite of modality-specific issues, the steps in the process of annotation and identification of RCCs in TID fairly resemble this approach. The advantage of using these multiple steps is that the procedure not only captures the discourse functions of the RCCs but also identifies different strategies for creating RCCs based on linguistic forms.

Keywords: relative clause constructions, Turkish Sign Language, prosody, non-manual elements

1. Introduction

The first study on RCCs as regarding sign languages was the Liddell (1978) study on ASL. Analyses on RCCs in German Sign Language (DGS) (Pfau and Steinbach, 2005) and in Italian Sign Language (LIS) (Branchini and Donati, 2009 among others) have also been put forward. Analysis of the variation among sign languages by Perniss et al. (2007) indicates that there may be non-manual markings on RCCs in common over these three sign languages, i.e. raised eyebrows. However, the aforementioned researchers emphasize that the syntactic contributions do not necessarily have to be the same: the manual markers can vary. For example, Pfau and Steinbach (2005) show that RCCs in DGS might have unique syntactic properties as compared to RCCs in the other sign languages that have been studied so far.

Indeed, the functions of relative clause constructions (RCC) should be ideally analyzed at the discourse level, since the occurrence of RCCs can be explained by looking at interlocutors' use of grammatical and intonational means (cf. Fox and Thompson, 1990). To date, RCCs in sign language have been analyzed at the syntactic level with a special focus on cross-linguistic comparisons (see e.g. Pfau and Steinbach, 2005; Branchini and Donati, 2009). However, to our knowledge, there is no systematic corpus-based analysis investigating discourse functions of RCCs in sign languages to date.

At the same time, corpus-based sign language studies have been conducted mostly at the lexical or morpho-syntactic levels. For example, at the lexical level, Johnston (2013) investigated pointing signs using corpus data in Auslan. Bank et al. (2013) describe mouthing and mouth gestures in NGT using various tiers including mouth (Dutch word that is mouthed), mouth type (mouthing or mouth gesture), mouth lemma (dictionary version of lemma) and mouth spreading (progressive or regressive spreading occurrences). At the morpho-syntactic level, Branchini et al. (2013) have discussed WH-duplication patterns in LIS by looking at occurrences of WH-signs in the LIS corpus. This paper aims towards a different approach: How it is possible to look at the bigger picture to identify a specific linguistic unit and its interconnection throughout a text through a corpus study.

Biber et al. (2007) state that corpus linguistic studies are in fact a type of discourse analysis because they cover the investigation of the functions of the linguistic forms within a particular context. Specifically, Biber et al. (2007) state corpus linguistic studies are generally considered to be a type of discourse analysis because they describe the use linguistic forms in context (p. 2). According to Biber et al., corpus studies take one of two perspectives: (i) looking at the distribution and functions of surface linguistic features and (ii) investigating the internal organization of texts. The researchers point out that corpus studies have, surprisingly, not attempted to combine these two perspectives. This study is an attempt to combine these perspectives, notwithstanding the confronted difficulties.

Following in the steps of Biber et al, the corpus-based analysis of RCCs in TİD follows the so-called "top-down" approach. In spite of issues specific to modality, there is an urgent need to develop a similar approach to investigate RCCs in sign languages. The advantage of using such an approach is that the procedure not only captures the discourse functions of RCCs but also identifies different strategies for creating RCCs based on their linguistic forms. Non-manual elements that have no independent linguistic function should be ideally covered by the "top-down" approach. This paper provides the details of these steps. The advantage of using these multiple steps is that the

procedure not only captures the discourse functions of the RCCs but also identifies different strategies for creating RCCs based on their linguistic forms.

2. Corpus study for RCCs in TİD

The data collection for the ongoing dissertation project (Kubus in progress) was conducted in two ways: (i) data obtained via elicitation and (ii) video clips shared publicly (in internet). The aim is here to obtain naturalistic, spontaneous data collected for the purpose of observing the nature of relativization.

Data elicitation (retelling stories) was conducted with three TİD signers (one native, two near-native signers). However, the data collected provided nine potential relative clauses. The amount of relative clauses thus fell short of expectations for the systematic analysis of RCCs. Obviously, there was a clear need for more relative clause samples in order to examine a wider variety of relative clause strategies that would allow for generalizations. Therefore, in addition to data obtained via elicitation, sixteen video clips, covering a wider range of potential RCCs, were selected for the annotation. The video clips are predominantly monologues signed by eleven participants (six female and five male). The entire data collection comprises of a total of twenty-one video clips consisting of approximately 3 hours of film. The sign language corpus on Turkish Signs is annotated using iLex ("integrated Lexicon"; Hanke, 2002). An annotation sample is given in Figure 1.

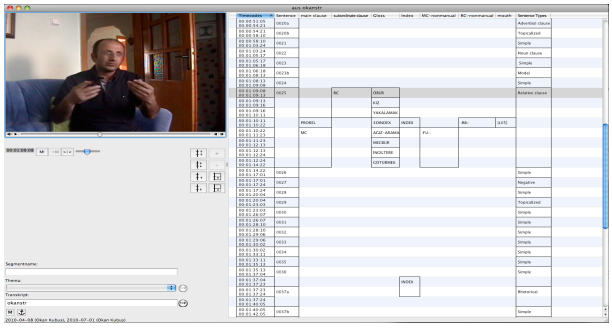


Figure 1: Data annotation (iLex)

The small-scale corpus in the ongoing dissertation project includes thirteen tiers (Table 1). Only one tier, labelled "chunks", is a structure tier and the tier "token" is a token tier. The other tokens are text tokens. Only the "chunk type" tier is subordinated under the "chunks" tier[1].

[1] Hanke and Storz (2010; p. 65) list different types of tiers. In the following, I present the three tiers, which are most often used in the ongoing dissertation project: (i) *token* tier, (ii) *structure* tier, and (iii) *text* tier.

Label	Function
Chunks	ID of each chunk
MC	The boundaries of matrix clause
RC	The boundaries of relative clause
Token	Glosses of both main clauses MC and subordinate clauses RC
Index	Marking index or other relative elements
NMM-MC	Non-manual markers for matrix clause (general)
NMM-RC1	Non-manual markers for relative clause part 1 (head movements)
NMM-RC2	Non-manual markers for relative clause part 2 (eyebrow)
NMM-RC3	Non-manual markers for relative clause part 3 (squint)
Mouth	Mouthings/ Mouth gestures specifying RC
Chunk Type	List of sentence types (e.g. declarative, interrogative, etc.)
Tr	Turkish translation equivalents of relative clauses
Eng	English translation equivalents of relative clauses

Table 1: The list of the tiers

2.1 "Top-Down" approaches in corpora study

Corpus linguistics covers various approaches with various goals for linguistic and especially discourse analyses (cf. Conrad, 2002). Conrad summarizes four corpus linguistics approaches for discourse analyses in spoken languages: (i) *Investigating characteristics associated with the use of a language feature* (p. 78), (ii) *Examining the realization of a particular function of language* (p. 81), (iii) *Characterizing a variety of languages* (p. 83) and (iv) *Mapping the occurrence of a language feature through a text* (p. 84). In the next paragraphs, each approach is described, and an argument is provided as to whether such an approach suits the current study.

According to Conrad (2002), the first approach is much more focused on a language feature, for example a word or a phrase or else a grammatical structure. In the ongoing dissertation project it is obvious that it is sought for RCC. However, due to the modality-specific properties, it turns out to be quite challenging to seek for a possible RCC in a specified corpus, since there is no previous research on this topic. Furthermore, there are no clearly spell-out words or phrases that can specify or hint such constructions. Rather, RCC seems to rely mostly on prosodic constituents of the sign language.

The second approach focuses on *a function of language and determines how it is realized in discourse* (Conrad, 2002; p. 81). For example, Biber et al. (1998) have investigated six characteristics: register, pronoun vs.

noun forms, given vs. new information status, type of reference, type of expression for anaphoric reference and the distance relationships among the characteristics. One of the findings in the ongoing dissertation project was that the type of referring expression and given/new information status relied on each other (as cited in Conrad, 2002). The present study follows this approach more by investigating RCCs and their functions in TİD. However, the challenge regarding sign language corpora which is mentioned in the previous paragraph persists. How this issue can be resolved will be explained in the next sections with the steps that are followed in the study.

In the third approach, the primary focus is *the language variety* (ibid, p. 83). For instance, Biber (1988) has developed a methodology called "multi-dimensional (MD) analysis" which includes a big scale of corpora with an automated analysis of linguistic features in more than two variables: for instance, various texts, text types, styles and/or registers (see also Biber, 1993). In this approach, multivariate statistical techniques are essential. In the ongoing dissertation project, three main discourse modes (cf. Smith, 2003) are investigated. However, since the primary focus is on RCC, it seems difficult to follow this approach with one linguistic feature variable in three different conditions. The quantity of the data and its uneven distribution over three modes makes it difficult to conduct statistical analyses. Rather, proportional (descriptive) and qualitative analysis are emphasized here.

The last approach is ... *one or more features are tracked through an entire text to determine how the features contribute to some aspect of the discourse development, such as its rhetorical organization*... (Conrad, 2002; p. 84). Indeed, this approach is closer to the approach in the ongoing dissertation project, with an exception: I am only focusing on RCC in TİD, and not on other linguistic elements. Such an approach is often related to the "top-down" approach.

2.2 The process of annotation in the "top-down" approach

The analysis and approach used in the ongoing dissertation project is inspired by the work of Biber and his colleagues. Even though there are some differences between the approach they define and the approach in the ongoing dissertation project, the core idea of the "top-down" approach is followed. It is essential to understand the structure of the RCCs in discourse analysis. In the ongoing dissertation project, not all of the signs were annotated. Rather, only the chunks that cover potential relative clauses are annotated in a detailed manner. Since this study is based on empirical research on relativization strategies, it would be too time-consuming if each segment was transcribed in a similarly detailed manner. Therefore, it is more practically efficient to follow the "top-down" approach, i.e. to specify first the possible relative clauses in TİD and then to annotate each of them.

The corpus-based approach in the ongoing dissertation project entails seven steps. First, the boundaries of discourse chunks are defined. Second, the possible sentence types included in these chunks are listed and the chunks with potential relative clauses are flagged. Then, tokens/types are constructed for each chunk, which includes possible relative clauses. Before the definition of the boundaries of each relative and matrix clause, the accompanying non-manual markers are defined. The sixth step is to translate the chunks covering the relative clauses into English and Turkish. The final step is to determine the referents in the RCC and its familiarity status within the text (i.e. if the referents have already been introduced to the text or not.).

2.2.1. Step 1: The determination of the boundaries of discourse chunks

The discourse units are narrowed down to smaller units, based on various non-manual and manual cues. Besides the prosodic cues, the meaningful smaller units are also based on semantic intuitions. It is preferred to label these smaller units as discourse chunks, because each chunk includes one or more sentences or clauses, which means that their definitions are open to discussion. The next step is to mark those chunks covering possible RCCs in order to investigate them more deeply.

2.2.2. Step 2: Selecting the chunks which include potential RCCs

RCCs in TİD are usually realized with specific non-manual markers such as raised eyebrows, tensed eyes and cheeks, some head movements and body lean. Tokens are marked with one of non-manual markers which may indicate RCCs.

Specifically, three criteria for marking RCCs in TİD are used: (i) the token includes two clauses, (ii) one clause is dependent on another clause in the selected token (iii) the token is realized with one of specific non-manual markers.

2.2.3. Step 3: Token/type constructions for the flagged discourse chunks

Only the discourse chunks which might include the potential RCCs are annotated. The entries for tokens and types are adapted from the transcription process used in Technical Sign Lexicon Projects (cf. Konrad, 2010), under the auspices of the Institute of German Sign Language and Communication of the Deaf (IDGS). According to Konrad (2010; pp. 28-29), this transcription is based on the distinction between tokens and types, i.e. each token refers to a distinctive type. In other words, *types should be uniquely or consistently identified*.

2.2.4. Step 4: Defining non-manual markers

The next step after annotating the tokens is to annotate non-manual markers for both relative clauses and matrix clauses. The cross-linguistic analyses of relative clauses in signed languages indicate that non-manual markers in relative clauses are generally accompanied by brow raise, tensed eyes/squint, and head movements if needed. Therefore, three tiers are constructed for annotating non-manual markers: (i) eyebrow movements, (ii) tensed

eyes/cheeks and (iii) head/body movements.
Common categorizations for eyebrow movements are (i) brow raise, (ii) neutral brow and (iii) furrowed brows (cf. Wilbur, 2000). Both brow raise and furrowed eyebrow raise are indicated by 'br' and 'fb' respectively and any other eyebrow movement assumes a neutral eyebrow code. Other non-manual markers are also involved, such as: tensed lips (i.e. ASL: Liddell, 1978), tensed eyes (i.e. LSC, Mosella, 2010), tensed cheeks (i.e. LIS, Branchini and Donati, 2009) and squint (i.e. Dachkovsky and Sandler, 2009). It is assumed that these four facial expressions resemble each other and I categorize them as squint which is coded as 'sq.' In addition, some head and torso movements may accompany relative clauses, even though they are not strong indicators. In order to mark these indicators, the third tier represents head and torso movements which include head tilt (back) 'ht', head nod (forward) 'hn', head shake 'hs', and body lean 'bl' (cf. Wilbur, 2000).
Non-manual expressions are not restricted to relative clauses. Different non-manual markers in matrix clauses may be observed as well. These markers may give a clue about sharp boundaries between relative clauses and matrix clauses (cf. Dachkovsky and Sandler, 2009). Also, these non-manual markers occurring in matrix clauses can be independent from the indication of relative clauses (e.g. negation, question). Therefore, another tier is constructed for the investigation of facial, head and torso movements in matrix clauses.
Furthermore, lower face movements may be significant for the realization of relative clauses. For instance, in TİD tensed lips and the mouthings 'o' and 'bu' are frequently observed. These are also coded separately.

2.2.5. Step 5: Defining boundaries of RCCs

After specifying the non-manual markers, the boundaries of relative and matrix clauses need to be specified as well. Boundaries are primarily based on non-manual markers such as brow raise and squint.

2.2.6. Step 6: Translation equivalents of potential RCCs

Turkish translation equivalents and Turkish glosses of Turkish Sign Language, as well as English glosses and English translation equivalents, are provided in a separate tier. Translation equivalents of some RCCs may not represent potential TİD RCCs exactly because of possible cross-language/cross-modal differences in syntactic constructions.

2.2.7. Step 7: Discourse analysis of RCCs

The referents that are used in RCCs are determined and interconnections between the referents are checked. This helps to understand the function of RCCs. This study focuses on the function of RCCs in various discourse modes from a linguistic point of view, in the framework of the Segmented Discourse Representation Theory (SDRT: Asher and Lascarides, 2003).
Aksu-Koç and Erguvanlı-Taylan (1998; p. 277, inspired by Fox and Thompson, 1990) specify two different references to the expressions (i) *head* and (ii) *modifying clause*. According to them, head can either be introduced into discourse for the first time, or else introduced again in the sense of the familiarity status of information. The information in a modifying clause can be realized in three different forms. If the modifying clause is made for clarifying the ambiguous content of the head, the clause has an *identification* function. If the content of the modifying clause has already been introduced earlier and is once again introduced into the discourse, it has been *re-identified*. Conversely, some modifying clauses may function as tools to express supplementary information about the head. Such clauses are regarded as *characterizing* modifying clauses. Using this categorization, in the ongoing dissertation project each head and modifying clause in flagged discourse chunks with potential RCCs in TİD is identified with underlying properties.

3. Advantages and disadvantages of the "top-down" approach

The annotation process in this dissertation project favors the "top-down" approach. This process has both advantageous and disadvantageous sides. The first advantage is that the "top-down" approach is primarily based on a specific research question and can focus on the findings and annotations that are related to this goal. The second advantage of this approach is the fact that it does not tokenize data which may not be related to the specific goal. The third advantage is that this approach allows deduction, i.e. from wider linguistic units to narrower units. For instance, this study looks at the discourse text first and divides it into possible discourse chunks and phonological utterances (cf. Sandler and Lillo-Martin, 2006). It also goes further into intonational phrases, phonological phrases and even prosodic words (i.e. here tokens). In addition, after deduction, it allows an inductive approach as well, e.g. in the ongoing dissertation project tokens may give a clue about the syntactic constructions.
However, this approach has disadvantages as well. If all discourse chunks are not treated equally, there is a danger of missing potential samples. For instance, in the ongoing dissertation project not all discourse chunks are glossed in terms of tokens/types and therefore other possible relative clauses may potentially be overlooked. In order to avoid such loss, each discourse type has been labelled with respect to its sentence types, as far as possible. This strategy may make up for the first disadvantage. The second drawback is that there is a need for a native signer with meta-linguistic awareness so that he/she may decide which chunks may include potential data related to the specific research aim.

4. Conclusion

Due to modality-specific properties, the "top down" approach can be seen as challenging to use for corpora in signed languages. No matter how large the corpus is, in order to understand the function of a linguistic element, a "top down" approach can assist in the obtaining of a bigger picture of the discourse development. As mentioned before, RCCs in TİD do not necessarily have a linguistic and manual form. Rather, RCCs in TİD mostly rely on prosodic constituents, which can vary. Starting from a text and dividing into smaller units with the help of non-manual expressions as well as semantic

intuitions made the analysis of RCCs in TİD possible.

The approach developed for this study might have some drawbacks and may benefit from further refinements; however, this approach might shine a light on the investigation of linguistic forms in signed languages which might not have a manual form, such as yes/no questions, topicalization and RCCs.

5. References

Aksu-Koç, A. and Erguvanlı-Taylan, E. (1998). The functions of relative clauses in narrative discourse. Lars Johanson (Ed.) T*he Mainz Meeting Proceedings of the Seventh International Conference on Turkish Linguistics Wiesbaden,* Harrasowitz Verlag, pp. 271--284.

Asher, N. and Lascarides. A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Bank, R. ; Crasborn, O. and van Hout, R. (2013). Alignment of two languages: The spreading of mouthings in Sign Language of the Netherlands. *International Journal of Bilingualism*. Published online before print, May 3, 2013, doi: 10.1177/1367006913484991.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representatives in corpus design. *Literary and Linguistic Computing*, 8, pp. 243--257.

Biber, D., Connor, U. and Upton, T.A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

Biber, D.; Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Branchini, C.; Cardinaletti, A.; Cecchetto, C.; Donati, C. and Geraci, C. (2013). WH-duplication in Italian Sign Language (LIS). *Sign language & linguistics*, 16(2), pp. 157--188.

Branchini, C. and Donati, C. (2009). Relatively different: Italian Sign Language relative clauses in a typological perspective." In A. Liptàk (Ed.), *Correlatives cross-linguistically* Amsterdam: Benjamins, pp. 157--191.

Conrad, S. (2002). Corpus linguistics approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, pp. 75—95.

Dachkovsky, S. and Sandler, W. (2009). Visual intonation in the prosody of a sign language. *Language and Speech,* 52(2/3), pp. 287--314.

Fox, B. and Thompson, S. (1990) A discourse explanation of the grammar of relative clauses in English conversation. *Language* 66, pp. 297--317.

Hanke, T. (2002). iLex - A tool for sign language lexicography and corpus analysis. In M. González Rodriguez, & C. Paz Suarez Araujo (Eds.): *Proceedings of the third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain. Paris: ELRA, pp. 923--926.

Hanke, T. and Storz, J. (2010). iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In Crasborn, Onno et al. (Eds.): LREC 2008. *6th International Conference on Language Resources and Evaluation. Workshop Proceedings. W25. 3rd Workshop on the Representation and Processing of Sign Languages*. Sunday 1st June 2008, Marrakech – Morocco. Paris: ELRA, pp. 64--67.

Johnston, T. (2013). Formational and functional characteristics of pointing signs in a corpus of Auslan (Australian sign language): are the data sufficient to posit a grammatical class of 'pronouns' in Auslan? *Corpus Linguistics and Linguistic Theory*, 9(1), pp. 109-159.

Konrad, R. (2010). Die Erstellung von Fachgebärdenlexika am Institut für Deutsche Gebärdensprache (IDGS) der Universität Hamburg (1993-2010). Universität Hamburg. URL: http://www.sign-lang.uni-hamburg.de/projekte/mfl/Konrad_2010_FachgebLexika.pdf (last accessed on January 06 2012).

Kubus, O. (in progress) Relative Clause Constructions in Turkish Sign Language (TİD) Ph.D. dissertation, Universität Hamburg.

Liddell, S. (1978). Non-manual signals and relative clauses in ASL. In P. Siple (Ed.), *Understanding language through sign language research*. New York: Academic Press.

Mosella Sanz, M. (2011). The position of fronted and postposed relative clauses in Catalan Sign Language. *Paper presented at Venice FEAST Colloquium, Formal and Experimental Advances in Sign language Theory*, Venice, June 20-22, 2011

Perniss, P.; Pfau, R.; M. Steinbach (2007). Can't you see the difference? Sources of variation in sign language structure. In P. Perniss, R. Pfau & M. Steinbach (Eds.), *Visible variation: Comparative studies on sign language structure*. Berlin: Mouton de Gruyter, pp. 1--34.

Pfau, R. and Steinbach, M. (2005). Relative clauses in German Sign Language: Extraposition and reconstruction. In L. Bateman and C. Ussery (Eds.), *Proceedings of the North East Linguistic Society (NELS 35)*, Vol. 2. Amherst, MA: GLSA, pp. 507--521.

Sandler, W. and Lillo-Martin, D. (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.

Smith, C. S. (2003). *Modes of Discourse*. Cambridge University Press.

Wilbur, R. B. (2000). Phonological and prosodic layering of nonmanuals in American Sign Language. In K. Emmorey & H. Lane (Eds.), *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima*. Mahwah, NJ: Lawrence Erlbaum.

Signing thoughts!

Andrea Lackner & Nikolaus Riemer
A methodological approach within the semantic fieldwork used for coding nonmanuals
which express modality in Austrian Sign Language (ÖGS)
E-mail: andrea.lackner@aau.at, nikolaus.riemer@ling.su.se

Abstract

Signing thoughts gives the possibility to express unreal situations, possibilities and so forth. Additionally, signers may express their attitude on these thoughts such as being uncertain about an imagined situation. We describe a methodological approach within the semantic fieldwork which was used for identifying nonmanuals which tend to occur in thoughts and which may code (epistemic and deontic) modality in Austrian Sign Language (ÖGS).
First, the process of recording short stories which very likely include lines of thoughts is shown. Second, the annotation process and the outcome of this process are described. The findings show that in almost all cases the different annotators identified the same non-manual movements/positions and the same starting and ending points of these nonmanuals in association with the lexical entries. The movement direction was allocated to one direction of the three body axes. Furthermore, some nonmanuals were distinguished due to intensified performance, size of performance, speed of performance, additional movement components, or additional body tension. Finally, we present nonmanuals which frequently occur in signed thoughts. These include various epistemic markers, a deontic marker, indicators which show the hypothetical nature of signed thoughts, and an interrogative marker which differs from interrogative markers in direct questions.

Keywords: nonmanuals, signing thoughts, (epistemic and deontic) modality, Austrian Sign Language

1. Expressing modality[1] by signing thoughts

Expressing thoughts is an excellent way of abstracting away from the here and now. Using this way of expressing oneself gives the possibility to speak/sign about unreal situations, wishes, possibilities, conditions and so forth. When doing so, also attitudes on these thoughts such as being certain or uncertain of the realization of a situation can be expressed.
We present a methodological approach for producing, identifying and analyzing nonmanuals which code (epistemic and deontic) modality, implemented in the framework of semantic fieldwork[2]. Phase 1 comprises the implementation of producing a type of signed context in which frequently non-manual means for coding modality occur. Phase 2 includes the process of identifying these nonmanuals by Deaf annotators. In Phase 3 these elements are analyzed with regard to their context of occurrence and their co-occurrence with other (non-manual) elements.

2. Producing signed thoughts

In Phase 1 the Deaf informants were asked to sign a longer action, e.g. hiking or driving. Furthermore, they were told that during this longer ongoing action they should think about a certain situation and wonder whether this or that situation will/would occur or to express possible conditions about the imagined situation. These trains of thoughts were then expressed with different attitudes or knowledge about the imagined situation such as being unaware of certain circumstances in this situation, being uncertain about the occurrence of a situation, being full of hope that the imaginations will come true and so on. The instruction was given twice, once by a video in which a Deaf lecturer described the task and once by a Deaf participant who coordinated the video recording process and who constantly guided the Deaf informants through the task. After giving the instructions, informants were asked to sign informal stories as a kind of warming up. After about 10-15 minutes, they were asked to sign stories which should also include trains of thoughts. The recordings were implemented in sitting and standing position. The Deaf informants were instructed to sign in standing position and afterwards to repeat (in general) signed contents while sitting. The narrations (longer and shorter stories) had to be signed twice in the particular positions.
As the recordings took place in the informants' Deaf club, a location with which the informants are very familiar, the following situation occurred: The part of the club where the recording took place was just one of the various places in the Deaf club, where the participants were busy signing. Thus, being visible to the others resulted in being watched by the other club visitors for a while or being interrupted by the others; also the signers who were doing the recordings started to chat with others and then continued signing for the camera. To be precise, the recording location was just one 'scene of communication' in the Deaf club and consequently, a well-ordered production of stories including lines of thoughts expressed with the first attitude on these thoughts, the second attitude on these thoughts and so on did not take place. However, compared to recording in a studio, this situation offered the possibility to record a very natural way of signing.
After analyzing the recordings, the outcome shows that six out of nine informants really produced lines of

[1] The term 'modality' is used, as it refers to the semantic domain while the term 'mood' is avoided as it is mostly associated with grammatical categories like indicative and subjunctive.
[2] An introduction/description on methodology in semantic fieldwork is given by Matthewson (2004).

thoughts while signing short stories. When producing thoughts, a topic was chosen (e.g. hiking in the mountains) to which different short stories were signed. Before telling the outcome of the story, these short stories included lines of thoughts. Furthermore, the data show that the participants were inspired by the topics being produced by the other signers such as going hiking and visiting a hut, playing cards, and so forth. Thus, the positive effect of being visible to the other informants was that the instructions were clear to most of the informants and the contents of the signed texts were quite similar. This resulted in data which was excellent to compare with each other. For instance, a scenario which was signed by all participants was that somebody is hiking and thinking about a hut which might be open or closed. This scenario was then expressed with different attitudes on this situation such as wondering, being certain, or being uncertain whether the hut is open or closed.

What is more, the recordings show that the various informants did not produce the same order of stories as the proceeding informant, nor the same kind of thoughts. So, they produced in their lines of thoughts declaratives, interrogatives and conditionals as well as various epistemic modalities in highly diverse orders.

To conclude, this procedure guaranteed us that the productions from the various informants were not strongly biased from previous signers' expressions as it is very unlikely that an informant remembers the exact non-manual configuration used to express one of the types of epistemic modality after a 20-30 minutes recording session produced in such an interactive setting.

With regard to data, the entire recordings last five hours in total. From these recordings 40 minutes were annotated by four (partly five) Deaf annotators. These annotated recordings include short stories in which six informants expressed their thoughts.

3. Identifying nonmanuals occurring in signed thoughts

In Phase 2 the recordings were annotated in ELAN[3] by four (partly five) Deaf annotators. To be precise, the signs were glossed by the first annotator. Afterwards each communicatively relevant non-manual element was described with regard to its form and meaning/function in the particular context by four/five annotators per recording. The template for the annotators included a tier for each non-manual articulator which may code communicatively relevant information. In sum, besides the parameters gloss-left-hand and gloss-right-hand the template included tiers for coding mouth movement(s)/position (including a separate tier for 'mouthing' and 'mouth gesture'), eye gaze movement(s), eye aperture and eye brow position/movement(s) and facial movement(s)/positions. With regard to the articulators head and body, each communicatively relevant instance of head position or movement(s) and body position or movement(s) along a body axis was annotated in separate tiers. The set of head and body tiers included: tilt-forward/backward, chin up/down, head tilt-right/left, head turn-right/left, head rotation/etc.; body turn-left/right, body lean forward/backward, body lean-sideward/sways/shifting of weight/step, shoulder(s)/body straitening-up.

When doing the annotation the Deaf annotators were instructed to identify the nonmanuals' characteristics which are:

- the kind/sequence of motion –i.e. whether the particular non-manual element is/are movement(s) or a position of a particular articulators,
- the exact beginning and ending points of these non-manual means
- the direction of motion for this non-manual element (e.g. positioning the head forward versus positioning the head backward)
- the intensified performance, the size of performance and/or the speed of performance of the identified non-manual element, if relevant for the annotator
- additional co-occurring factors such as the degree of body tension or additional movement components, if relevant to the annotator
- and the current possible meaning of the identified non-manual element in the particular context.

In order to compare the annotations of the different annotators, each of them got a separate list of non-manual tiers. When annotating, the annotations of the others were concealed, only the glossing tier was visible to everyone. This process resulted in at least four different annotations of the various non-manual tiers which were compared afterwards as illustrated in Figure (1).

Figure (1) shows an example of the annotated data. For reasons of clarification, the annotations of each of the four Deaf annotators (A to D) are edged red, green, blue and yellow. It is shown that the annotators identified the same movement/position (here the marker 'head forward; encircled red) as well as the same starting and endpoint of the non-manual element (encircled green). Also their descriptions of the semantic meaning of these elements were quite similar.

All nonmanuals which were identified by at least three of the four (partly five) annotators were taken for the analysis. To be precise, nonmanuals which had an inter-annotator agreement of at least three annotators were adduced as instanced for the analysis[4].

In conclusion, the striking outcome of this procedure was:

- First, in almost all cases the different annotators identified the same non-manual movements/positions. For instance, as shown in Figure (1) all annotators identified the same distinctive marker – i.e. 'head forward'.

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

[4] As the focus of this investigation was the identification of nonmanuals which had been unknown so far and as for this research human resources were limited, the statistic evaluation is limited to general data information.

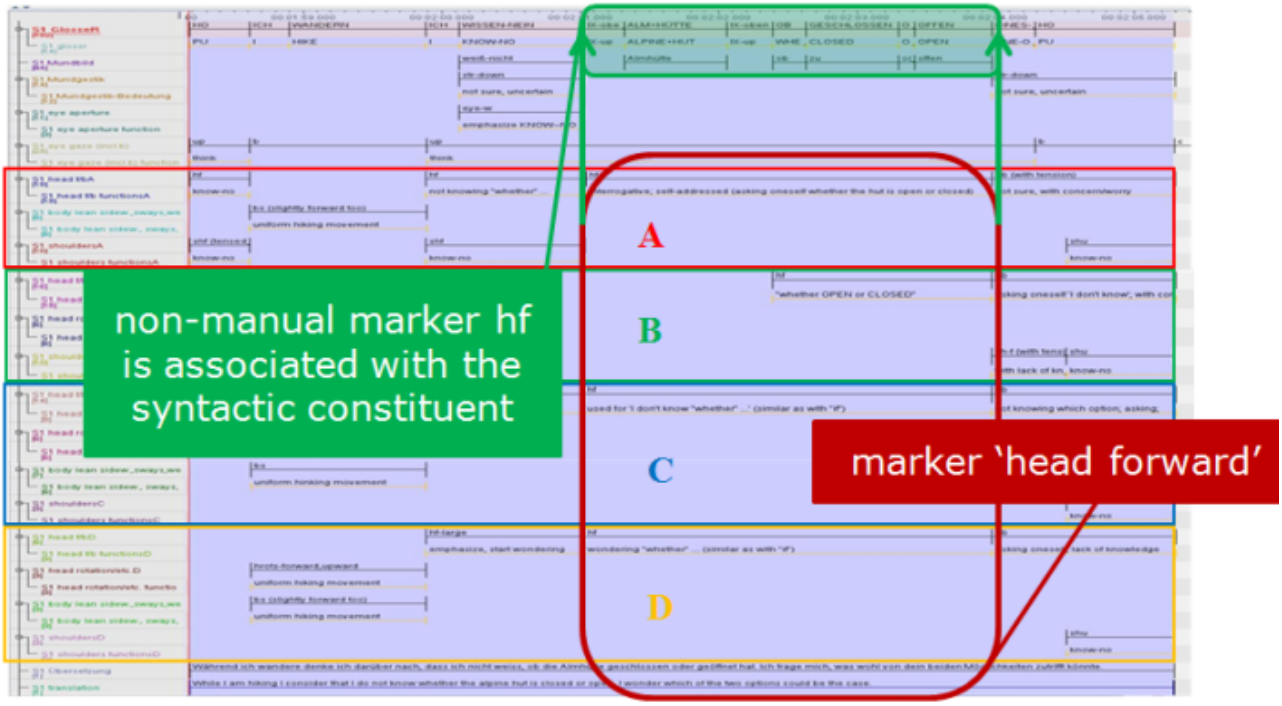


Figure (1): Identified non-manual marker 'head forward', associated with the syntactic constituent by annotator A, B, C and D (Lackner 2013, 70)

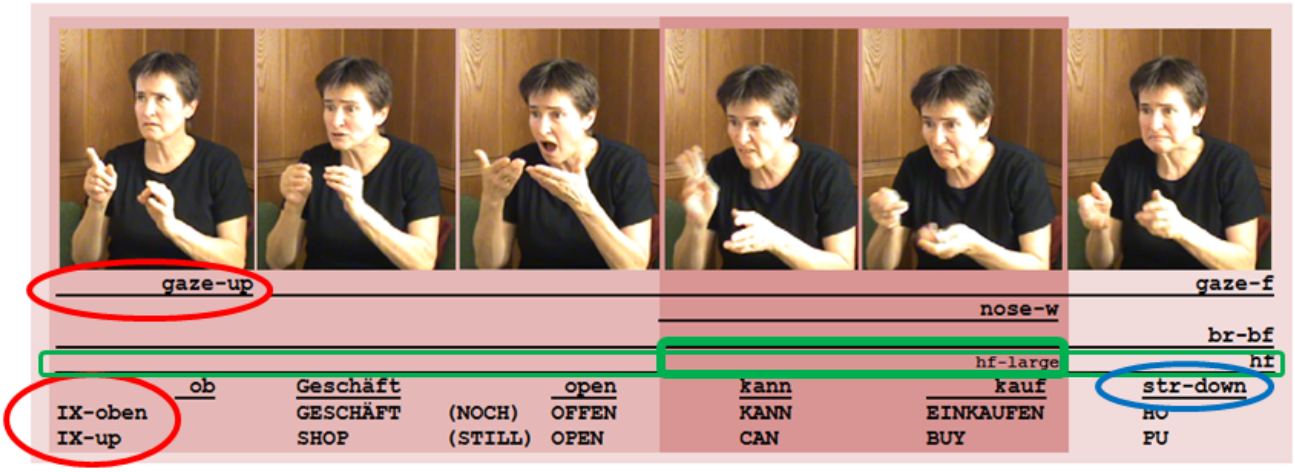
- Second, in the majority of instances the annotators determined (not influenced by each other) the same starting and ending points of these non-manual movements/positions in association with the lexical entries. This result shows that there must be a high tendency in sign languages (SLs) of alignment between non-manual components with lexical entries, which they associate to with regard to the production but, most notably in the perception of the signing flow. Also, the annotated ÖGS-data show that a variety of these nonmanuals are associated with the syntactic constituent, as illustrated in Figure (1).
- According to the annotators' feedback, each non-manual element showed the following characteristics: The kind/sequence of motion was perceived as 'movement' or 'position' (e.g. constantly forward movements of the head versus positioning the head forward). The direction of motion was in the majority of cases distinguished by a contrast of movement/position (e.g. positioning the head forward versus positioning the head backward). Some nonmanuals were distinguished due to intensified performance or the size of performance (e.g. positioning the head forward versus positioning the head forward in an intensified way, or producing headshakes with a small radius versus headshakes with a large radius), the speed of performance (e.g. producing fast headshakes versus producing slow headshakes), the degree of body tension (e.g. producing non-tensed headshakes or performing head nods in an trembling way with a tensed body), and an additional movement component (e.g. head nods with trembling movement or headshakes with alpha-movement).

All these new insights were implemented in our annotation conventions. In brief, when annotating nonmanuals, first, an abbreviation of the articulator is given (e.g. 'h' for the head). Second, the direction of movement is added (e.g. 'hf' for head positioning forward). Third, additional information is attached with a hyphen (e.g. 'hf-large' for positioning the head forward in an intensified way). Also, the information whether the identified element is a position or movement(s) is attached, if that information is of relevance for the annotator (e.g. 'hn' for a single head nod while 'hns' for several nodding movements). What is more, we realized that with articulators such as the head and the body more non-manual elements could co-occur. For instance, it is possible to produce nods together with putting the head forward and tilting the head to the side. The annotators allocated to all of these co-occurring movements/positions of the head a certain meaning/function. This finding resulted in creating a template for ELAN which includes for each possible direction of movement of the head and body a separate tier.

4. Analyzing nonmanuals occurring in signed thoughts

Related to analyzing signed thoughts (Phase 3), the most striking finding was that the annotated data showed that various nonmanuals are used for coding modality.

Using modal verbs for coding modality (both deontic and epistemic modality) has been described for other Sign Languages (SLs) (Wilcox & Shaffer 2006 in American SL or Pfau & Quer 2004 in German SL and Catalan SL). Non-manual elements (face, head, body) that can co-occur with these modality verbs have been described to some extent as well. Also, modal particles occurring in



(Während ich mit dem Zug fahre überlege ich,) ob das Geschäft noch offen hat und ob ich noch einkaufen kann. Ich bin mir aber sehr unsicher.
(While I'm going by train I wonder) whether the shop will still be open and whether there will still be time to go shopping. But I am very uncertain.

Figure (2): Signed thought in ÖGS

SLs have been described (Hermann 2013, for German SL).

Our data show that in ÖGS a second modality system exists which comprises nonmanuals used to code modality. First, this is a set of non-manual markers which are used to code epistemic modality. They serve to mark the signer's knowledge and/or degree of confidence of the true value of a proposition and are labeled by Lackner (2013, 324-347): assertive marker, non-assertive maker, dubitative marker, and trembling marker. Second, the annotations show that a head marker which is also used to show contrast or alternatives, is used to express deontic modality. To be precise, tilting the head to the side is used to express the possibility/probability of realization of an imagined situation. Third, there are further means (most of them are nonmanuals) which also frequently occur when expressing unrealized thoughts or when wondering about an unreal situation. These are different indicators which refer to a higher place in the signing space, labeled as 'hypothetical space' (Lackner 2013, 260). These elements (co-)occur in the initial position of the thought or co-occur with the entire thought. These are indexing (pointing) upward, gaze-up, chin-up and displacement of the sign's place of articulation into a higher signing space. Moreover, the data show that an interrogative marker which is different to the interrogative markers used in direct questions or constructed dialogues occurs when wondering about an imagined situation. Interestingly, the same head marker is used as conditional marker by all informants. Finally, there are other non-manual markers which also frequently occur in signed thoughts, but which require further investigations. First, the marker 'squinted eyes', which is frequently associated with knowledge or lack of knowledge by the various annotators, needs to be looked at more closely. The second identified non-manual marker is 'wrinkled nose' which occurs in the majority of conditionals which include negativity. According to the annotators' feedback this marker might express the negative attitude on an imagined situation.

Some of these indicators which code modality meaning in ÖGS are shown in the following Figure (2).

Figure (2) shows a line of thought in which the signer wonders whether the shop will be open and whether there will still be time to go shopping. This is followed by showing the signer's uncertainty, expressed by the mouth action 'closed mouth, lips stretched, corners slightly go down' (encircled blue). The indicators referring to the 'hypothetical space' are looking and indexing upward, both produced in the beginning of the line of thought (encircled red). The questionability/interrogativity of the entire utterance is expressed by positioning the head forward (encircled green), in an intensified way while signing CAN BUY (encircled green in bold), co-occurring with winkled nose which might express the negative attitude on the probability of realization this imagined situation.

5. Conclusion

To sum up, our study shows a methodological approach used to identify various indicators which code modality meaning in ÖGS.

To begin with, our solution for receiving recordings which comprise various means of coding (epistemic and deontic) modality was to let the signers express their thoughts by signing a short story. Embedding signed thoughts in short stories as well as offering a familiar atmosphere (where the informants could see each other) was the right setting to get data which contains a lot of information coded by nonmanuals and various elements which code modality meaning.

Then, we instructed all Deaf annotators to identify the nonmanuals' characteristics such as the kind/sequence of motion, the exact beginning and ending points of these non-manual elements and so on. In doing so, we gained the insights that the different annotators identified the same non-manual movements/positions, the same starting and ending points of these non-manual

movements/positions in association with the lexical entries, and further characteristics of these non-manual elements such as size or speed of performance.

Our findings show that in ÖGS various nonmanuals exist which express modality meaning. In particular epistemic modality is coded by various non-manual markers when signing thoughts. The findings also show that there are other nonmanuals which frequently occur in signed thoughts such as indicators for expressing the hypothetical nature of thoughts or an interrogative marker which differs from interrogative markers used in direct questions or constructed dialogues.

6. References

Lackner, Andrea (2013) Functions of head and body movements in Austrian Sign Language (ÖGS).A corpus-based analysis. PhD thesis at the University Graz.

Hermann, Annika (2013) Modal and focus particles in sign languages. A cross-linguistic study (Sign Languages and Deaf Communities (SLDC) 2). Berlin: Mouton de Gruyter.

Matthewson, Lisa (2004) On the methodology of semantic fieldwork. In: International Journal of American Linguistics 70(4), 369-415.

Pfau, Roland & Quer, Josep (2004) On the syntax of negation and modals in Catalan Sign Language and German Sign Language. 26th Annual Meeting of the German Linguistic Society (DGfS), Mainz, February 27th 2004.

Wilcox, Sherman & Shaffer, Barbara (2006) Modality in ASL. In: Frawley, William (ed.) The Expression of Modality. Berlin: Mouton de Gruyter, pp. 207-238.

Estimating Head Pose and State of Facial Elements for Sign Language Video

Marcos Luzardo*, Ville Viitaniemi*, Matti Karppa*, Jorma Laaksonen*, Tommi Jantunen†
*Department of Information and Computer Science
Aalto University School of Science, Finland
firstname.lastname@aalto.fi

†Sign Language Centre, Department of Languages
University of Jyväskylä, Finland
tommi.j.jantunen@jyu.fi

Abstract

In this work we present methods for automatic estimation of non-manual gestures in sign language videos. More specifically, we study the estimation of three head pose angles (yaw, pitch, roll) and the state of facial elements (eyebrow position, eye openness, and mouth state). This kind of estimation facilitates automatic annotation of sign language videos and promotes more prolific production of annotated sign language corpora. The proposed estimation methods are incorporated in our publicly available SLMotion software package for sign language video processing and analysis. Our method implements a model-based approach: for head pose we employ facial landmarks and skins masks as features, and estimate yaw and pitch angles by regression and roll using a geometric measure; for the state of facial elements we use the geometric information of facial elements of the face as features, and estimate quantized states using a classification algorithm. We evaluate the results of our proposed methods in quantitative and qualitative experiments.

Keywords: head pose estimation, facial state recognition, sign language analysis

1. Introduction

Currently there is an increasing need of automatic video analysis and annotation tools to support linguists in their studies of sign language (SL). Henceforth, studies focusing on automatic annotation of SL videos and non-manual gestures are continuously developing. In this work we study methods for automatic estimation of three head pose angles (yaw, pitch, and roll) and the state of facial elements (eyebrow position, eye openness, and mouth state). Our main motivation is to facilitate automatic annotation of SL videos and promote more prolific production of annotated SL corpora. The estimation methods proposed in this work are incorporated in the SLMotion software package (Karppa et al., 2014) for SL video processing and analysis.

We propose an approach for head pose estimation from images based on two kinds of visual features. The first group of features is formed by facial landmarks extracted using the flandmark software library (Uřičář et al., 2012). Secondly, as novel additional features we use tonal segmentation masks of skin-like colors within the face area. The yaw and pitch angles are estimated using separate Support Vector Regressors (Smola and Schölkopf, 2004). The roll angle is estimated using a geometric approach based on the location of the eye landmarks.

Our method for estimating eyebrow position, eye openness, and mouth state is based on the construction of an extended set of facial landmarks that are not part of the flandmark output. The proposed landmark detection algorithm employs different techniques designed for each facial element. For comparison, we also consider landmarks detected using the Supervised Descent Method (Xiong and De la Torre, 2013). The extended landmarks are used to compute a set of geometric features which are further post-processed using Principal Component Analysis. The processed features function as input for the Naive Bayes and Support Vector Machine classifiers in order to produce quantized estimates of the state of facial elements.

The estimation performance of the head pose and the state of facial elements are evaluated quantitatively and qualitatively. Motion capture data from a SL recording session is used for quantitative evaluation of the head pose. The state of facial elements uses manually annotated data from SL video sequences. In both cases the qualitative evaluation is performed from a linguistic point of view.

The rest of the paper is arranged as follows: in Section 2 the state of recent research in estimation of head pose and state of facial elements is presented. In Section 3 details of the head pose estimation method are presented. The estimation of states of facial elements is elaborated in Section 4. Conclusions drawn from this work are summarized in Section 5.

2. Related work

In addition to the activity of the hands, an important part of signing is the layered activity of the non-manual (NM) articulators such as the head and its components: eyebrows, eyes, and mouth. In signing, the activities of these articulators express various linguistically significant functions (Pfau and Quer, 2010). For example, a head shake is the primary means through which SLs mark sentence-level negation; head nods, in turn, are used in SLs to signal, for instance, affirmation, existence, and emphasis. The functions of the activities of eyebrows, eyes, and the mouth are equally important. For example, the various states of eyebrows and eyes mark both domains and boundaries of syntactic constituents. The activity of the mouth, on the other hand, is often used morphologically to modify the basic meaning of signs.

2.1. Head pose estimation in sign language

Head pose is determined by three angles: horizontal movement or *yaw*, vertical movement or *pitch*, and rotational movement or *roll* (Figure 1a). The angles can be estimated with either model-based approaches using a number of facial features, or with appearance model approaches that use the entire image of the face. While several methods have reported good results using appearance-based approaches, more advanced model-based methods use appearance models to learn shape variations.

A popular approach has been to interpret pose detection as a classification problem and train a set of pose-specific classifiers for recognizing pose angle ranges (Whitehill and Movellan, 2008). The opposite approach has been to directly estimate the pose angles, e.g. with methods such as regression in combination with dimensionality reduction techniques. We are not aware of any previous SL studies where visually estimated pose would have been compared with a ground truth obtained from motion capture.

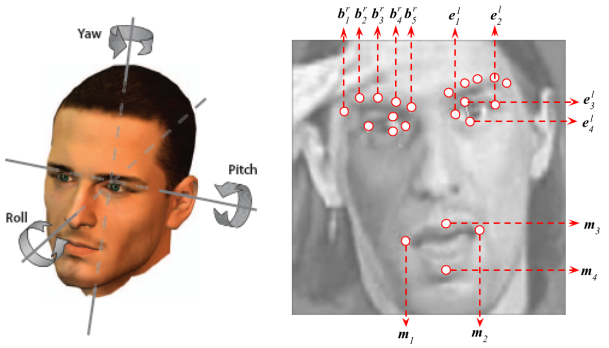
2.2. State of facial elements estimation in sign language

Studies in the state identification and tracking of individual facial elements are strongly related to facial expression analysis. The use of facial expression analysis for NM marker estimation has been reported for a defined set of facial movements (Metaxas et al., 2012). Research on comprehensive sign-to-text/speech translation system have also incorporated NM marker estimation (Dreuw et al., 2010; Campr et al., 2010). However, the maturity of these systems is still low.

Isolated studies for eyebrow estimation are scarce; early studies in eyebrow movement demonstrated that some facial expressions can be identified by the eyebrow position alone. Recently, a method trained to detect eyebrow articulations and other NM facial gestures for American Sign Language (ASL) was reported in (Liu et al., 2013) with promising results. Eye openness and blinking estimation has been of special interest for hypo-vigilance detection in a varying range of applications (Hansen and Ji, 2010). Blink detection from video sources has been benchmarked against electrooculography (EOG) approaches, where it has been demonstrated that robust results can be achieved (Picot et al., 2012). Estimation of mouth shapes has been done primarily for gesture recognition, lip reading, and for hypo-vigilance. Estimation methods aimed at aiding lip reading typically extract the shape produced by the lips' outer boundaries to improve detection rates in speech recognition tasks (Gómez-Mendoza, 2012).

3. Head pose estimation

In this section we present a method for automatic estimation of head pose from images. Head pose is defined here as having three angles of movement: yaw, pitch, and roll. We follow a model-based approach to estimate the three head pose angles. Facial landmarks and a skin mask are extracted from a set of training images and combined to form a feature vector. The resulting features are used as input data to estimate pitch and yaw using Support Vectors Regression (SVR) (Smola and Schölkopf, 2004) with radial basis functions as kernels.



(a) (b)

Figure 1: (a) Degrees of freedom of the human head described by rotation angles (Murphy-Chutorian and Trivedi, 2009). (b) Facial landmarks for geometric feature extraction. The eyebrow and eye landmarks have the same left-to-right numerical ordering on both sides.

We estimate roll angles by a geometric approach using the image plane with the assumption that the facial landmarks have been correctly approximated and the camera is aligned at zero degrees. The roll angle is determined by simple trigonometry from the angle between the image axis and an imaginary line drawn connecting the eye centers.

The Pointing04 image database (Gourier et al., 2004) is used for training the SVRs, the selected images are within *near frontal* angles. The different combinations of facial landmark points, their normalizations and combination with facial skin area information are tested to find an optimal set of features that can provide reliable pose angle information. Finally, the model is used to estimate head pose from a SL video where the ground truth pose angles are available from a motion capture recording.

3.1. Feature extraction

This section details the different features employed for head pose estimation. The $(x_0, y_0), (x_1, y_1)$ coordinates that define the face area *bounding box* are also included as part of the features.

3.1.1. Landmark detection

Facial landmarks are extracted using the flandmark package (Uřičář et al., 2012). The package is based on Deformable Part Models: given an appearance fit and deformation cost functions, the facial points are constrained to fit within a structured component graph. The flandmark output is composed of $8 \times (x, y)$ coordinates points. Since face location and size vary across images, the landmarks are normalized into the range of $(x, y) \in [0, 1] \times [0, 1]$ with respect to the bounding box.

3.1.2. Skin mask

As a novel technique for aiding the identification of the head pose, a skin-tone mask was extracted from each image. The skin mask consists of tonal segmentation of skin-like colors images. The binary mask is used to calculate four additional values for regression: the fractional areas of

non-skin pixels on the left and right side of the face bounding box, L and R, respectively, and similarly the top and bottom areas T and B, all in the range $[0, 1]$.
In the evaluation, we have used the four fractional non-skin areas as such, but also considered coordinate normalization by offsetting the point coordinates with respect to the mask areas. For yaw and pitch angle estimation, we displace the landmark (x, y) coordinates independently in proportion to the left/right (yaw) and top/bottom (pitch) mask areas to get the *offset normalized coordinates* (x', y') as

$$x' = x - L + R \ , \quad (1)$$
$$y' = y - T + B \ . \quad (2)$$

3.2. Experiments

The performance of the proposed head pose estimation method was evaluated in two experimental settings. In the first series of experiments a subset of the Pointing04 data was used to measure the accuracy of the trained yaw and pitch regressors. In the second experiment head pose was estimated from a video of continuous signing during a motion capture session and the estimates compared with the ground truth values from the motion capture recording.

3.2.1. Data

The selected images from the Pointing04 database have angles in the ranges $\pm 45°$ in yaw, and $\pm 30°$ in pitch. The Pointing04 data used for training does not include non-zero roll angles. The angle differences are $15°$ from one pose to the other. Two sets of feature vectors with different angular distributions were selected for training the regressors. The first set, A, results from 684 images for which the landmark detection had been successful and consecutively has an emphasis on the near frontal poses. The second set, B, contains $29 \times 7 \times 5 = 1015$ feature vectors equally distributed in all poses. This set was generated by adding 366 synthetic samples based on pose-specific pixel location means and variances from set A. The synthetic samples were created as $x = \mu + r\sigma$ with mean μ, standard deviation σ and a random factor r in the range ± 0.75, and similarly for y.

3.2.2. Classification experiment

Sixteen experiments were performed for both data sets A and B to find the best combination of facial features, and to determine the usefulness of the skin masks. All SVRs were evaluated independently for yaw and pitch for both data sets with leave-one-sample-out cross validation. We quantized the regressors outputs to the nearest values in $0, \pm 15, \pm 30, \pm 45$ degrees for yaw and $0, \pm 15, \pm 30$ degrees for pitch.
The results (Table 1) indicate that for yaw, ignoring the face center landmark increases the accuracy whereas for pitch it provides important reference information. The results also show that it is always better to use both coordinates for estimating the angles. It is clearly beneficial to use the offset normalized coordinates (x', y') for yaw, but not so much for pitch. The best results were, however, obtained when the skin area pixel counts are used as such in the feature vector. It seems that, for yaw, training with the set A mostly produces better results whereas, for pitch, the additional synthetic values in set B bring improvement.

Point set	Yaw$_A$	Yaw$_B$	Pitch$_A$	Pitch$_B$
$8 \times x, y$	50.29	49.71	45.18	46.35
$8 \times x, y + L, R, T, B$	66.81	66.96	**51.75**	52.63
$8 \times x', y'$	68.28	67.69	47.66	45.76
$8 \times x', y' + L, R, T, B$	68.72	64.91	47.22	48.25
$7 \times x, y$	48.98	48.83	44.74	45.61
$7 \times x, y + L, R, T, B$	68.86	**69.29**	49.56	**54.24**
$7 \times x', y'$	**69.15**	67.69	44.44	46.78
$7 \times x', y' + L, R, T, B$	69.15	66.08	44.15	47.81
$8 \times x\,(y)$	49.71	46.49	44.15	45.76
$8 \times x\,(y) + L, R\,(T, B)$	64.47	61.55	45.76	46.93
$8 \times x'\,(y')$	63.89	60.38	45.76	44.74
$8 \times x'\,(y') + L, R\,(T, B)$	63.60	63.74	47.81	45.91
$7 \times x\,(y)$	47.52	42.84	44.01	45.18
$7 \times x\,(y) + L, R\,(T, B)$	62.87	59.06	45.91	46.49
$7 \times x'\,(y')$	64.62	62.43	42.84	45.91
$7 \times x'\,(y') + L, R\,(T, B)$	63.74	63.74	46.20	46.78

Table 1: Classification accuracy with different feature vectors and training data. In the third and fourth vertical blocks only the x coordinates were used for yaw, and only the y coordinates for pitch. In training set A the images had a stronger distribution near the central poses, in set B poses were equally distributed. All values are percentages.

	MAE		Classification
Model	Yaw	Pitch	Accuracy %
FL+SVR A	6.2°	8.8°	{69.2, 51.8}
FL+SVR B	6.2°	8.8°	{69.3, 54.2}

Table 2: Performance of fine pose estimation and pose angle classification. Listed methods use 13 discrete poses for yaw and 9 for pitch. Our work uses 7 discrete poses for yaw and 5 for pitch.

The angle classification errors and mean absolute errors (MAE) were calculated for our best methods (Table 2) using the Pointing04 data set as similar studies have done (Murphy-Chutorian and Trivedi, 2009). The results are not directly comparable as our method has been limited to near frontal angles only. Nevertheless, the proposed method shows improved classification accuracy for the yaw angle and similar accuracy for the pitch angle, compared to previously reported studies.

3.2.3. Sign language video experiment

In our final head pose experiment, the best regressors were used to estimate the yaw and pitch angles in a SL video. The roll angles were obtained using the previously described geometric approach. The video was obtained during a motion capture recording session and comprises continuous signing with a variety of naturally occurring head movements and poses. The estimated angles were visualized using a gyroscope plot to aid the interpretation of the results (Figure 2).
The estimated angles were low-pass filtered using a FIR filter of order five to reduce the observed noise. These smoothed values are compared (Figure 3) with the ground truth obtained from the recorded motion capture data (Jan-

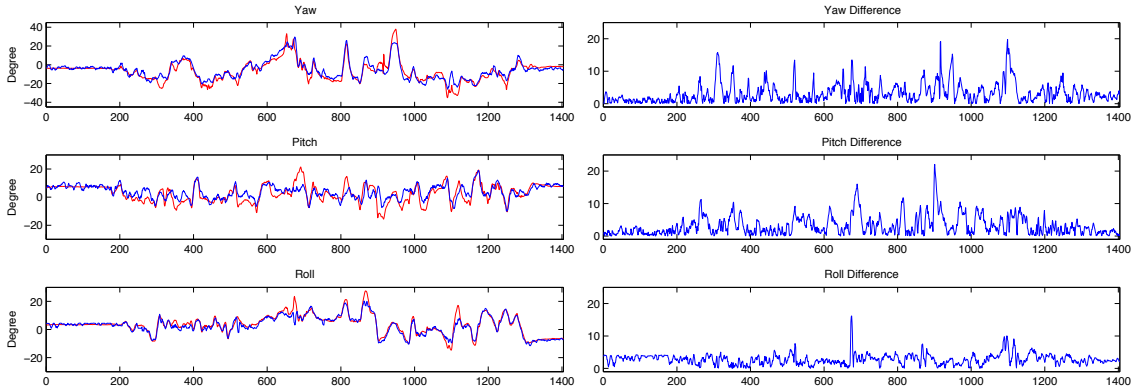


Figure 3: Left: Estimated pose angles from a sign language video in blue and ground truth angles from motion capture in red. Right: Absolute difference between the visually estimated angles and the motion capture ground truth.

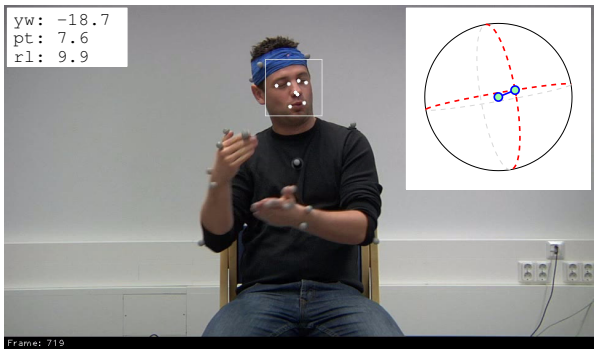


Figure 2: A frame from the motion capture video experiment with the estimated head pose angles yaw, pitch and roll. In this frame, the landmark points from flandmark are super-imposed on the signer. The headband ball markers are used in the motion capture system. Top right: Gyroscope visualization of the estimated pose.

	Correlation			Difference σ		
Model	Yaw	Pitch	Roll	Yaw	Pitch	Roll
FL+SVR A	**0.92**	0.72	**0.95**	**4.29**	4.30	**2.19**
FL+SVR B	0.85	**0.74**	**0.95**	5.55	**4.17**	**2.19**

Table 3: Correlation and standard deviation σ of the signal difference for angle estimation and motion capture data for the best trained models.

tunen et al., 2012). We considered only the four markers attached roughly symmetrically to the signer's head with a headband. The locations of these markers were used to infer ground truth values by computing the corresponding roll, pitch, and yaw angles trigonometrically.

The selected SVRs trained with data set A (FL+SVR A) had a strong correlation with the motion capture data especially for yaw (Table 3). For the pitch angle estimation, regressors trained with data set B had a slight improvement over those of set A. Roll angles show the highest correlation with the motion capture data, demonstrating the strength of the geometric approach.

In the results of Figure 3, around frames 490–510 there is a very subtle negative head shake which is captured perfectly by the yaw angle. Moreover, between frames 385–400 and 460–470 there are boundary-marking head nods, the latter of which has also an affirmative function, that are clearly identified by the pitch angle of the pose estimate. Approximately between frames 930–1150 there are several linguistically significant roll movements captured. Roll movements, together with simultaneous yaw and pitch movements, serve here to demonstrate changes in perspective from which the signer narrates the actions of the characters in the story.

4. Estimating state of facial elements

In this section, we present details of the proposed method for estimating eyebrow position, eye openness, and mouth state. The method is based on the construction of geometric features computed from an extended set of facial landmarks. The landmark detection algorithm employs an ensemble of techniques for each facial element. The extended set of landmarks is intended to determine the position of eyebrows, eyelids, and upper and lower lip boundaries which are not part of the flandmark output. For comparison we also consider landmarks detected using the Supervised Descent Method (SDM) implemented in the IntraFace library (Xiong and De la Torre, 2013). The best landmark algorithms are combined into a model, and qualitative analysis of the annotations produced by the system is performed on randomly selected videos.

The proposed facial state categorization utilizes quantized states for eyebrow position, eye openness, and mouth state. The states are categorized in *absolute* or *progressive* types: absolute states are binary and can be defined as either open or closed whereas progressive states include intermediate steps between the open and closed states (Table 4).

4.1. Landmark detection

In this section we detail the two different methods for landmark detection: the first is the proposed Landmark Ensemble Method (LEM), and the second is SDM. In both cases the extended landmark set consists of 22 points (Figure 1b).

	Eyebrow	Eye	V Mouth	H Mouth
Absolute	0:neutral	0:closed	0:closed	1:neutral
	1:shifted	1:open	1:open	2:shifted
Progressive	0:down	0:closed	0:closed	0:relaxed
	1:neutral	1:squint	1:open	1:narrow
	2:raised	2:open	2:wide	2:wide
		3:wide		

Table 4: Categorization values for each facial element.

4.1.1. Ensemble method

The LEM algorithm requires an initial estimate of the facial element area to compute the landmarks. This area estimate does not need to be exact, but it must contain the facial element studied. The approximate location of facial elements is obtained from the area surrounding the geometric center of the right and left eye landmarks from the flandmark detector, and similarly for the mouth.

To minimize the influence of shadows, the gray-scale eye image area is processed with an illumination invariant filter, in this case Single-Scale Retinex (SSR) (Jobson et al., 1997). Non-skin pixels are eliminated with a simple skin color filter model.

Eyebrow landmarks The horizontal separation limit of eyebrow and eye is the global maximum between the two lower local minimums of the vertical projection of the obtained image. Given the separation limit, the eye area is divided in two parts: the eyebrow RoI and the eye RoI. The darkest eyebrow pixel is obtained from the global maximum of the horizontal projection of the eyebrow RoI. From the estimated eyebrow seed location, a 1×3 window is used to form a path of pixels with the lowest intensity difference towards the left and right edges of the image. A cumulative sum of the intensity values in the estimated eyebrow path is computed towards both edges and scaled to the $[0, 1]$ range. Based on the available training data the center-most landmark point of the eyebrow resides where the cumulative sum exceeds 0.35. The outermost eyebrow point is similarly found at the cumulative sum value of 0.45.

Eye landmarks The eye landmark estimation starts by using a *radial symmetry* transform (Timm and Barth, 2011) to identify the pupil. The transform takes an image as input, computes the vertical gradients, and evaluates all pixels as potential centers of radial shapes. The output of the transform consists of a matrix of values indicating how likely each pixel is of being surrounded by a radial pattern.

Iris and pupil pixels appear darker and show narrower intensity value distribution than skin pixels. Our interest is then only the intensity changes from low to high to focus on dark radial patterns. Therefore, we threshold the search space to the lowest 10% pixel intensities of the image.

Following the location of the pupil, eye corners are computed using oriented projections. The eye RoI is divided in two subregions delimited by the horizontal location of the pupil. Within each of the subregions, the eye corner is estimated as the global maximum of the oriented projections.

Mouth landmarks Mouth landmark estimation is based on a color transformation by means of *pseudo hue* variations. All mouth RoIs are preprocessed with the *gray world* algorithm (Finlayson et al., 1998) for color normalization. Two color components are used: the *pseudo hue* component H, and the luminance component L.

The luminance component L from the LUX color space (Liévin and Luthon, 2004) is used in order to take advantage of the shadows produced by the mouth and improve estimated lip boundaries. The relative luminance in the image can be computed from the RGB channels as:

$$L = (R+1)^{0.6}(G+1)^{0.3}(B+1)^{0.1} - 1. \quad (3)$$

The component H takes advantage of the red and green pixel value difference between lip and skin colors. H is computed by an approximation of the component U from the LUX color space such that $H \approx U$:

$$H = \begin{cases} G/R & \text{if } R > G, \\ 1 & \text{otherwise.} \end{cases} \quad (4)$$

Following (Stillittano et al., 2013) we combine the information of the vertical gradients of H and L as follows (H and L are scaled to the $[0, 1]$ range):

$$R_{\text{top}} = \nabla_y (H - L) \quad (5)$$

$$R_{\text{mid}} = (\nabla_y H)L \quad (6)$$

$$R_{\text{low}} = \nabla_y H \quad (7)$$

In the mid and low image gradients we ignore values greater than zero (changes from dark to light), this is represented as R^*. A combined edge image R is constructed from the set of gradient images and scaled to the $[0, 1]$ range as:

$$R = R_{\text{top}} - R^*_{\text{mid}} - R^*_{\text{low}} \quad (8)$$

A lip mask is computed from H, and a second one from R. In both masks, post-processing steps are applied. Morphological closing of disk of size 3×3 is used to connect marginally separated regions. An oval mask with its axes aligned to the mouth RoI edges is used; pixels outside the oval mask are eliminated as lip pixel candidates. Connected components with size less than 10% of the total lip candidate pixels are ruled out, as well as those connected to the image border. Landmarks are finally estimated from the horizontal and vertical projections of the lip masks from R and H respectively.

4.1.2. Appearance-based method

The appearance based method used in this work is the *Supervised Descent Method* (SDM) (Xiong and De la Torre, 2013), a face alignment algorithm provided by the IntraFace software package. During training the SDM algorithm learns a sequence of optimal descent directions with a supervised approach. The optimal descent directions are computed using SIFT features (Lowe, 1999) extracted from known landmark locations at sampled images. We use only a subset of the landmarks available in IntraFace.

4.2. Geometric features

In this section a geometric feature set is proposed for estimating the facial states from previously detected facial landmarks. The features describe several geometrical properties of the eyebrow, eye, and mouth. These features are

post-processed to reduce the observed noise using PCA. The PCA-processed feature vector has the dimensionality of 10: 4 for the eyebrow, 2 for the eyes, 1 for the vertical mouth, and 3 for the horizontal mouth.

Eyebrow features For eyebrow position we use features o^{B0} to o^{B4} from (Araujo et al., 2012). The features measure the distance between eyebrows, distance between eyebrow corners and eye corners, eyebrow slope, and area of the eyebrow region. Additionally, we propose the eyebrow feature o^{B5} that uses the eye center as a reference point. Features o^{B1} to o^{B5} are computed for both left and right eyebrows, leading to a total of 11 eyebrow features. With w_b^l (w_b^r) the width and h_b^l (h_b^r) the height of the left (right) eyebrow, the features are computed for the left eyebrow as:

$$o^{B0} = ||\boldsymbol{b}_5^r - \boldsymbol{b}_1^l|| \tag{9}$$

$$o^{B1} = ||\boldsymbol{b}_1^l - \boldsymbol{e}_1^l|| \tag{10}$$

$$o^{B2} = ||\boldsymbol{b}_5^l - \boldsymbol{e}_2^l|| \tag{11}$$

$$o^{B3} = (b_{5y}^l - b_{1y}^l)/(b_{5x}^l - b_{1x}^l) \tag{12}$$

$$o^{B4} = w_b^l h_b^l \tag{13}$$

$$o^{B5} = \frac{\|e_{\mu y}^l - \rho e_{\mu x}^l - (b_{\mu y}^l - \rho b_{\mu x}^l)\|}{\sqrt{\rho^2 + 1}} \tag{14}$$

here ρ is the slope of the face with respect to the horizon, points $\boldsymbol{e}_\mu^l$ and $\boldsymbol{b}_\mu^l$ are the mean of the landmark coordinates of the left eye corners and eyebrow respectively. The slope ρ is estimated using the mouth corners. Features o^{B1}, o^{B2}, and o^{B5} are scaled according to the average feature value of the first five video frames.

Eye features Using the extended landmarks (Figure 1b) the eye openness feature is defined as (independently for each eye):

$$o^E = h_e/w_e \tag{15}$$

where $h_e = ||\boldsymbol{e}_4 - \boldsymbol{e}_3||$, $w_e = ||\boldsymbol{e}_2 - \boldsymbol{e}_1||$ and $||\cdot||$ stands for the Euclidean distance.

Mouth features The mouth features use the landmarks that define the lip shape as:

$$o^{Mw} = w_m/w_{m0} \tag{16}$$

$$o^{M1} = h_{m1}/w_m \tag{17}$$

$$o^{M2} = h_{m2}/w_m \tag{18}$$

with $w_m = ||\boldsymbol{m}_2 - \boldsymbol{m}_1||$, $h_{m1} = ||\boldsymbol{m}_3 - \boldsymbol{\mu}_{w_m}||$ and $h_{m2} = ||\boldsymbol{m}_4 - \boldsymbol{\mu}_{w_m}||$, where $\boldsymbol{\mu}_{w_m}$ is the geometric center of the two landmarks describing the mouth corners. Here w_{m0} represents the average w_m of the first five video frames. We also include features from (Tang and Deng, 2007):

$$o^{M3} = w_m/(h_{m1} + h_{m2}) \tag{19}$$

$$o^{M4} = h_{m1}/h_{m2} \tag{20}$$

4.3. Experiments

The performance of our facial element state estimators is evaluated in a quantitative and qualitative type of experiments. In the first experiment we manually annotated the facial states in videos taken from the SUVI dictionary of Finnish Sign Language (Suvi, 2003). The annotations were performed frame-by-frame on basis of the visual appearance of the isolated frame. For the qualitative experiments we compare our results with linguistic annotations prepared for a subset of the SUVI material.

	Eyebrow			Eye			
	0	1	2	0	1	2	3
Train	39	258	41	26	50	229	33
Test	42	1275	365	135	280	1079	188

Table 5: Distribution of annotated video frames for eyebrow and eye states. See Table 4 for the explanation of the states

	V Mouth			H Mouth		
	0	1	2	0	1	2
Train	228	77	33	240	14	84
Test	1034	487	161	1191	137	273

Table 6: Distribution of annotated video frames for mouth states. See Table 4 for the explanation of the states

We use the Naive Bayes (NB) probabilistic classifier and the Support Vector Machine (SVM) classifier for the experiments. The NB classifier uses the Gaussian density function for the likelihood estimation, while a Gaussian radial basis function (RBF) is used as the kernel for the SVM. The SVM implementation used for the experiments is provided in the LIBSVM package (Chang and Lin, 2011).

4.3.1. Data

The video data used consists of a set of selected video captures of signed sentences from SUVI, where linguistic analysis is available for the selected videos (Jantunen, 2007). Three video sequences were used for training, and twelve were used for testing (Tables 5 and 6).

4.3.2. Performance measure

The performance of the classifiers is evaluated using the Matthew's Correlation Coefficient (MCC) (Powers, 2011). MCC provides a good single measurable result whereas using other performance metrics would have required per-class analysis of each test. We take into consideration the distribution of the MCC coefficients, as well as their variances. This is achieved using *box-and-whisker* diagrams with median, 25th and 75th percentiles and 99.3% boundaries for graphic evaluation.

4.3.3. Results

The MCC box and whiskers plots (Figure 4) show that the performance difference of the classifiers between the LEM and SDM algorithms is small for all the facial elements. The eye and vertical mouth annotations display strong results while eyebrow and horizontal mouth are relatively weak, nevertheless the correlation is above 0.25 in most measurements. The variation of the results between the test videos suggests that the estimates are noisy.

For the qualitative evaluation, the SDM landmarks and best classifiers (NB for eyebrow and vertical mouth, and SVM for eye and horizontal mouth) were used. A timeline plot was generated to show each facial element's activity in the

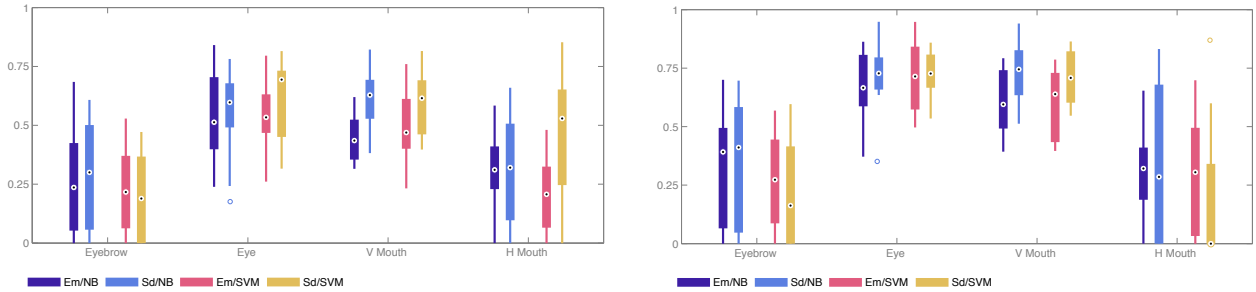
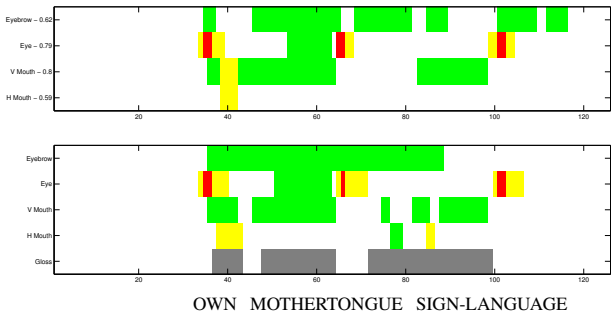


Figure 4: Classification performance: MCC distributions in (left) multiclass and (right) two-class configurations.

(a)



(b)

Figure 5: SUVI video 051703 'My mother tongue is sign language' with SDM landmarks. (a) Frame 51 with overimposed symbols representing estimated states. (b) Top: timeline representation of estimations. Bottom: ground truth annotations. Colors as in Table 7. Gray=sign gloss.

	Eyebrow	Eye	V Mouth	H Mouth
Red	lowered	closed		
Yellow		squint	wide	narrow
White	neutral	open	closed	relaxed
Green	raised	wide	open	wide

Table 7: Color coding of quantized facial states.

tested videos. A median filter of 5 frames of length has been applied to remove noisy detections.

In the included example (Figure 5), the eyebrow estimations coincided with the linguistic annotations except in the non-linguistic visual changes or perspective illusions (head tilting) visible in frames 103–117. The fading-out phase of the raised eyebrows in frames 72–88 is not deemed linguistically significant, but is still detected. For the eye openness estimations, the blinks are correctly detected around frames 38, 64 and 102, and the same holds for the widening of the eyes in frames 56–62. The mouth MCC is high in the vertical movements, activity was detected from frame 37 to 63, but the section in frames 83–98 showing open lips with closed jaw was only partially detected. The horizontal mouth movement estimation detected activity in frames 39–43, however latter frames were not detected.

5. Conclusions

In this work, head pose estimation was proposed using a model-based approach aiming at analysis and interpretation of SL videos. Facial landmark locations, face bounding box coordinates, and skin mask areas were used as features. Head pose estimation was applied in an experiment showing strong correlation of the estimated angles with SL motion capture ground truth data. We also considered a classification scheme for the position of the eyebrows, openness of the eyes, and mouth state. Geometric properties of facial landmarks were used as features. Our algorithm showed comparable results against the SDM landmark detector. The facial state estimates can be regarded useful enough for linguistic studies of eye and mouth vertical openness, further work is required for eyebrow estimation.

Our results suggest that, in the future, these methods may be used, for example, for quantitative studies of phonetics of sign languages and to aid annotation of non-manual activity in videos containing natural signing. Future work will be focused on increasing the estimation range for head pose, and the performance of eyebrow position estimates.

6. Acknowledgment

This work has been funded by the following grants of the Academy of Finland: 140245, Content-based video analysis and annotation of Finnish Sign Language (CoBaSiL), 251170, Finnish Centre of Excellence in Computational Inference Research (COIN), 34433, Signs, Syllables, and Sentences (3BatS), and 269089, The aspects of the Grammar and Prosody of FinSL (ProGram).

7. References

Araujo, R., Miao, Y.-Q., Kamel, M. S., and Cheriet, M. (2012). A fast and robust feature set for cross individual

facial expression recognition. In *Computer Vision and Graphics*, pages 272–279. Springer.

Campr, P., Dikici, E., Hruz, M., Kindiroglu, A., Krnoul, Z., Ronzhin, A., Sak, H., Schorno, D., Akarun, L., Aran, O., et al. (2010). Automatic fingersign to speech translator. *Proceedings of eNTERFACE*.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Dreuw, P., Forster, J., Gweth, Y., Stein, D., Ney, H., Martinez, G., Llahi, J. V., Crasborn, O., Ormel, E., Du, W., et al. (2010). Signspeak–understanding, recognition, and translation of sign languages. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23.

Finlayson, G. D., Schiele, B., and Crowley, J. L. (1998). Comprehensive colour image normalization. In *Computer VisionECCV'98*, pages 475–490. Springer.

Gómez-Mendoza, J.-B. (2012). *A contribution to mouth structure segmentation in images aimed towards automatic mouth gesture recognition*. Ph.D. thesis, L'institut national des sciences appliquées de Lyon.

Gourier, N., Hall, D., and Crowley, J. (2004). Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*.

Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500.

Jantunen, T., Burger, B., De Weerdt, D., Seilola, I., and Wainio, T. (2012). Experiences from collecting motion capture data on continuous signing. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*, pages 75–82, Istanbul, Turkey.

Jantunen, T. (2007). The equative sentence in finnish sign language. *Sign Language & Linguistics*, 10(2):113–143.

Jobson, D. J., Rahman, Z.-u., and Woodell, G. A. (1997). Properties and performance of a center/surround retinex. *Image Processing, IEEE Transactions on*, 6(3):451–462.

Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J., and Jantunen, T. (2014). SLMotion: An extensible sign language oriented video analysis tool. In *Proceedings of the Nine International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).

Liévin, M. and Luthon, F. (2004). Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video. *Image Processing, IEEE Transactions on*, 13(1):63–71.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., and Neidle, C. (2013). Recognizing eyebrow and periodic head gestures using CRFs for non-manual grammatical marker detection in ASL. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee.

Metaxas, D., Liu, B., Yang, F., Yang, P., Michael, N., and Neidle, C. (2012). Recognition of nonmanual markers in american sign language (ASL) using non-parametric adaptive 2d-3d face tracking. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Murphy-Chutorian, E. and Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626.

Pfau, R. and Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. *Sign languages (Cambridge Language Surveys)*, pages 381–402.

Picot, A., Charbonnier, S., Caplier, A., and Vu, N.-S. (2012). Using retina modelling to characterize blinking: comparison between EOG and video analysis. *Machine Vision and Applications*, 23(6):1195–1208.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Smola, A. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Stillittano, S., Girondel, V., and Caplier, A. (2013). Lip contour segmentation and tracking compliant with lip-reading application constraints. *Machine Vision and Applications*, 24(1):1–18.

Suvi. (2003). Suomalaisen viittomakielen verkkosanakirja [the online dictionary of FinSL]. Kuurojen Liitto ry.

Tang, F. and Deng, B. (2007). Facial expression recognition using aam and local facial features. In *Natural Computation, 2007. ICNC 2007. Third International Conference on*, volume 2, pages 632–635. IEEE.

Timm, F. and Barth, E. (2011). Accurate eye centre localisation by means of gradients. In *VISAPP*, pages 125–130.

Uřičář, M., Franc, V., and Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. In *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, volume 1, pages 547–556. SciTePress — Science and Technology Publications.

Whitehill, J. and Movellan, J. R. (2008). A discriminative approach to frame-by-frame head pose tracking. In *FG*, pages 1–7. IEEE.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.

Addressing the Cardinals Puzzle: New Insights from Non-Manual Markers in Italian Sign Language

Lara Mantovan[1], Carlo Geraci[2], Anna Cardinaletti[1]
Ca' Foscari University of Venice[1], Institut Jean-Nicod CNRS[2]
Address: Lara Mantovan, Dorsoduro 1075, Fondamenta Tofetti, 30123 Venezia
Email: laramantovan@unive.it, carlo.geraci76@gmail.com, cardin@unive.it

Abstract

This paper aims at investigating the main linguistic properties associated with cardinal numerals in LIS (Italian sign language). Considering this issue from several perspectives (phonology, prosody, semantics and syntax), we discuss some relevant corpus and elicited data with the purpose of shedding light on the distribution of cardinals in LIS. We also explain what triggers the emergence of different word/sign orders in the noun phrase. Non-manual markers are crucial in detecting two particular subcases.

Keywords: cardinal numerals, nonmanuals, Italian sign language, noun phrases, sign order

1. Background

In this paper we focus on cardinal numerals functioning as modifiers in the nominal domain and expressing a certain quantity. The cardinal system in Italian sign language (LIS) uses both hands and is a base-10 system.

In this respect, the distribution of cardinals in LIS reveals a puzzling picture. On the one hand, recent corpus data from 162 LIS signers reveal that in spontaneous narratives the majority of cardinals appears before the noun (Mantovan & Geraci, 2013), as reported in Table 1.

Word order	n	%
Card > N	278/353	79%
N > Card	75/353	21%

Table 1: Distribution of cardinal numerals in corpus data

On the other hand, the existing literature claims that cardinals are consistently or even exclusively postnominal (Bertone, 2007; Branchini, 2007; Cecchetto, Geraci & Zucchi, 2009; Brunelli, 2011). An example from Bertone (2007) is reported below for expository purposes.

[Bertone, 2007:84]

(1) BOOK NEW TWO DEM MINE
'These two new books are mine.'

Why do we observe such an important difference between corpus data and elicited data? In what respect is Card>N different from N>Card (and vice versa)? In the remainder of the paper we will offer an explanation for these two newly discovered puzzles. Our working hypothesis is that part of the sign order variability is due to the definite/indefinite character of the noun phrase, that is marked both by prosodic (i.e. non-manual) features and sign order manipulation.

2. Methods

The data for this study mainly come from the LIS corpus (Geraci et al., 2011). The annotated cardinals amount to 353 tokens. Additional data have been collected through picture-based narration tasks and elicitation of grammaticality judgments.

The materials used as stimuli for the picture-based narration tasks are wordless comic strips illustrated by Plauen (2000). Plauen's illustrations are generally self-explanatory and do not give rise to interlinguistic influences since they do not contain any written text.

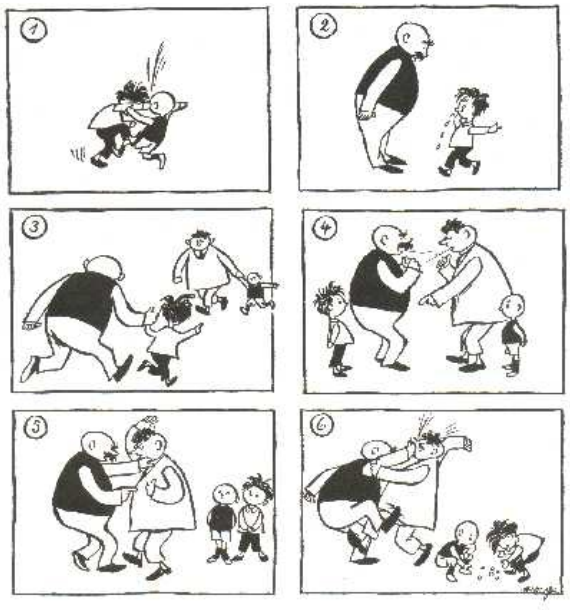


Figure 1: Extract from "Vater und Sohn" (Plauen 2000)

The story represented in Figure 1 is interesting because it triggers the production of cardinal TWO in two different contexts. In the first panel two children are represented for the very first time. Being first-mentioned referents, they are expected to be introduced in the discourse by an

indefinite noun phrase. On the contrary, the two children represented in the fifth panel are pre-established referents, therefore they are expected to be referred to by using a definite noun phrase.

Data annotation has been conducted by using the annotation software ELAN (Johnston & Crasborn, 2006). Manual and non-manual features have been carefully annotated on separate tiers. The coding scheme associated to the non-manual markers (NMMs) relevant for this study is illustrated in (2). The duration of NMMs has been measured as the time interval intervening between start and end points.

(2) a. NM-Head: left, right, raised, down, forward, back
b. NM-Eyebrows: lowered, raised
d. NM-Body: left, right, down, forward, back
e. NM-Eyes: blink, squint, close, wide, track-hands eye-gaze

To illustrate how ELAN has been used for data annotation, a representative screenshot is shown in Figure 2.

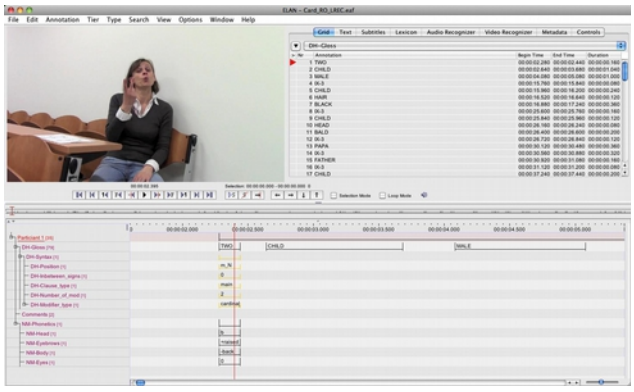


Figure 2: ELAN screenshot

Finally, grammaticality judgments have been elicited from three native signers of LIS (Rosella Ottolini, Gabriele Caia and Mirko Santoro), whom we thank enormously.

3. Results

A deeper investigation of the distribution of cardinals as emerging from the LIS corpus (see Table 1 above) revealed the presence of a confounder, namely the potentially ambiguous status of the sign ONE, and the special behavior of a subclass of cardinals, namely the ones contained in measure phrases. We discuss each of them in turn.

3.1. The sign ONE

Similarly to "uno/una" in Italian, the LIS sign ONE is ambiguous between a cardinal and an indefinite determiner. In our corpus, ONE mainly occurs in prenominal position (almost 90% of the cases) irrespectively of the syntactic/semantic function. The distribution of determiner ONE and cardinal ONE can be observed in examples (3) and (4), respectively (see also Figure 3 and Figure 4).

Corpus data (middle-aged signer from Rome)

(3) ONE MATE SCHOOL IX-3_POSS IX-3 JEALOUS STRONG
'A schoolmate of mine was extremely jealous.'

Corpus data (middle-aged signer from Rome)

(4) REFECTORY EAT FINISHED, REFECTORY ARRANGE TURN, ONE WEEK IX-1, THEN WEEK IX-3
'After we finished eating at the refectory, we took turns arranging things, one week it was my turn, then it was someone else's turn.'

As originally suggested by Bertone (2007), NMMs help distinguish the two functions. Figure 3 shows the facial expressions associated with determiner ONE in sentence (3). The most remarkable features are backward-tilted head and raised eyebrows.

Figure 3: ONE as indefinite determiner

Figure 4 shows the realization of cardinal ONE in sentence (4). In this latter case, eyebrows are in neutral position and the head is not backward tilted.

Figure 4: ONE as cardinal numeral

Once the occurrences of ONE are removed from the counting, we obtain the distribution represented in Table 2.

Word order	n	%
Card > N	184/252	73%
N > Card	68/252	27%

Table 2: Distribution of cardinal numerals without ONE

3.2. Cardinals within Measure Phrases

Let's now turn to the special case of cardinals included in measure phrases referring to time, capacity, weight, length, temperature, currency (e.g. SIX WEEK, SEVENTY KILOGRAM, THIRTY KILOMETER). According to corpus data, they display a categorical distribution: they always precede the noun, as shown in examples in (5) and (6).

Corpus data (young signer from Lamezia)

(5) NOW IX-3_POSS WIFE PREGNANT FIVE MONTH
'Now my wife is five months pregnant.'

Corpus data (old signer from Florence)

(6) HOUSE NEAR, FOUR-HUNDRED METER IX-3
'The house is in the neighborhood, about four hundred meters away.'

This piece of data has been confirmed by grammaticality judgments, as exemplified in (7).

(7) a. IX-1 REPEAT++ TWO-HUNDRED-THOUSAND TIME
'I repeated it two hundred thousand times.'
b. * IX-1 REPEAT++ TIME TWO-HUNDRED-THOUSAND

Without considering these two special cases, the distribution of cardinals, shown in Table 3, looks considerably different from the ones reported in Table 1 and Table 2. As a result, the percentage of postnominal cardinals becomes more prominent and it is now perfectly balanced with prenominal cardinals.

Word Order	n	%
Card > N	67 /135	50%
N > Card	68 /135	50%

Table 3: Distribution of cardinal numerals without ONE and measure phrases

The picture that emerges is even more intricate, showing an apparently uncontrolled variability. We now turn to the narration tasks and grammaticality judgment elicitation in order to address this issue.

3.3. The distribution of cardinals

The data collected during the narration tasks and elicitation reveal that the position of cardinals may be influenced by information structure. New-discourse information (e.g. first-mentioned referents) can be conveyed by both orders (i.e. Card>N and N>Card), whereas old-discourse information (i.e. already-mentioned referents) is compatible with N>Card only. The former is illustrated in the first panel of the comic strip, shown here in Figure 5; the latter in the fifth panel, shown here in Figure 6.

Figure 5: First-mentioned referents (new-discourse information)

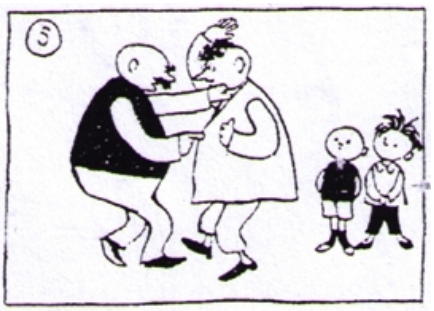


Figure 6: Already-mentioned referents (old-discourse information)

When the children are first mentioned we observe both orders Card>N and N>Card, while in further mentioning only the N>Card order is found.

This is further confirmed by the informants' assessment of their own productions. When explicitly asked about the order possibility in the two distinct contexts, only the new-information situation allows for the two sign order options, as exemplified in (8). On the contrary, in the old-discourse context only the N>Card order is possible, as illustrated in (9).

(8) New-information context
a. TWO CHILD
b. CHILD TWO
'Two children'

(9) Old-information context
a. * TWO CHILD
b. CHILD TWO
'The two children'

It is worth noting that the relative order of cardinals with respect to the noun is not crucial to distinguish the two discourse functions, as the sequentially identical data in (8)b and (9)b demonstrate. Rather, we found that it is the NMM component that plays a crucial role here. If the signer is dealing with a new referent, the prenominal or postnominal cardinal is usually accompanied by backward-tilted head and raised eyebrows (see Figure 7).

Figure 7: TWO as new-discourse information

If the referent has already been mentioned in the discourse, then the postnominal cardinal is compatible with squinted eyes and/or lowered eyebrows (see Figure 8).

Figure 8: TWO as old-discourse information

We tentatively associate the new/old discourse information with the [±definite] character of the noun phrase. Interestingly, when the noun phrase is new information, it is introduced by the same NMMs as indefinite ONE and the prenominal syntactic position is available for cardinals.

From a syntactic point of view, in the spirit of Cardinaletti and Giusti (2006), the former cardinal functions as a proper quantifier, whereas the latter, being compatible with a definite environment, should be rather considered as a quantity adjective.

4. Conclusions

In this study we combined both quantitative and qualitative data with the purpose of capitalizing on the advantages of each source. When analyzing cardinals in LIS, two special cases (i.e. ONE and cardinals within measure phrases) need to be examined separately. Syntactic positions and, most importantly, NMMs convey crucial information on the definite or indefinite nature of the nominal expression containing cardinal numerals.

5. References

Bertone, C. (2007). La struttura del sintagma determinante nella Lingua dei Segni Italiana (LIS). Ph.D. Thesis in Linguistics. University Ca' Foscari of Venice.

Branchini, C. (2007). On Relativization and Clefting in Italian Sign Language (LIS), Ph.D. Thesis in Linguistics. University Ca' Foscari of Venice.

Brunelli, M. (2011). Antisymmetry and Sign Languages. A comparison between NGT and LIS, Utrecht: LOT Publications.

Cardinaletti, A. & Giusti, G. (2006). The syntax of quantified phrases and quantitative clitics. In M. Everaert, H. van Riemsdijk (eds.), The Blackwell Companion to Syntax, Oxford: Blackwell, vol. 5, pp. 23-93.

Cecchetto, C., Geraci, C. & Zucchi, A. (2009). Another way to mark syntactic dependencies: The case for right-peripheral specifiers in sign languages. In Language, 85:2, pp. 278-320.

Geraci, C., Battaglia, K., Cardinaletti, A., Cecchetto, C., Donati, C., Giudice, S. & Mereghetti, E. (2011). The LIS Corpus Project. A Discussion of Sociolinguistic Variation in the Lexicon. In Sign Language Studies, 11:4, pp. 528-574.

Johnston, T. & Crasborn, O. (2006). The use of ELAN software annotation software in the creation of sign language corpora. Presentation at the E-MELD workshop on digital language documentation. Michigan State University in East Lansing, Michigan, US, June 2006.

Mantovan, L. & Geraci, C. (2013). A round trip from Theory to Corpus. The Case of Universal 20 in LIS. Poster presented at the TISLR 11 Conference, London, July 2013.

Pfau, R. & Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. In D. Brentari (ed.), Sign Languages, Cambridge: Cambridge University Press, pp. 381-402.

Plauen, E. O. (2000). Vater und Sohn. In Politische Karikaturen, Zeichnungen, Illustrationen und alle Bildgeschichten “Vater und Sohn” (Gesamtausgabe), Konstanz: Südverlag GmbH.

Analysis for Synthesis: Investigating Corpora for Supporting the Automatic Generation of Role Shift

John McDonald, Rosalee Wolfe, Robyn Moncrief, Souad Baowidan

DePaul University School of Computing

243 S. Wabash Ave, Chicago IL 60604 USA

E-mail: jmcdonald@cs.depaul.edu, wolfe@cs.depaul.edu, rkelley5@mail.depaul.edu, sbaowida@mail.depaul.edu

Abstract

In signed languages, role shift is a process that can facilitate the description of statements, actions or thoughts of someone other than the person who is signing, and sign synthesis systems must be able to automatically create animations that portray it effectively. Animation is only as good as the data used to create it, which is the motivation for using corpus analyses when developing new tools and techniques. This paper describes work-in-progress towards automatically generating role shift in discourse. This effort includes consideration of the underlying representation necessary to generate a role shift automatically and a survey of current annotation approaches to ascertain whether they supply sufficient data for the representation to generate the role shift.

Keywords: sign language synthesis, avatar technology, corpus annotation guidelines, role shift

1. Introduction

In signed languages, role shift is a process that can facilitate the description of statements, actions or thoughts of someone other than the person who is signing. It is an important structure in many signed languages, and thus sign synthesis systems must be able to portray it. Animation is only as good as the data used to create it, which is the motivation for using corpus analyses when developing new tools and techniques. This paper describes work-in-progress towards automatically generating role shift in discourse. In order to complete this effort, we need to address two questions:

- What underlying representation is necessary to generate a role shift automatically?
- Can current corpora supply sufficient data for the representation to generate the role shift?

2. Linguistic theory

Role shift has been a topic of study in signed language linguistics almost since the inception of the discipline. This section is a condensed review of the history of linguistic theory concerning role shift. For a more comprehensive treatment, see (Lillo-Martin, 2012).

Friedman (1975) observed that when reporting a dialog in American Sign Language (ASL), a signer can designate a protagonist via a third-person referent and then assume the role of that protagonist. Analyzing the phenomenon further, Liddell & Metzger (1998) noted that a role shift in ASL could convey constructed action as well as thoughts or dialog, and introduced the concept of "mental spaces" as a framework to account for constructed action.

Morgan (1999) described a framework of three spaces in British Sign Language (BSL). The first, *narrator space*, was used by signers to introduce protagonists and plot motivation. The second, *fixed referential framework*, accounted for establishing scenes involving topographic space and setting up pronominal points toward spatial loci. Once these loci have been designated, the signer can exploit them to form agreement verbs. This space interacts with the third framework, called the *shifted referential framework*, which is used to describe dialog, actions, and thoughts of the protagonists. When performing a role shift, the signer uses the shifted referential framework, but can still make use of other loci previously designated in the fixed referential framework (Figure 1). Thus the spaces interact during discourse.

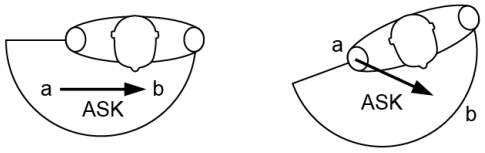


Figure 1: Fixed and Shifted Referential Frameworks.

When considering the depiction of objects and events in ASL, Dudis (2004) further explored the concept of interacting spaces. He noted that different spaces will scale (size) the depictions differently. He used an example of a motorcyclist climbing a hill. When the signer assumes the role of the motorcyclist gripping the handles, the motorcycle is life-sized. However, when the signer uses a vehicle classifier to show the steep slope of the road, the motorcycle shrinks to the size of the signer's hand. Further, only the signer's hand portrays the vehicle classifier while the rest of the signer's body is still riding the motorcycle. Thus the two spaces interact, in what he called a *blend*.

In a study of spatial coherence in German Sign Language (DGS), Perniss (2007) introduced the terms *observer perspective* and *character perspective* to describe the two spaces and to motivate the types of scaling. Observer perspective is analogous to having an imaginary camera set sufficiently far away with a field-of-view wide enough to encompass the entire space. Since the entire space is visible through the imaginary viewfinder, the depicted distances between entities are small. On the other hand, in character

perspective, the signer assumes the role of a previously-designated protagonist. In this space, an imaginary camera would have the same view and perspective as the protagonist, and the distances between objects would be much closer to life-sized.

The metaphor of a camera is also useful when discussing Janzen's research (2004) on space rotation and perspective shift in ASL. He described a narrated story of a police officer and a driver. Although the two characters would have been face-to-face during the incident, the narrator did not shift to assume the roles of the driver and the officer. Janzen described this as mentally rotating "their conceptualized space so that third-person referents realign with the signer's own stance. Body shifts toward a designated space do not occur."(Janzen, p. 149) In other words, the narrator was performing each protagonist as seen through a camera from the addressee's vantage point.

3. Visual indications of a role shift

Findings from linguistic theory yield a rich set of information describing the visual aspects of a role shift. They are a set of specifications, or a metric against which we can evaluate the quality of animations from signed language synthesis systems. Thus an important question to consider is, "What needs to be portrayed in an animation to convey a role shift? "

Early studies emphasize the gross motor movement of the spinal column. Friedman (1975) mentions the orientation of a signer's body or the turning of the head to distinguish one third-person referent from another. Liddell & Metzger (1998) describes the types of constructed action that can occur within a role shift (Table 1). These give the scope of the animation required.

Protagonist actions	What they indicate
Articulation of words or signs or emblems	What the protagonist says or thinks
Direction of head and eye gaze	Direction protagonist is looking
Facial expressions of affect, effort, etc.	How the protagonist feels
Gestures of hands and arms	Gestures produced by the protagonist

Table 1: Types of constructed action.

The phenomenon has been studied in many sign languages. In 2000, Cuxac presented several classes of "personal transfer" in French Sign Language (LSF) similar to role shifts in ASL (Meurant, 2004). When investigating role shift forms to convey non-direct speech in the Sign Language of Southern Belgium (LSFB) Meurant (2004) found that eye gaze is the main mechanism, rather than body leans or tilts for reference. Quer (2005) analyzed role shift in Catalan Sign Language (LSC) and made cross linguistic comparisons with studies of ASL, Italian Sign Language (LIS) and Danish Sign Language (DTS) data. The following are a list of nonmanual markings that may indicate a role shift:

- a slight body shift towards the locus of the previously-designated protagonist;
- a change in eye gaze contact from the actual to the purported addressee of the reported dialog;
- a change in head orientation;
- facial expression (linguistic and affective) associated with the protagonist.

Although this list enumerates a diverse set of nonmanual markings, Herrmann & Steinbach (2012) have found that only the change in eye gaze is obligatory for marking a role shift in DGS, and body shifts and changes in head orientation are optional.

4. Corpus studies involving role shift

The introduction of multimedia annotation tools such as iLex (Hanke, 2002) and ELAN (Crasborn & Sloetjes, 2008) and the establishment of transcription systems such as HamNoSys (Hanke, 2004) and annotation guidelines, including the ECHO conventions (Nonhebel, Crasborn, & van der Kooij, 2004) and the Auslan corpus guidelines (Johnston, 2009) have facilitated a blossoming of signed language corpus research, including investigations involving role shift. Both sets of annotation guidelines specify a role (or role/constructed action) tier. Annotations in the tier indicate start and end times of a role shift in addition to the character being assumed by the signer. Using Johnston's annotation guidelines, de Beuzeville (2008) investigated verb modification and recorded the frequency of co-occurrence of constructed action (role shift) with modified and unmodified signs.

Other researchers created customized tiers for their study of role shift but were mindful of the challenges of using consistent annotations. In their study of iconic representations (depictions), Dudis et al. (2008) developed a flowchart to guide annotation. In ELAN, they used two tiers, one to annotate instances of character perspective and another to annotate instances of event depictions in observer perspective.

Several recent studies of role shift have carefully analyzed eye gaze. While building a corpus for a cross-linguistic project investigating the signed languages of Germany, Ireland, and the Netherlands, Herrmann (2008) discovered that previously established annotation guidelines for eye gaze did not provide sufficient precision for her investigation. One of her goals was to use an annotation protocol that was as precise and as detailed as possible without ascribing to any particular theory. She proposed an approach which would accurately and continuously annotate eye gaze and blinks. This new approach opens the possibility for studying blink and gaze contribution to role shift.

The question of using theory-neutral annotations, as contrasted with those that are theory-dependent, is an ongoing issue that affects studies of role shift. The method

that Zwitserlood, Özyürek & Perniss (2008) used was to separate the coding into two sets of tiers. The analytic tiers contain theory dependent interpretations. The descriptive tiers are annotated in terms of phonetic / phonological forms only and are theory neutral. An analytic tier contains referent annotations. These are connected to annotations on descriptive tiers by virtue of their co-occurrence.

In a study of BSL, Cormier & Smith (2011) defined a set of eight tiers to study constructed action. Six of these are dedicated to forms (articulators) used to support role shift / constructed action and include tiers for eye gaze, head, face, torso, dominant arm/hand and nondominant arm/hand. The remaining two tiers specify the primary role and secondary role. For the primary role (Role1), the narrator is the default; otherwise the tier indicates the protagonist whose role the signer assumes. The second role (Role2) could be the narrator if Role1 is designating a protagonist. In this way, they can accommodate the blended spaces such as the motorcyclist story described by Dudis (2004).

For eye gaze, Herrmann and Cormier & Smith use a coding system that is similar to the ECHO guidelines (Nonhebel, Crasborn, & van der Kooij, 2004), which is reproduced in Figure 2.

r-	to the right, close to 90 degrees of MSP
r	to the right, close to 45 degrees of MSP
l-90	to the left, close to 90 degrees of MSP
l	to the left, close to 45 degrees of MSP
lh	to the left hand (for RH tier)
rh	to the right hand (for LH tier)
u	upward, higher than lexical default height
d	downward, lower than lexical default height
a	ahead, more to the front than lexical default
s	towards the signer, close to the signer
p	toward a person present

Figure 2: Coding for eye gaze, ECHO guidelines

There are four options for a lateral gaze that are not directed at the hands, two to the right and two to the left of the midsagittal plane (MSP). In contrast, as seen in Figure 3, Zwitserlood (2008) uses a streamlined scheme involving a single deviation on either side of the MSP.

These annotations for eye gaze are a good starting point for creating a computer system capable of automatically generating animations depicting role shift. The next section presents previous discussions of role shift in computer animation systems.

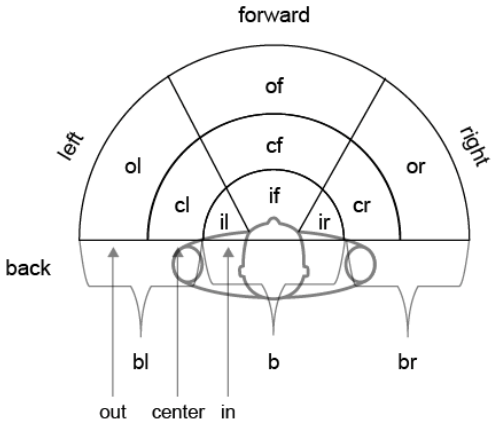


Figure 3: A 3D location grid, used facilitate specification of eye gaze direction. Zwitserlood (2008)

5. Previous efforts in synthesizing role shifts

Several research groups have included the portrayal of role shift in their animation systems. At LIMSI, Braffort and Dalle (2008) created a representation that closely reflects current linguistic theory. From an animation perspective, role shift is related to fixing referent loci and proforms. For these entities, they can accommodate characteristics such as location in signing space, orientation, shape and size, and other, syntactical (functional), semantic or cognitive features. They observe that loci for referents are always placed relative to the signer. Therefore they adopted a system of coordinates centered on the avatar and anchored on the avatar's pelvis, in order to deal with role shifts that require pelvis rotation.

In a study on improving spatial reference, Huenerfauth (2009) created "16 paragraph-length" animations that included, among other constructs, role shift. However, there was no mention of internal representation or implementation details.

The SignCom project (Duarte & Gibet, 2010a, Duarte & Gibet, 2010b) allows for the annotation of synchronized video and motion capture (mocap) data to facilitate both synthesis and analysis of LSF. For synthesis, sign data can be retrieved from different mocapped recordings and linked together via transitions created by an animation engine. The engine is capable of creating a transition that includes a role shift. This maintains discourse accuracy and comprehension.

6. Synthesizing Role Shift

From a synthesis perspective there are several problems to be solved. Our system relies on procedurally generated keys to create the basic movements of a role shift, which layer on top of signs animated as sets of sparse key-frame data. The procedures seamlessly transform the signs created in the fixed referential framework to the shifted referential framework of a constructed dialog or action.

Application of this shift is not limited to key frame data created by an animator since it layers over any previously existing avatar motion. It could also be applied to avatars

that rely on motion capture data for their base animation. All that is needed is a separate set of controls for the spine, neck and eyes that allow a procedure to add the rotations of these joints onto the existing animation. As long as the motions in the sign are not extreme, adding in the small amount of rotation in the spine necessary to shift the coordinate frame will not generally cause the spine, shoulders or arms to rotate beyond their natural motion limits.

In addition to this basic transformation rooted at the spine, the system must consider eye gaze. Per Lillo-Martin (2012), transfer of eye gaze begins a role shift, and as has been noted, role shift can be indicated entirely by eye gaze, even without the torso twist that usually accompanies it. This shift in eye gaze will depend both on the referents that have been placed previously in the fixed referential framework and on the orientation of the body in the shifted referential framework.

To create a role shift with an avatar, a synthesis system must be capable of representing frameworks of reference. As a first step towards this, we will consider the frameworks from Morgan (1999) with the goal of incorporating the interaction with co-occurring representational frameworks in the future. We will focus this discussion on the mathematical modeling and the implementation required to portray the nonmanual markings comprising the first three items of Quer's (2005) inventory, including body, gaze and head orientation.

7. Spinal column and eye gaze in role shift

The first aspect that must be modeled is the transition into the shifted referential framework for a constructed dialog. To do this, the system will need to know the protagonists and addressees in the constructed dialog and where those speakers are placed in the fixed referential framework.

For indexing and verb conjugation, our system uses a collection of four key referential points spaced radially around the avatar in the fixed referential framework (Toro, 2004). These participants are placed at angles of approximately 15º and 30º on the strong side of the avatar and 30º and 45º on the weak side, relative to the midsagittal plane. The extra angular shift on the weak side is necessary because of the twist in the torso that naturally occurs when reaching across the body to point towards these loci.

Thus, for the avatar to assume the role, the protagonist will either need to:

- Be explicitly indexed in space in the discourse, in which case the system will have positions for each protagonist predefined in the fixed referential framework as one of the loci in Figure 4.
- Be inferred by the system according to the speaking order in the constructed dialog. The system will then choose contrastive positions for the protagonists on the strong/weak sides of the body.

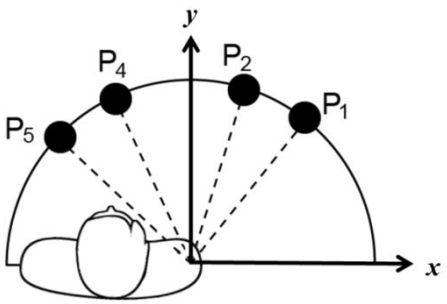


Figure 4: Loci for referents.

Given these loci, the system must manipulate the avatar to clearly indicate both the protagonist and addressee. An important aspect of this transformational model is that the avatar does not need to rigidly assume the precise position of the locus previously defined for the protagonist. All that is needed is a contrastive shift in the direction the locus sufficient to mark the transition from fixed to shifted referential framework.

The human action of turning the torso is a subtle and complex process due to the multiple participating joints involved including the pelvis, the lumbar and thoracic spinal joints, and also the sternoclavicular joints in the shoulders. In addition, the cervical spine will rotate the neck, and the eyes will shift as well. This is further complicated by the fact that, unlike most models of animation in computer graphics, these joints do not rotate in complete concert, but will cascade in a natural progression. In our motion studies of human torso movement, we found that the joints will typically begin their movement in the following sequence:

- The eyes rotate towards the direction that the body will eventually face.
- The neck rotates to facilitate the eye gaze.
- The pelvis rotates to begin the torso motion.
- The lumbar and thoracic spines follow in sequence to pull the shoulders into the desired orientation.
- The sternoclavicular joints will further rotate the shoulders to complete the transition.

Our studies of signers indicate that the eyes will turn to focus on the addressee's locus before the body rotates. In fact, this action will precede the body's rotation by up to a half second. Moreover, the actual direction that the eye gaze will assume in the role shift will depend on the addressee of the constructed dialog. Eye gaze consists of two actions that combine to orient the eyes comfortably toward a locus, namely neck and eye rotation. Ultimately, we need to take into account loci of both the protagonist and the addressee. However, when initially breaking gaze, the neck and eyes of the avatar must be turned to face the addressee, because the body has not yet turned to assume the role of the protagonist.

Since referents in signed language are indicated by direction, not by position in space, the rotation in world coordinates (fixed framework) required is precisely the angle between the addressee locus and the midsagittal

plane. Thus, the angle can be computed as

$$v' = \left(P_i - S_d\right)$$
$$v = \left\langle v_x', v_y', 0 \right\rangle$$
$$\theta = \arccos\left(\frac{v_y'}{\sqrt{\left(v_x'\right)^2 + \left(v_y'\right)^2}} \right)$$

Where P_i is the locus of referent i, and S_d is the position of the dominant shoulder. The z-coordinate is ignored here because we assume that the protagonist and the addressee are of equal height.

The actual division of this angle between the eyes and the neck will change dynamically over the course of the eye gaze shift. The eyes will move first, and then the neck will follow. As long as the angles for the eyes, neck and spine sum to θ, the eyes will maintain the proper orientation towards the addressee.

Ultimately, the rotation of the shoulders will have the dominant share of θ because they define the shifted coordinate frame. It is important to note that although the motion begins at the pelvis, it is actually the orientation of the shoulders that form the shifted coordinate frame. This action, which follows the eye and neck rotations defined above, is composed of a lean in the avatar towards the locus, and spinal column twist to orient the shoulders toward the addressee. A full discussion of this spinal algorithm can be found in McDonald, Wolfe, Schnepp, & Toro, (forthcoming).

8. Annotation to support synthesis

Both analytic and descriptive tiers are enormously valuable for synthesizing role shift. Analytic tiers give us the referent needed to synthesize narratives, while the descriptive tiers are essential for study to build the requisite mathematical models. For example, there is consensus that eye gaze contributes to marking role shift, but without analytic annotation, it is difficult to understand whether a particular eye movement coded in a descriptive tier is functioning as part of a role shift.

When generating eye gaze, sign language synthesis systems need to take into account the fact that many of the gaze codes in descriptive tiers are contrastive rather than geometrically literal. When applied literally, the codes in the ECHO conventions yield geometric interpretations of gaze that are too extreme. A "near 90º" eye gaze is difficult to produce, particularly at normal conversational speed. This is particularly true for adults -- it is not a motion that is easily performed as it requires a rotation of the neck of at least 60º with the remainder of the angle being carried by an extreme sideways glance in the eyes. This is close to the comfort limit for a human both on the neck and the eyes (Washington State Department of Social & Health Services, 2003). A total 45-60º gaze shift is more reasonable as an upper limit, and so synthesis systems should not interpret these annotations literally, but should consider the actual ranges of human motion.

However, both video and motion capture corpora can be extremely valuable for synthesis of eye gaze marking for role shift if they have certain minimal elements coded in their annotations. The protagonists in the conversation need to be identified, and if they have been specifically indexed by the signer, the referent locus for each protagonist needs to be specified in the annotation. For each role shift, both the protagonist and the intended addressee(s) need to be included in the annotation.

If these data are not supplied, then any synthesis system would be forced to estimate the best placements for the protagonists in a narrative, which could lead to inconvenient positions that yield awkward animation. Without these data, a corpus becomes less useful for building and refining procedural techniques.

9. Conclusions and future work

Efforts to synthesize role shift can benefit greatly from annotated corpora. This is true whether the synthesis uses a sparse key technique such as ours or a motion capture system such as the one described in Awad, Courty, Duarte, Le Naour, & Gibet, (2010). Motion capture utilizes large sets of captured sequences of sign that have been annotated for linguistic structure within the fixed referential framework. In contrast, the sparse key technique relies heavily on theory to make decisions on how to manipulate the keys in order to generate the shifted referential framework, and studies of corpora are essential to building the procedural algorithms.

The discussion presented here is an algorithm for producing eye gaze in role shift within a sparse-key animation system. Further study is necessary to refine the algorithm and to extend it to include facial non-manual components of role shift such as personality.

10. Acknowledgements

We thank Ronnie Wilbur and Julie Hochgesang for valuable discussions regarding the linguistic theories of role shift and insights into coding strategies.

11. References

Awad, C., Courty, N., Duarte, K., Le Naour, T., & Gibet, S. (2010). A combined semantic and motion capture database for real-time sign language synthesis. In *Intelligent Virtual Agents,*, 432-438.

Beuzeville, L. de. (2008). Pointing and Verb Modification: the expression of semantic roles in the Auslan corpus. *Workshop on the Representation and Processing of Sign Language, at the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 13-16). Marrakech, Morocco: European Language Resources Association (ELRA).

Braffort, A., & Dalle, P. (2008). Sign language applications: preliminary modeling. *Universal access in the information society,* 6 (4), 393-404.

Cormier, K., & Smith, S. (2011). Defining and annotating constructed action, constructed dialogue and role shift. *Sign Language Discourse workshop.* Göttingen.

Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, (pp. 39-43).

Duarte, K., & Gibet, S. (2010). Corpus Design for Signing Avatars. *"Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language.* Valetta, Malta.

Duarte, K., & Gibet, S. (2010). Heterogeneous Data Sources for Signed Language Analysis and Synthesis: The SignCom Project. *Seventh International Conference on Language Resources and Evaluation. 2*, pp. 1-8. Valetta Malta: ELRA.

Dudis, P. (2004). Body partitioning and real-space blends. *Cognitive Linguistics*, 15 (2), 223-238.

Dudis, P., Mulrooney, K., Langdon, C., & Whitworth, C. (2008). Annotating Real-Space Depiction. *Workshop on the Representation and Processing of Sign Language, at the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 54-57). Marrakech, Morocco: European Language Resources Association (ELRA).

Friedman, L. (1975). Space, Time and Person Reference in American Sign Language. *Language, 51*(4), 940-961.

Hanke, T. (2002). iLex-A tool for Sign Language Lexicography and Corpus Analysis. *Third International Conference on Language Resources and Evaluation (LREC'02).* Las Palmas: ELRA.

Hanke, T. (2004). HamNoSys -- Representing sign language data in language resources and language processing contexts. *Fourth International Conference on Language Resources and Evaluation (LREC'04). Representation and Processing of Sign Languages Workshop* (pp. 1-6). Paris: European Language Resources Association.

Herrmann, A. (2008). Sign language corpora and the problems with ELAN and the ECHO annotation conventions. *Workshop on the Representation and Processing of Sign Language, at the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 68-73). Marrakech, Morocco: European Language Resources Association (ELRA).

Herrmann, A., & Steinbach, M. (2012). Quotation in Sign Languages – A Visible Context Shift. In I. Alphen, & I. Buchstaller (Eds.), *Quotatives: Cross-linguistic and Cross Disciplinary Perspectives* (pp. 203-228). Amsterdam: Benjamins.

Huenerfauth, M. (2009). Improving spatial reference in American sign language animation through data collection from native ASL signers. *Universal Access in Human-Computer Interaction. Applications and Services*, 530-539.

Janzen, T. (2004). Space rotation, perspective shift and verb morphology in ASL. *Cognitive Linguistics,* 15 (2), 149-174.

Johnston, T. (2009, November). *The Auslan Corpus Annotation Guidelines.* Retrieved December 28, 2011, from www.auslan.org.au: http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf

Liddell, S., & Metzger, M. (1998). Gesture in sign language discourse. *Journal of Pragmatics, 30*, 657-697.

Lillo-Martin, D. (2012). Utterance reports and constructed action. In R. Pfau, M. Steinbach, & B. Woll (Eds.), *Sign Language: An International Handbook HSK 37* (pp. 365-387).

McDonald, J., Wolfe, R., Schnepp, J., & Toro, J. (forthcoming). A kinematic model for constructed dialog in American Sign Language. *Sixth Conference of the International Society for Gesture Studies.* San Diego.

Meurant, L. (2004). Role Shift, Anaphora and Discourse Polyphony in Sign Language of Southern Belgium (LSFB). *Signs of the time. Selected papers from TISLR (2004)*, (pp. 319-351). Barcelona.

Morgan, G. (1999). Event packaging in British Sign Language discourse. In E. Wiinston (Ed.), *Storytelling & Conversation: Discourse in Deaf Communities* (pp. 27-58). Washington, DC: Gallaudet University Press.

Nonhebel, A., Crasborn, O., & van der Kooij, E. (2004, January). *Sign language transcription conventions for the ECHO Project.* Retrieved December 30, 2011, from European Cultural Heritage Online: http://sign-lang.ruhosting.nl/echo/docs/ECHO_transcr_conv.pdf

Perniss, P. (2007). Achieving spatial coherence in German Sign Language narratives: The use of classifiers and perspective. *Lingua,* 117 (7), 1315-1338.

Quer, J. (2005). Context shift and indexical variables in sign languages. *Proceedings of SALT*, *15*.

Toro, J. (2004). Automated 3D Animation System to Inflect Agreement Verbs. *Sixth High Desert Linguistics Conference.* Albuquerque, New Mexico.

Washington State Department of Social & Health Services. (2003, January). *Range of Join Motion Evaluation Chart DSHS 13-585A (REV. 08/2002) (AC 01/2003).* Retrieved February 10, 2014, from Washington State Department of Social and Health Services: http://www.dshs.wa.gov/pdf/ms/forms/13_585a.pdf

Zwitserlood, I., Özyürek, A., & Perniss, P. (2008). Annotation of Sign and Gesture Cross-linguistically. *Workshop on the Representation and Processing of Sign Language, at the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (pp. 185-190). Marrakech, Morocco: European Language Resources Association (ELRA).

The "how-to" of integrating FACS and ELAN for analysis of non-manual features in ASL

Kristin Mulrooney, Julie A. Hochgesang, Carla D. Morris, Katie Lee
Gallaudet University
Department of Linguistics
800 Florida Avenue, NE
Washington, DC 20002
Email: Kristin.mulrooney@gallaudet.edu, julie.hochgesang@gallaudet.edu, carla.morris@gallaudet.edu, katie.lee@gallaudet.edu

Abstract
The process of transcribing and annotating non-manual features presents challenges for sign language researchers. This paper describes the approach used by our research team to integrate the Facial Action Coding System (FACS) with the EUDICO Linguistic Annotator (ELAN) program to allow us to more accurately and efficiently code non-manual features. Preliminary findings are presented which demonstrate that this approach is useful for a fuller description of facial expressions.

Keywords: American Sign Language, FACS, ELAN, non-manual features

1. Introduction

The process of transcribing and annotating non-manual features presents challenges for sign language researchers. This paper describes the approach used by our research team to integrate the Facial Action Coding System (FACS) with the EUDICO Linguistic Annotator (ELAN) program to allow us to more accurately and efficiently code non-manual features.

Since 2010, researchers in the Department of Linguistics at Gallaudet University have collaborated with avatar developers VCom3D, Inc. The most recent collaboration is part of VCom3D's *Mobile Signing Math Dictionary with Mouth Morphemes* project, which was established because "[e]xisting animations of facial expressions and speech fall short of addressing the full range of "visible speech" and mouth morphemes." The Gallaudet research team has two main tasks on this project. One, to provide feedback on avatar animations as to the accuracy and naturalness of the facial behaviors. The second, to analyze naturally produced ASL discourse in a variety of settings (classroom, narratives), and identify the most frequent range of the most commonly used facial expressions.

Initially, the team utilized a notation system which was supplied by VCom3D. This system provided to the Gallaudet team, which was accompanied by a video clip collection of a model demonstrating each label, "bundled" the actions of individual face muscles, and presents them as one unified behavior. This system used global labels such as 'surprise' or 'oo' to describe an entire facial expression. However, facial expressions, more often than not, contain more than one meaningful part conveying multifaceted information. For example, a facial expression could contain the meaning of both surprise and a WH question, or anger and a WH question. The two meaningful parts being generated via distinct muscle movements in different areas of the face.

The Gallaudet team concluded that the "bundled" approach motivated by complex meanings would not accurately describe the facial behaviors used to create a whole facial expression. This moved the team to create an alternative system, the basis of which included dividing the parts of the face (on separate ELAN tiers), describing what each of the parts did (using FACS), and when the different movements occurred (separate annotations on independent ELAN tiers). By doing so, it has enabled the team to annotate actions of the separate facial muscles, such as the eyebrows, mouth, and cheeks, as they moved independently (or dependently), and identify when those movements occur synchronously or not.

2. Facial Action Coding System

The Facial Action Coding System (FACS), by Paul Ekman, Wallace V. Friesen, and Joseph C. Hager (2002), is a system to taxonomize human facial movements by their appearance on the face. In other words, FACS is a coding system used to define groups of facial behaviors and movements on the basis of shared characteristics and giving names to those defined groups. In FACS, movements of individual facial muscles are encoded based on slight changes in outward facial appearances.

FACS allows researchers to code nearly any anatomically possible facial expression, deconstructing them into specific Action Units (AU). The temporal sequence of those segments result in unique meaningful expressions. The FACS manual defines AU as a contraction or relaxation of one or more of the facial muscles, which are constrained by physical limitations based on the muscular structure of the face and skull. These AUs, by definition, are "independent of any interpretation," which means that the code assigned to each contraction/relaxation, or combination of them, is based on form alone and not on the function or meaning of the physical behavior. This system removes subjectivity from the description of facial expressions, making the resulting transcription a more reliable source for research information.

Researchers interested in examining facial expressions used in sign languages have found FACS an effective tool. Charlotte Baker-Shenk was an early adopter and applied it to her research on questions in ASL (1983) and others have

used it since (Corina, Belludi, Reily, 1999; Dachkovsky & Sandler, 2009). It has been used with other programs to code sign languages such as SignStream. (Grossman & Shepard-Kegl 2006). In this work we are further extending the coding system by using it with another annotation tool, ELAN.

3. EUDICO Linguistic Annotator

ELAN (EUDICO Linguistic Annotator) is a time-alignable video/audio annotation tool that can be used with different transcription systems with different analytical goals (i.e., from phonetics to discourse). From the online manual, the developers explain that ELAN "is an annotation tool that allows you to create, edit, visualize and search annotations for video and audio data." Developed at the Max Planck Institute for Psycholinguistics, ELAN was designed "to provide a sound technological basis for the annotation and exploitation of multi-media recordings. ELAN is specifically designed for analysis of language, sign language, and gesture, but it can be used by everybody who works with media corpora, i.e., with video and/or audio data, for purposes of annotation, analysis and documentation".

4. Incorporating FACS in ELAN

The challenge the team faced when providing feedback to VCom3D about the avatars expression of 'natural-like' facial behaviors was that the avatars' facial expressions did not appear to be dynamic. Facial expressions are essentially dynamic, in that expressions are created by different parts of the face, and these parts move independently of one another (e.g., eyebrows are raised while the lower face remains static). The Gallaudet team required a way of demonstrating this discrepancy to VCom3D, and agreed that the best way to do so would be to provide them with 1) a more accurate description of the timing of when the parts of the face changed configuration, and 2) a shared 'language' to describe what the parts on the face did. Coding the expressions in ELAN using FACS facilitated these goals.

To begin, the team created a "No Match" dependent tier system to more accurately analyze the annotations previously identified as not matching the model examples/labels provided by VCom3D. For this process three dependent tiers were established under a parent "No Match" tier; annotations for the parent tier were created based on previous transcription of VCom3D "expression" tiers, wherein the token present in the natural data did not match the model in the VCom3D system. On the dependent tiers, the team was able to provide more detailed information about each "No Match" token. Figure 1 is an example of the ELAN template used.

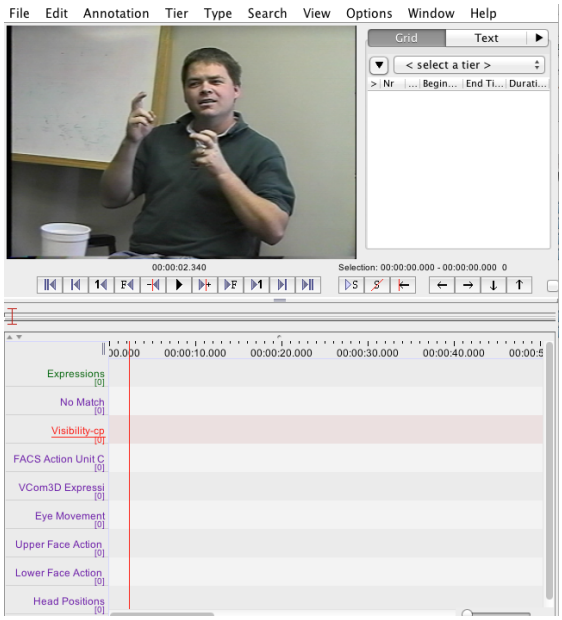


Figure 1: ELAN Template

The first dependent tier used a controlled vocabulary containing all 91 of the VCom3D "expression" labels. Tokens on this tier were coded based on the "closest approximate match" of an expression demonstrated by the VCom3D model and the sign seen in the natural data.

The second dependent tier then provided more detail about what features in the natural data did not match the closest approximate match (which were coded on the first dependent tier). The controlled vocabulary for the second dependent tier was based on material adapted from the FACS manual. Ultimately, there were 15 items in this tier's controlled vocabulary. To create this controlled vocabulary, four broad Action Units (AU) categories were adapted from the FACS manual:

Upper Face Action Units
Lower Face Action Units
Head Positions
Miscellaneous

These 4 categories comprise the initial controlled vocabulary items for this tier; however, combination items were needed for the co-occurrence of features in the data. An additional 10 items were created, from 2 and 3 Action Unit feature combinations, and were created via the use of multi-dimensional tables. Tables 1 and 2 below show the production process for these combination controlled vocabulary codes.

No Match	Upper Face Actions	Lower Face Actions	Head Positions	Miscellaneous Actions	2 features 'No Match'	
Upper Face Actions	o	x	x	x		3
Lower Face Actions	x	o	x	x		2
Head Positions	x	x	o	x		1
Miscellaneous Actions	x	x	x	o		Total of 6 combos

Table 1: Two feature combinations

No Match	Upper Face Actions	Lower Face Actions	Head Positions	Miscellaneous Actions	
Upper Face Actions & Lower Face Actions	o	o	x	x	2
Upper Face Actions & Head Positions	o	x	o	x	1
Upper Face Actions & Miscellaneous	o	x	x	o	
Lower Face Actions & Upper Face Actions (redundant)	o	o	x	x	
Lower Face Actions & Head Positions	x	o	o	x	1
Lower Face Actions & Miscellaneous (redundant)	x	o	x	o	
Head Positions & Upper Face Actions (redundant)	o	x	o	x	
Head Positions & Lower Face Actions (redundant)	x	o	o	x	
Head Positions & Miscellaneous (redundant)	x	x	o	o	
Miscellaneous & Upper Face Actions (redundant)	o	x	x	o	
Miscellaneous & Lower Face Actions (redundant)	x	o	x	o	
Miscellaneous & Head Positions (redundant)	x	x	o	o	Total of 4 combos

3 features 'No Match'

Table 2: Three feature combinations

Shown in the above tables, the boxes outlined in red highlight the single instance of 2 and 3 feature combinations (6 and 4, respectively).[1] A third table was generated, wherein 4 feature combinations were shown; however, the combinations in this table are all redundant of each other, since there are only 4 feature categories, and the 4 features co-occurring equates to a full "No Match." Thus, a 15th controlled vocabulary item was added: "Full No Match."

The third dependent tier was also designed to provide detail about which features in the natural data did not match the closest approximate match. This last tier was added to provide information about the visibility of the sign in the data, which was suspected to be the root cause of some "No Match" annotations thus far. The controlled vocabulary for this tier was adapted from the FACS manual. The FACS manual contained 4 codes in its "Miscellaneous" category, which were reallocated for use in this tier. Each of these codes relates information about the visibility of the face in the data. The FACS manual had 4 such codes:

Visibility 70 - Brows and Forehead not visible
Visibility 71 - Eyes not visible
Visibility 72 - Lower Face not visible
Visibility 73 - Unscorable

These four codes were used directly in the controlled vocabulary. Three items were added to accommodate combinations of 'hidden' features. These combinations were generated following a similar multi-dimensional table as that for the preceding tier. In order to aid interpretation of transcript results (i.e. token counts, when images of the data are not provided), an 8th item was added to code whether or not the face in the data had full visibility.

Examples of these controlled vocabularies in application are shown in figure 2 below.

[1] Table items not highlighted red are redundant occurrences of the same feature combinations.

Figure 2: Example of VCom3D model and natural data

The three tokens seen in the middle of figure 1 are examples of instances in which the "closest approximate expression match" is the same for each token (shown in upper half of figure 1, and in ELAN annotations below the tokens coded in blue). They also each match in full visibility (shown on bottom ELAN tier coded in yellow). But, the tokens' FACS tier shows that each token mismatches the VCom3D model in a different way. The left-most token mismatches in the actions of the lower face (coded in gold), particularly in the final segment of production. The center token mismatches in the actions of the lower face, upper face, and head position (coded in green). Finally, on the same dependent tier as the other two tokens, the right-most token mismatches the model in the actions of the lower face and head position. By reviewing portions of ELAN transcripts in this fashion, the team has noted that these dependent tier annotations highlight patterns in the natural data, providing detailed information about the "No Match" tokens.

5. Preliminary Findings

To perform a preliminary test of the No Match dependent tier system, the Gallaudet team chose to code a sample of the natural ASL video data, applying the new tiers to the preexisting ELAN transcripts. The first 60 seconds of a narrative were coded. Within this sample size, the narrative contained a total of 106 facial expression tokens. Of these, the No Match tokens equaled 85 (80% of the total facial expressions). Thus, while at first glance the temporal duration of the preliminary data sample may seem minute at the macro-level, via the coding process (at the micro-level) it becomes clear that this sample size is rich with content and sufficient enough for initial analysis and testing of this new coding system.

As mentioned above, from the sample data 85 No Match tokens were coded for a corresponding 106 facial expression tokens; which equates to 80% of the facial expressions appearing in the natural data not matching those of the VCom3D model. More interestingly, however, are the patterns that emerged in the dependent tiers using the new coding system. First, the coding of the closest approximate match revealed distinct frequencies of 15 different VCom3D "expressions" within the total 85 No Match tokens. Table 3 lists the closest approximate matches and ranks them (left to right) based on their frequency in the data sample.

Total No Match tokens	85				
none	23	asl_008_smile	3	asl_064_pah	3
asl_034_ab	7	asl_044_bop	3	asl_033_regular	2
asl_035_ahh	7	asl_050_fafa	3	asl_040_bah	2
asl_016_relativeclause	5	asl_060_mm	3	asl_047_eee	2
asl_090_disgust	4	asl_063_ooo	3	asl_055_gagaga	2

Table 3: Closest Approximate Match token counts[2]

As can be seen in table 3 above, 23 tokens were identified to have "none" of the VCom3D expressions as their closest approximate match; which equates to 27% of the No Matches. While for those to which an approximate match was identifiable, "ab," "ahh," "relative clause," and "disgust" were the most frequent.

Next, patterns also emerged in the coding of the FACS Action Units tier. Table 4 displays the frequency of the AUs and combinations of AUs that were observed to cause the mismatch between the natural data tokens and their closest approximate matches (VCom3D "expressions").

	Vcom Expressions	FACS Categories	Visibility
No Match	asl_008_smile.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_008_smile.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_008_smile.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_016_relativeclause.mp4	Lower Face Actions & Head Positions	Full Visibility
No Match	asl_016_relativeclause.mp4	Lower Face Actions & Head Positions	Full Visibility
No Match	asl_016_relativeclause.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_016_relativeclause.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_016_relativeclause.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_034_ab.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_035_ahh.mp4	Lower Face Actions	Full Visibility
No Match	asl_035_ahh.mp4	Lower Face Actions	Full Visibility
No Match	asl_035_ahh.mp4	Lower Face Actions & Head Positions	Full Visibility
No Match	asl_035_ahh.mp4	Lower Face Actions & Head Positions	Full Visibility
No Match	asl_035_ahh.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_035_ahh.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_035_ahh.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_044_bop.mp4	Lower Face Actions	Full Visibility
No Match	asl_044_bop.mp4	Lower Face Actions	Visibility 73 - Unscorable
No Match	asl_044_bop.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_050_fafa.mp4	Lower Face Actions & Head Positions	Full Visibility
No Match	asl_050_fafa.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_050_fafa.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_060_mm.mp4	Head Positions	Full Visibility
No Match	asl_060_mm.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_060_mm.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_063_ooo.mp4	Lower Face Actions	Full Visibility
No Match	asl_063_ooo.mp4	Upper Face Actions & Lower Face Actions	Full Visibility
No Match	asl_063_ooo.mp4	Upper Face Actions & Lower Face Actions	Visibility 73 - Unscorable
No Match	asl_064_pah.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_064_pah.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_064_pah.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility
No Match	asl_090_disgust.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_090_disgust.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_090_disgust.mp4	Upper Face Actions & Head Positions	Full Visibility
No Match	asl_090_disgust.mp4	Upper Face Actions & Lower Face Actions & Head Positions	Full Visibility

Table 4: Frequencies of FACS Action Units and Visibility for "expressions" with 3 or more instances of No Match[3]

In table 4, we can see that for each of the VCom3D expressions there were consistent patterns for how the natural data differed. Namely, the natural data differed by Upper Face Actions, Lower Face Action, Head Positions, or some combination of the three.

Also displayed in table 4 is a portion of the visibility results. This final finding, and perhaps the most definitive thus far, is that 77 of the 85 No Match tokens exhibited full visibility in the natural data (90%). In other words, for identification of the signer's non-manual features and facial expressions, particularly those that did not match the VCom3D model, all were unobstructed visually with the exception if 6, of which only 2 were unscorable. This means that, contrary to our previous supposition, a lack of visibility of the signer's face is not the source of the mismatches we had been noticing up until now. Supporting, more so, that it is the form of the signer's natural facial configurations and movements that are the key to the discrepancies in avatar development.

6. Conclusions

Although this new method of notation has only been applied to a small data set thus far, the team has already been able to find comprehensible patterns within the data that add clarification to the discrepancies previously experienced with the VCom3D notation system. By applying this new set of tiers in the transcription process, the team has been able to identify which model examples/labels are most frequently mismatching with the natural data, which portions of the face are "triggering" the mismatches, and that, despite previous supposition, visibility is not the root of the mismatches.

The representation of non-manual features presents a special set of challenges, and has not received much widespread attention in the field of sign language research. As has been demonstrated in this paper, treating independent non-manual features individually by using FACS in the ELAN transcript allows us to more accurately represent their behavior and better understand their function in language.

7. Acknowledgements

This research was supported by a subcontract agreement from VCom3D under the Department of Education SBIR Phase II Award entitled "Mobile Signing Math Dictionary with Mouth Morphemes" Award number H1335120066, awarded on October 1, 2012 to Vcom3D, Inc.

Thank you also to Sara Malkowski a research assistant on this project as well as students in our LIN 480 Linguistics Research Experience course for their efforts coding and other substantive contributions: Dana Baldiviez, Maryna Epstein , and Eddy Morrison.

8. References

Baker-Shenk, C. (1983). A Microanalysis of the Nonmanual Component of Questions in American Sign Language. PhD diss., University of Cali- fornia–Berkeley.

Castor, A., Pollux, L.E. (1992). The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1), pp. 37--53.

Corina, D. P., Bellugi, U., & Reilly, J. (1999). Neuropsychological studies of linguistic and affective facial expressions in deaf signers. *Language and Speech*, 42, pp. 307-31.

Dachkovsky , S and Sandler , W. (2009). Visual Intonation in the Prosody of a Sign Language. *Language and Speech* 287 – 314.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial Action Coding System: The Manual on CD ROM*. A Human Face: Salt Lake City.

ELAN http://tla.mpi.nl/tools/tla-tools/elan/

Grossman, R. and Shepard-Kegl (2006). To Capture a Face: A Novel Technique for the Analysis and Quantification of Facial Expressions in American Sign Language. *Sign Language Studies*, (6), 3, pp. 273-305.

[2] "Expressions" with only one instance are not included here.

[3] Tokens coded as "none" not included here.

Computer-based Tracking, Analysis, and Visualization of Linguistically Significant Nonmanual Events in American Sign Language (ASL)

Carol Neidle*, Jingjing Liu, Bo Liu**, Xi Peng**, Christian Vogler***, Dimitris Metaxas****

*Boston University / **Rutgers University / ***Gallaudet University

Boston University, Linguistics Program, 621 Commonwealth Avenue, Boston, MA 02215
Rutgers University, Computer and Information Sciences, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019
Gallaudet University, Technology Access Program, 800 Florida Ave NE, Washington, DC 20002

E-mail: carol@bu.edu, jl1322@cs.rutgers.edu, lb507@cs.rutgers.edu, px13@cs.rutgers.edu, christian.vogler@gallaudet.edu, dnm@ cs.rutgers.edu

Abstract

Our linguistically annotated American Sign Language (ASL) corpora have formed a basis for research to automate detection by computer of essential linguistic information conveyed through facial expressions and head movements. We have tracked head position and facial deformations, and used computational learning to discern specific grammatical markings. Our ability to detect, identify, and temporally localize the occurrence of such markings in ASL videos has recently been improved by incorporation of (1) new techniques for deformable model-based 3D tracking of head position and facial expressions, which provide significantly better tracking accuracy and recover quickly from temporary loss of track due to occlusion; and (2) a computational learning approach incorporating 2-level Conditional Random Fields (CRFs), suited to the multi-scale spatio-temporal characteristics of the data, which analyses not only low-level appearance characteristics, but also the patterns that enable identification of significant gestural components, such as periodic head movements and raised or lowered eyebrows. Here we summarize our linguistically motivated computational approach and the results for detection and recognition of nonmanual grammatical markings; demonstrate our data visualizations, and discuss the relevance for linguistic research; and describe work underway to enable such visualizations to be produced over large corpora and shared publicly on the Web.

Keywords: American Sign Language (ASL), nonmanual grammatical marking, computer-based sign language recognition

1. Overview

The linguistic annotation that has been carried out over the last 20 years or so by the American Sign Language Linguistic Research Project (ASLLRP) on video data collected from native users of American Sign Language (ASL) has included close attention to facial expressions and head gestures that can convey essential linguistic information. We have annotated, for example, events involving changes in eyebrow configuration, eye aperture, and head position—distinguishing the "onset" and "offset" phases, where relevant, of types of specific events (such as raised or lowered eyebrows, or head nods/shakes). Furthermore, we have labeled the linguistic information signaled by various combinations of these behaviors (topics, negation, multiple types of questions, *if/when* clauses, relative clauses, and so on) (Neidle 2002; Neidle 2007).

Our annotated corpora have formed a basis not only for linguistic research, but also for research to automate sign language detection by computer (e.g., Dreuw et al. 2008; Neidle et al. 2000). The ability to recognize linguistic information conveyed nonmanually is, of course, essential for computer-based sign language recognition and other types of applications (including, but not limited to, automatic translation) that rely upon such recognition. The general approach described here to recognition of nonmanual grammatical markers in ASL would be applicable, as well, to other signed languages.

In our earlier work, we tracked the position of the head and deformations of the face, and we used computer learning, based on the annotations of human transcribers from high-quality video images of native ASL signers, to develop the ability to discern and differentiate markings of topics, conditional clauses, negation, wh-questions, and yes-no questions, and we achieved fairly good success (Liu et al. 2013; Metaxas et al. 2012; Michael et al. 2011).

Our ability to detect, identify, and temporally localize the occurrence of nonmanual grammatical markings in ASL videos has recently been improved by incorporation of two principal innovations: (1) Newly developed techniques for deformable model-based 3D tracking, from a single video track, of head position and facial expressions (Liu et al. in press); and (2) A computational learning approach incorporating 2-level Conditional Random Fields (CRFs (Lafferty, McCallum, and Pereira 2001)) that is suited to the multi-scale spatio-temporal characteristics of the data (Liu et al. 2014). The computational analyses also enable us to produce visualizations showing the positions, over time, of the major articulators.

In Section 2, we summarize our current, linguistically motivated, computational approach, and the overall success rates now achieved for detection and discrimination of nonmanual grammatical markers. Section 3 addresses the computer-generated visualizations that we are now able to produce and their potential

value for linguistic research. In Section 4, we briefly describe work now underway to enable such visualizations to be produced over large corpora and shared publicly on the Web—as an extension of the interface described in (Neidle and Vogler 2012).

2. Computational Approach

Our current approach is summarized here. For further details about the methods and results, see Liu et al. (2014).

2.1. New tracking methods

Precise analysis of facial expressions, requiring the capture of spatio-temporal characteristics of facial events, has long been a challenging problem in computer vision. Most previous methods have been developed in controlled laboratory environments, with near-frontal faces and hardly any occlusions. For obvious reasons, these methods cannot be applied directly to ASL videos more generally, since large head movements and partial occlusions frequently occur while a subject is signing. Large and varied head movements would result in serious feature distortions of facial events. To address this problem, we take a 3D approach whereby facial expressions can be represented in a pose-invariant way.

We use a 3D deformable model-based face tracker that twines facial point localization and head pose estimation in a unique 3D shape model. Our two-stage cascaded 3D deformable shape face model localizes facial landmarks, allowing large head pose variations (Yu et al. 2013). For deformation, the first step uses mean-shift local search with a constrained local model (CLM) to achieve the global optimum. The second step uses component-wise deformable models to refine the subtle shape variation. From a single video track, we obtain 2D image coordinates of 66 facial landmarks, the corresponding 3D face shape, as well as 3 head rotations (i.e., pitch, yaw, and roll). Then feature extraction, representations, and comparisons are carried out in 3D space.

Our face tracker is capable of tracking facial expressions in the presence of large head rotations (over 30 degrees) and occlusions of the face by the hands that may occur during signing. The use of the 3D face model eliminates the alignment procedure required in 2D approaches (e.g., Active Shape Models (ASM) (Arı, Uyar, and Akarun 2008) and Active Appearance Models (AAM) (Forster et al. 2012)), which often leads to errors in head pose and expression features, restricting use of such 2D approaches to videos with small head pose variations.

See Yu et al. (2013) for comparisons of the tracking accuracy of our current 3D face tracker with that of some state-of-the-art 2D techniques, including those we have used in our previous work on recognition of nomanual markers in ASL. In all cases, the 3D method reduces the error rate by at least 50%. When tested on three public datasets (LFW (Huang et al. 2007), LFPW (Belhumeur et al. 2011), and AFW (Zhu and Ramanan 2012)), the multiple-ASM tracker (the best of the 2D trackers) and our current 3D tracker had mean average pixel errors for the facial landmark image locations as shown in Table 1.

Dataset tested	Multiple-ASM (2D) tracker	Our 3D tracker
LFW	**8.53**	**3.64**
LFPW	**17.33**	**7.37**
AFW	**20.33**	**9.13**

Table 1. Multiple-ASM (2D) tracker vs. our 3D tracker: comparison of mean pixel error rate when tested on three public corpora

We cannot provide definitive validation of the tracking for this ASL dataset, since ground truth of the locations of the facial landmarks is not available. However, the tracking appears to be working well (based on human observations) except in 12 extreme cases out of 161, where it fails: 10 video clips had severe occlusions (in which 60% of the face is occluded for over 15 frames), and 2 had large head rotations (over 60 degrees). In these cases, however, because we are using a model-based tracker, we know that the tracking has failed (because of abrupt shape changes to the model). We, therefore, are able to reinitialize the tracker, as compared with 2D methods, where this is not possible. Thus, our face tracker provides a timely tracking failure alarm and recovers quickly from temporary loss of track, thereby resulting in significantly better tracking accuracy.

2.2. Computational learning approach

Whereas previous approaches to detection of linguistic information expressed nonmanually have generally focused on low-level appearance-based features found in individual video frames (e.g., Grossman and Kegl 2006; Michael, Neidle, and Metaxas 2010; Nguyen and Ranganath 2008; Piater, Hoyoux, and Du 2010; Rodomagoulakis et al. 2011), temporal patterning over domains of variable length is also extremely important. For example, periodic head movements (nods and shakes) are an important component in the expression of many types of linguistic information. However, evaluation of a head nod or head shake requires consideration of a pattern that occurs over a time period that can vary considerably in length. Thus, we need an approach that is well suited to the multi-scale spatio-temporal characteristics of the data, one that combines low-level appearance-based features and high-level features that involve recognition of particular types of gestures—such as events involving raised or lowered eyebrows, head nods, or head shakes—and linguistically motivated evaluation of their specific characteristics and temporal phases.

We use a computational learning approach that incorporates 2-level Conditional Random Fields (CRFs

(Lafferty, et al. 2001)). At the first level of the CRF, we attend to the low-level features, based on facial geometry and appearance as well as head pose, obtained through accurate 3D deformable model-based tracking. At the second level, we learn to recognize some of the major component events that are typically found as part of the nonmanual expressions that convey specific types of grammatical information, such as raised/lowered eyebrows and head nods/shakes. Furthermore, we partition these events into their temporal phases, so that we can, for example, separate out the anticipatory movements (as the articulators get into position) from the linguistically significant region of the event; see Figure 1. We also identify the relevant characteristics of the various types of events. For example, for periodic head movements, variations in frequency and amplitude can correlate with different types of grammatical markings. Negation typically involves a side-to-side head shake; however, this head shake differs in appearance from the slight rapid head shake that is sometimes found over at least part of a wh-question; see Figure 2. We then use this multi-scale, spatio-temporal combination of low- and high-level features, in combination with the linguistically annotated corpus, to learn to detect specific linguistically important markers and to determine the temporal extent of those markings (Liu, et al. 2014; Liu, et al. 2013). Our current overall framework is shown schematically in Figure 3.

2.3. Recognition of NMMs

The new tracking and computational learning methods described above provide substantial improvements over previous methods in identification, discrimination, and temporal localization of nonmanual grammatical markers in ASL. Compared with a baseline method using only low-level features, the use of the 2-level CRF improved recognition accuracy by 20%.

Currently we validate our system on recognition of 5 major types of NMMs in 85 utterance-length videos collected at Boston University by C. Neidle and her research group. The recognition results were evaluated on a test set that contained 55 instances of topic/focus marking, 16 conditional/*when* clauses, 35 negations, 7 wh-questions, and 5 yes/no questions. As shown in the confusion matrix in Table 2, about 90% of those NMMs were correctly detected and identified; 4% were not picked up; and 6% were detected but misidentified (and all those examples involved confusion between conditional/*when* clauses and either topics (5 cases) or a yes/no question (1 case); these markings are very similar in appearance, all including raised eyebrows). In addition, there were 3 instances of false positives, where NMMs were detected that had not been identified as such in the annotations.

For details about improvements in temporal accuracy resulting from the use of the new methods, and for comparisons with the success rates for NMM recognition obtained from previous methods, see Liu et al. (2014).

		Results of detection					
		Wh	**Neg**	**Top**	**Y/N**	**C/W**	**NM**
From annotations	**Wh**	6	0	0	0	0	1
	Neg	0	34	0	0	0	1
	Top	0	0	46	0	6	3
	Y/N	0	0	0	5	0	0
	C/W	0	0	0	1	15	0
	NM	1	0	1	1	0	-

Table 2. Confusion matrix of NMM recognition: **Wh** (Wh-question), **Neg** (Negation), **Top** (Topic/Focus), **Y/N** (Yes/no question), **C/W** (Conditional/*when* clause), **NM** (no marker)

3. Computer-generated Visualizations

3.1. Visualizations that can now be produced

Figure 4 shows graphs for two example sentences illustrating degrees of eyebrow height and eye aperture, as well as 3D head position. The purple lines in the bottom graphs represent the temporal extent of manual signs, for which English-based glosses are also displayed. The 5 types of NMMs that we are currently detecting are also displayed in the visualizations that are produced from the computational analysis. Although still images are illustrated in this figure, these are actually videos that can be advanced frame by frame, with the video alignment indicator marking the current frame in the graphs.

3.2. Potential value for linguistic research

The nonmanual channel plays a vital role in the expression of various kinds of linguistic and paralinguistic information. Although this has received a fair amount of attention in the linguistic literature since about the 1970's (Baker 1976; Baker 1979; Baker and Cokely 1980; Baker and Padden 1978; Liddell 1978; Liddell 1980; Neidle, et al. 2000; Sandler 2010; Wilbur 2000, among many others), precise analysis over large data samples has been limited by the unavailability of appropriate tools.

The need to quantify observations has been felt. This has led to various approaches involving painstaking techniques for measurement and annotation by humans. For example, Grossman and Kegl (2006) used SignStream® to record impressionistically-assigned numerical values for degrees of eyebrow height; Weast (2008) used a "Screen Calipers tool" to measure pixels, by hand, in order to determine eyebrow height; the 3500 measurements for this study of 270 sentences took about 170 hours.

The possibility of producing computer-generated measurements of nonmanual components of sign language in temporal relation to the production of manual signs, for substantial data sets, opens up exciting possibilities for types of linguistic research on signed languages that have never before been possible, as well

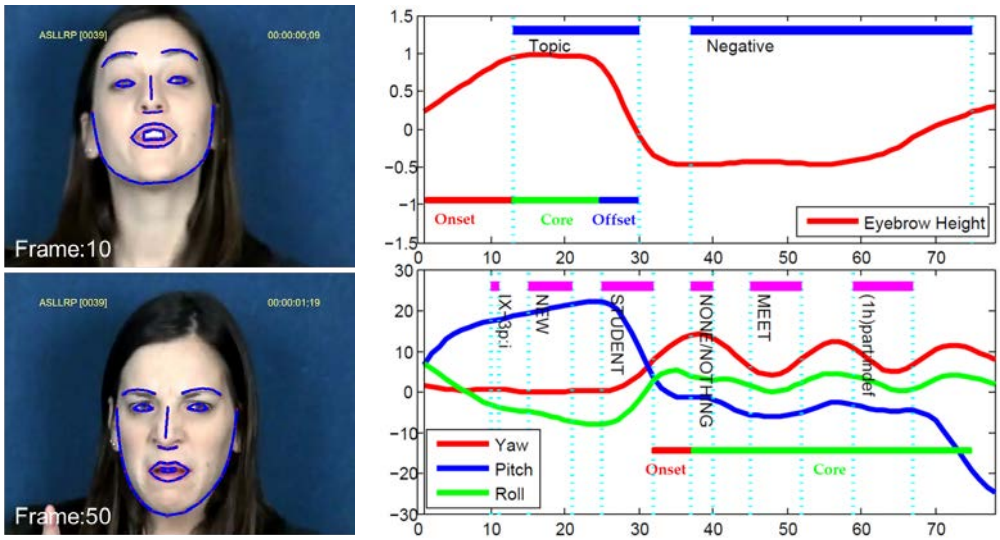


Figure 1: Detection of high-level (linguistically motivated) events—such as periodic head movements (here: head shake) and eyebrow gestures (here: raised eyebrows)—and partitioning of events into temporal phases to enable identification of the portion(s) that are of linguistic significance.

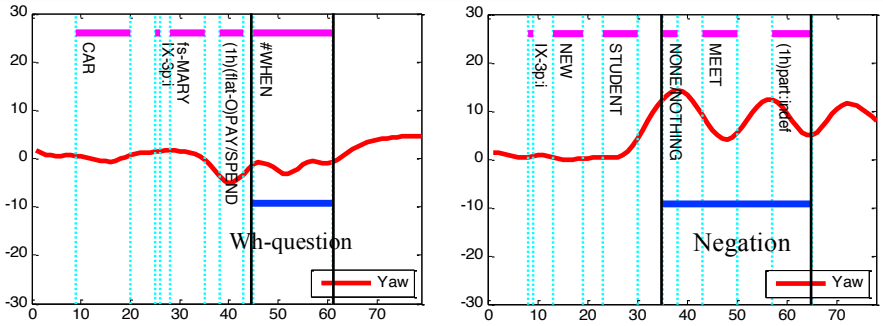


Figure 2: Analysis of the temporal patterns and properties of such detected events. For example, the head shake that occurs with negation is quite different (with respect to amplitude, velocity, peak value) from the slight rapid head shake that is sometimes found within wh-questions.

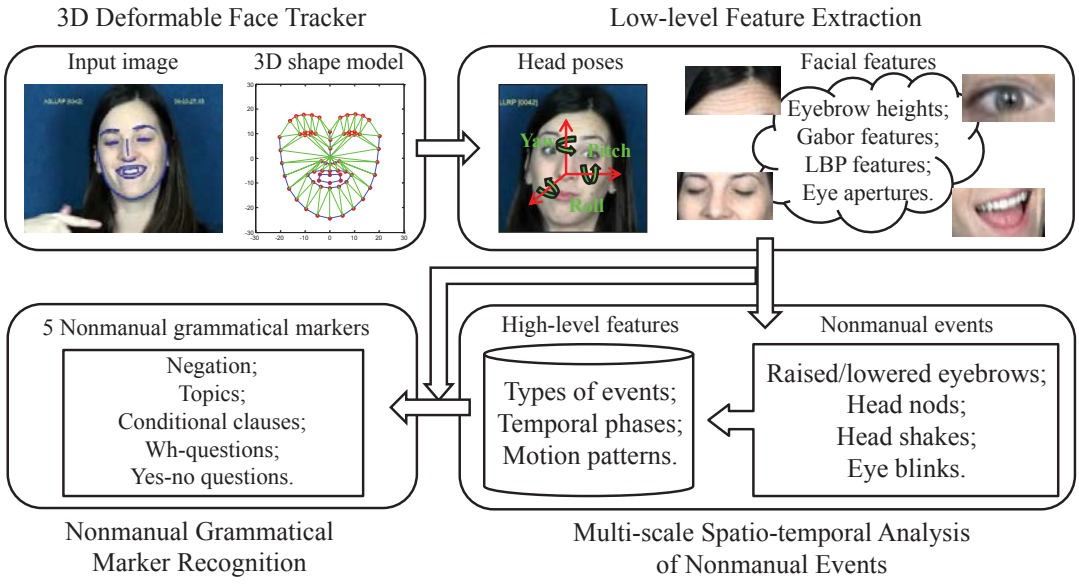


Figure 3. Overview of our current approach

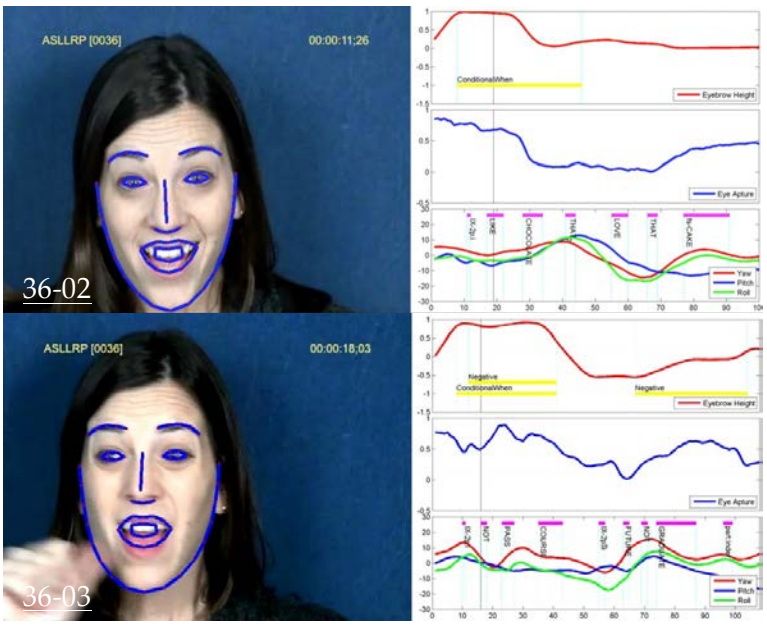


Figure 4. Visualizations of tracking data: eyebrow height (top -- in red); eye aperture (middle -- in blue); head position in three dimensions (bottom -- yaw in red, pitch in blue, roll in green

as for cross-modality comparisons of various kinds (e.g., comparing the changes in eyebrow height in signed languages with intonation contours in spoken languages). These methods would be similarly applicable to the analysis of facial expressions and head gestures in spoken language, and thus to various types of comparisons between modalities with respect to the use of the nonmanual channel. These results can also be applied to improving the linguistic realism of signing avatars, for which the unnaturalness of nonmanual expressions has been a serious issue (Kacorri, Lu, and Huenerfauth 2013).

4. Sharing Computer-generated Analyses and Visualizations

The ASLLRP corpora are shared publicly through a web-based Database Access Interface (DAI) described by Neidle and Vogler (2012). This interface allows easy searching and download of the corpora by gloss, sign type, classifier, and part of speech. Utterance and sign videos in the corpus can be viewed online in real time. The DAI is currently being extended to allow searching for utterances by nonmanual grammatical markers and nonmanual features, and to display the graphs of the computer analysis in the results list. The user can then drill down into each individual result and play back a full video of the computer analysis with the associated graphs.

Figures 5-7 illustrate a representative use case for the extended functionality: A researcher is interested in the kinematics of raised eyebrow movements in ASL, which are an important component of quite a few different NMMs. Starting with the retrieval of examples of topic markers, she selects the "topic/focus" option in the search form (Figure 5). Because eyebrows are the feature of interest, she elects to display thumbnails of the eyebrow graphs in the search results list (Figure 6; other display options are eye aperture and 3D head pose). Together with the rough glosses in this list, the graphs allow the researcher to see at a glance where in the utterances the topic markers occur, and if they exhibit the typical eyebrow movement pattern. The thumbnail with the dual occurrence of topic markers catches her attention, and she would like to investigate this utterance in more detail. She clicks on the graph to bring up a full-resolution video showing the graphs and tracking of the facial markers in detail, frame by frame (Figure 7). She can subsequently repeat the process for the other grammatical constructions of interest and see at a glance whether the eyebrow movement patterns are similar to the ones seen for topics, or whether they differ.

Sharing the data via the web-based DAI, rather than merely making the annotations and video files available for download, offers several compelling advantages. First, it makes the data accessible to a much wider audience, including those who have no expertise in using linguistic annotation software, and it works out of the box in a web browser, which everyone has installed, as opposed to requiring the installation of special-purpose software. Second, the DAI has been designed for efficient search and retrieval over large corpora, and correlating linguistic phenomena across different annotation files and videos is much quicker and easier than it is with standalone software. Third, because of the nature of the web, referencing a specific linguistic phenomenon (e.g., a topic marker seen in a specific utterance) is as simple as sharing a link with a collaborator or student, which allows them to bring up the utterance in question with a single click; bringing up the same utterance in annotation software, in contrast, takes many more steps.

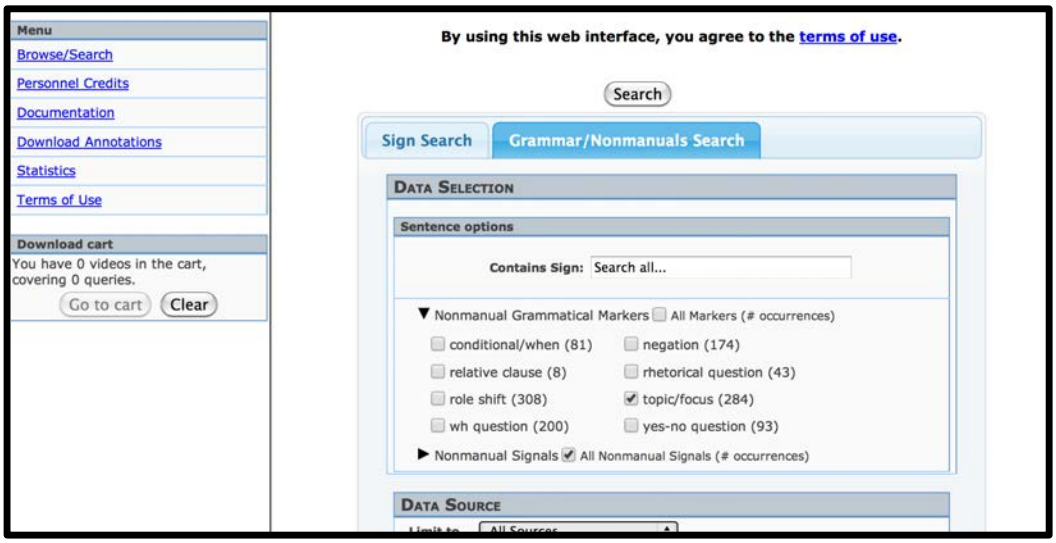


Figure 5. Search interface for nonmanual events or grammatical markers

Figure 6. Illustration of Search Results

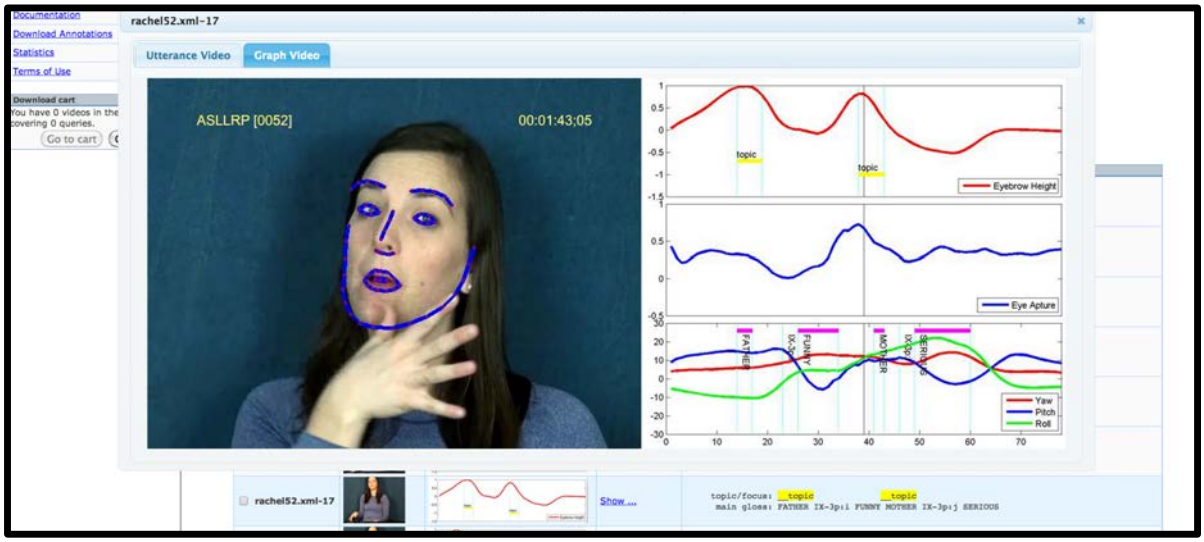


Figure 7. Video playback screen: the alignment indicator in the graph shows the position of the current video frame

5. Conclusions

We have summarized here new methods for computer-based detection of nonmanual grammatical markers in ASL, reported in greater detail in Liu et al. (2014). Such methods could readily be applied, as well, to the analysis of other signed languages, as well as to the production of nonmanual expressions used in conjunction with spoken languages.

These methods rely on computational learning, using a 2-level CRF that incoporporates both low-level features and linguistically motivated higher-level features associated with types of head motion and eyebrow events that occur over varying spatio-temporal scales. The extraction of both the low- and high-level features benefits from a new 3D deformable face tracker, which achieves greater accuracy in tracking facial landmarks and head position than has been possible with even the best 2D approaches.

Visualizations of the results of the computational analyses, which can be run on large corpora, can also be generated. We plan to make these publicly available in conjunction with our web-accessible corpora. The availability of such materials offers great potential for use in linguistic research on the nonmanual components of ASL.

6. Acknowledgments

The research reported here was partially funded by grants from the National Science Foundation (CNS-1059281, IIS-1064965, IIS-0964597, IIS-1065013, CNS-0964385, EIA-9528985, and CNS-1059218). We thank Iryna Zhuravlova, SignStream® programmer, and gratefully acknowledge the participation of the many signers and students who have worked on related projects over many years. Thanks especially to Rachel Benedict, Braden Painter, Jonathan McMillan, Jessica Scott, Indya Oliver, Corbin Kuntze, Tory Sampson, Emma Preston, Donna Riggle, Amelia Wisniewski-Barker, and Joan Nash for their invaluable contributions.

7. References

Arı, İ., Uyar, A. and Akarun, L. (2008). Facial Feature Tracking and Expression Recognition for Sign Language. Paper presented at the 23rd International Symposium on Computer and Information Sciences, 2008.

Baker, C. (1976) What's not on the other hand in American Sign Language. In Mufwene, C. and Steever, S., (eds.) *Papers from the 12th Regional Meeting of the Chicago Linguistic Society*, Chicago: University of Chicago Press. pp. 24-32.

Baker, C. (1979) *Nonmanual components of the sign language signal*, Paper presented at the NATO Advanced Study Institute, Copenhagen.

Baker, C. and Cokely, D. (1980) *American Sign Language: A Teacher's Resource Text on Grammar and Culture,* Silver Spring, MD: T.J. Publishers.

Baker, C. and Padden, C. A. (1978) Focusing on the nonmanual components of American Sign Language. In Siple, P., (ed.) *Understanding language through sign language research*, New York: Academic Press. pp. 27-57.

Belhumeur, P., Jacobs, D., Kriegman, D. and Kumar, N. (2011) Localization parts on faces using a consensus of exemplars. *CVPR*.

Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S. and Ney, H. (2008). Benchmark Databases for Video-Based Automatic Sign Language Recognition. *Proceedings of the 6th international Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco.*

Forster, J., Schmidt, C., Hoyoux, T. and Koller, O. (2012). RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey.*

Grossman, R. B. and Kegl, J. (2006) To Capture a Face: A Novel Technique for the Analysis and Quantification of Facial Expressions in American Sign Language. *Sign Language Studies, 6*(3), pp. 273-305.

Huang, G., Ramesh, M., Berg, T. and Miller, E. (2007) *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Technical Report 07-49. University of Massachusetts, Amherst, October, 2007.

Kacorri, H., Lu, P. and Huenerfauth, M. (2013) Evaluating Facial Expressions in American Sign Language Animations for Accessible Online Information. *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion, Lecture Notes in Computer Science Volume 8009*, pp. 510-519.

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Paper presented at the Intl. Conference on Machine Learning.

Liddell, S. K. (1978) Nonmanual signals and relative clauses in American Sign Language. In Siple, P., (ed.) *Understanding language through sign language research*, New York: Academic Press. pp. 59-100.

Liddell, S. K. (1980) *American Sign Language Syntax,* The Hague: Mouton.

Liu, B., Liu, J., Yu, X., Metaxas, D. and Neidle, C. (2014). 3D Face Tracking and Multi-scale, Spatio-temporal Analysis of Linguistically Significant Facial Expressions and Head Positions in ASL. Paper presented at the Language Resources and Evaluation Conference, LREC 2014, Rejkjavik, Iceland.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N. and Neidle, C. (2013) Non-manual Grammatical Marker Recognition based on Multi-scale Spatial Temporal Analysis of Head Pose and Face. In *Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai, China, April 25, 2013.*

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N. and Neidle, C. (in press) Non-manual Grammatical Marker Recognition based on Multi-scale Spatio-temporal Analysis of Head Pose and Facial Expressions. *Image and Vision Computing Journal, The Best of Face and Gesture 2013 (Special issue)*.

Metaxas, D., Liu, B., Yang, F., Yang, P., Michael, N. and Neidle, C. (2012) Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, the Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey.

Michael, N., Neidle, C. and Metaxas, D. (2010). Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation. *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, the Language Resources and Evaluation Conference, LREC 2010, Malta.*

Michael, N., Yang, P., Liu, Q., Metaxas, D. and Neidle, C. (2011) A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language. *BMVC*.

Neidle, C. (2002) *SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project.* Boston, MA: American Sign Language Linguistic Research Project Report No. 11, Boston University.

Neidle, C. (2007) *SignStream™ Annotation: Addendum to Conventions used for the American Sign Language Linguistic Research Project*. Boston, MA: American Sign Language Linguistic Research Project Report No. 13, Boston University.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B. and Lee, R. G. (2000) *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure,* Cambridge, MA: MIT Press.

Neidle, C. and Vogler, C. (2012) A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface (DAI) In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, the Language Resources and Evaluation Conference, LREC 2012*, Istanbul, Turkey.

Nguyen, T. D. and Ranganath, S. (2008). Tracking facial features under occlusions and recognizing facial expressions in sign language. *Proceedings of the IEEE Conference on Automatic Face & Gesture Recognition.*

Piater, J., Hoyoux, T. and Du, W. (2010). Video Analysis for Continuous Sign Language Recognition. *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, the Language Resources and Evaluation Conference, LREC 2010, Malta.*

Rodomagoulakis, I., Theodorakis, S., Pitsikalis, V. and Maragos, P. (2011). Experiments on Global and Local Active Appearance Models for Analysis of Sign Language Facial Expressions. *Proceedings of the 9th International Gesture Workshop (GW 2011): Gesture in Embodied Communication and Human-Computer Interaction, May 2011, Athens, Greece.*

Sandler, W. (2010) Prosody and Syntax in Sign Languages. *Transactions of the Philological Society, 108*(3), pp. 298-328.

Weast, T. P. (2008) *Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows*. Unpublished Doctoral dissertation, University of Texas, Arlington.

Wilbur, R. B. (2000) Phonological and prosodic layering of nonmanuals in American Sign Language. In Lane, H. and Emmorey, K., (eds.) *The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima*, Hillsdale, NJ: Lawrence Erlbaum, pp. 213-241.

Yu, X., Huang, J., Zhang, S., Wang, Y. and Metaxas, D. N. (2013) Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model. *ICCV*.

Zhu, X. and Ramanan, D. (2012) Face detection, pose estimation and landmark localization in the wild. *CVPR*.

Nonmanuals and markers of (dis)fluency

Ingrid Notarrigo and Laurence Meurant
F.R.S. – FNRS and University of Namur
Namur, Belgium
E-mail: ingrid.notarrigo@unamur.be, laurence.meurant@unamur.be

Abstract

This paper focuses on the analysis and annotation of non-manual features in the framework of a study of (dis)fluency markers in French Belgian Sign Language (LSFB). In line with Götz (2013), we consider (dis)fluency as the result of the combination of many independent markers ("fluencemes"). These fluencemes may contribute either positively or negatively to the efficiency of a discourse depending on their context of appearance, their specific combination, their position and frequency. We show that the non-manual features in LSFB make distinctions within pauses and palm-up signs in a consistent way and contribute to the value of the manual marker. The selection of a limited number of relevant combinations of nonmanuals, in the context of pauses and palm-up signs, proves to simplify the annotation process and to limit the number of features to examine for each nonmanual. The gaze and the head appear to be necessary and sufficient to describe pauses and palm-up signs accurately. Though these findings are limited to this pilot study, they will pave the way to the next steps of the broader research project on (dis)fluency markers in LSFB this work is part of.

Keywords: annotation, fluency, disfluency, nonmanuals, pauses, palm-up signs, stops, gaze, eyes, eyebrows, head, mouth

1. Introduction

This study focuses on fluency and disfluency in "normal", i.e. non-pathological, signing and sets apart the impressive amount of research on disfluency conducted in the areas of stuttering and aphasia (Marshall 2000; Atkinson et al. 2002). From a holistic perspective, fluency is associated with the impression of an overall discourse quality, a "smooth, rapid, effortless use of language" (Crystal 1987: 421), or "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language" (Lennon 2000: 26). However, "fluency does not always imply an uninterrupted flow of speech which is grammatically perfectly irreproachable" (Lehtonen 1978); in other words a successful communication or proficient/efficient speech does include disfluencies.

Götz (2013) noticed that disfluency can be considered not only as a signal of the speaker's difficulties to plan and encode his/her discourse, but also as a positive signal when speakers use disfluencies for rhetorical purposes for example. She pointed out that, depending on its context, its combination with other features, its position and frequency, the same feature can contribute either to the fluency or to the disfluency of a production. This study is in line with Götz's componential approach that sees (dis)fluency as the result of combinations of many independent markers ("fluencemes"), and is part of a PhD thesis that aims to identify fluencemes in French Belgian Sign Language (LSFB) and to observe their combinations within different contexts of speech. We expect to be able to identify fluency and disfluency profiles in terms of combinations of fluencemes, probably related to the type of speech context.

Two potential fluencemes of LSFB are focused on in this study, namely pauses and palm-up signs. Their non-referential contribution to the discourse makes them good (dis)fluency marker candidates. At a first glance, nonmanual features occurring with both pauses and palm-ups seem to convey important information related to (dis)fluency. A gaze can for instance interrupt communication temporarily, an 'erm' mouthing or head and eyebrows behaviours can express reflexion or hesitation. However, annotating each non-manual articulator (i.e. gaze, eyes, eyebrows, mouth, head) in detail is extremely time consuming; it may be worthwhile to test whether such precise annotation is relevant, i.e. whether non-manual information refine the information given by the manual marker (pause or palm-up). In this study, we address three main research questions: (1) What type of information do nonmanuals give about pauses and palm-up signs? (2) Is the annotation of each non-manual component needed for each pause and each palm-up sign? (3) How is it possible to code the potentially relevant nonmanuals?

2. Methodology

To answer these questions, we conducted a pilot study based on a 10-minute long corpus. The corpus consists in four excerpts of unprepared monologues produced by 2 native and 2 near-native signers of LSFB (see details in Table 1). The excerpts were selected from larger interviews or dialogues, but are considered as monologues because the interlocutor does not interrupt the signer's turns within the selected clips.

	Sex	Age	SL profile	Clip duration
Signer 1	M	33	Native	3 min
Signer 2	F	22	Native	2 min
Signer 3	M	25	near-native	2 min 30
Signer 4	F	28	near-native	2 min 30

Table 1: Signers and clips

Within these data, we first coded each pause and each palm-up sign. Then, we looked at their immediate context, and more precisely at the behaviour of the gaze, the eyes, the eyebrows, the head and the mouth, which were annotated in five separate tiers. In doing that we

improved our annotation guidelines for the nonmanuals and finally applied a template that appeared to be suited and efficient for our subject (see section 4).
With a multi-layer search in ELAN we extracted for each occurrence of a pause or of a palm-up sign its overlapping non-manual features. We finally queried the data in Excel and generated information about the non-manual features co-occurring with each manual marker. We tried to see whether some (combinations of) nonmanuals behave regularly when a pause or a palm-up appears, and whether these regularities draw boundaries between consistent groups of pauses and palm-ups. The absence of any regularity would contribute to the assumption that the behaviour of nonmanuals is not related to the pauses or the palm-ups they occur with, and therefore has not to be coded for its relation to each pause or palm-up occurrence. This pilot study alone can certainly not lead us to adopt this assumption conclusively, but it can determine the next steps of the investigation of the interaction between nonmanuals and manual markers of (dis)fluency.

3. Coding pauses and palm-ups

3.1 Pauses

In comparison with what is known about spoken language fluency, a first glance at our data reveals a strikingly small amount of unfilled pauses in the signing flow. In fact, this difference may be due to the breathing limits that constrain the speech flow, but above all it may be due to the scarce use of multimodal data for the description of spoken productions: access to the silent information conveyed by manual or non-manual components during speech productions would probably have led to give up the concept of unfilled pause. From videotaped sign language (SL) data, it becomes clear that the stops of the hands are inevitably "filled" with non-manual information. So, instead of distinguishing between filled and unfilled pauses, we considered all the stops of the hands as (dis)fluency markers, since all may help the signer to plan or reorganize the discourse, be it in a fluent or in a disfluent way.
We drew a distinction between stops during a sign (S1) and stops between signs (S2). The first group (S1) includes stops at the beginning, in the middle or at the end of a sign; they are recognizable by the fact that handshape and location of the signs are hold. We took these kind of stops into account when they lasted at least 5 frames[1], and we coded them S1:start, S1:middle and S1:end respectively. The second group (S2) covers all the cases of non-signing at all, or in other words cases where the hands do not show a meaningful handshape or movement. We divided S2 into three sub-groups depending on the position of the hands: crossed hands (S2:crossed), along the body (S2:body) and relaxed in the neutral space in front of the signer (S2:neutral). Table 2 provides an overview of these (sub-)groups and their respective tags.

Pauses	Stop in the hands flow
S1	Stop during a sign
S1:start	Stop at the beginning of a sign
S1:middle	Stop in the middle of a sign
S1:end	Stop at the end of a sign
S2	Stop between signs
S2:crossed	Stop with hands crossed
S2:body	Stop with hands along the body
S2:neutral	Stop with relaxed hands in the neutral space

Table 2: Pauses (sub-)groups and related tags

3.2 Palm-up signs

The "palm-up" sign is described in numerous sign languages (among others van der Kooij, Crasborn and Ros 2006 and van Loon 2012 for NGT). The form of the sign (an upward palm orientation sign articulated in the neutral space and resulting from a wrist location) and its various functions (expression of modality, backchannel signal, elicitation of evolvement, start or end of a turn, conjunction, interrogative particle or pause filler) are similar across sign languages. The spectrum of functions related to palm-ups prompted us to count them as a potential (dis)fluency marker.
Four groups of palm-ups have been distinguished, according to the hand(s) involved in the sign and to the handshape(s) taken by the hand(s). The canonical palm-up sign is articulated by the two hands in 5-handshape (PU). But the palm-up can also be articulated with only one hand in 5-handshape (PU-R for the right hand and PU-L for the left hand). In some cases, we saw a two-handed palm-up with one hand in 5-hanshape and one hand in I-handshape (PU-L (I)). See Table 3 for an overview of these groups.

Palm-up signs	Upward palm orientation sign in the neutral space resulting from a wrist rotation
PU	Palm-up sign with both hands in 5-handshape
PU-R	One 5-hanshape handed palm-up (right hand)
PU-L	One 5-handshape handed palm-up (left hand)
PU-L(I)	Palm-up sign with one hand in 5-handshape and one hand in L-handshape

Table 3: Palm-up groups and related tags

4. Coding nonmanuals

Once each pause (S) and each palm-up sign (PU) had been tagged, we coded the behavior of the non-manual components occurring in the close context of each S and PU: the gaze, the eyes, the eyebrows, the head and the mouth. We deliberately began with a quite extensive annotation grid based on existing protocols (Neidle 2002; Nonhebel, Crasborn and van der Kooij 2004; Johnston 2011) and refined it during the annotation process. One of the main changes we applied corresponds to the time

[1] This length of 5 frames (one frame is 1/50 sec.) does not come from an upstream decision, but rather from the downstream observation that, below a length of 5 frames, we could not detect the stop.

intervals we considered for each nonmanual. For example, we started to code the gaze components from two signs before to two signs after the manual marker (S or PU). But this interval appeared to provide noise, namely information that was obviously not related to the marker we were focused on but to the previous or next signs. Annotating the gaze behavior only one sign (300-500 milliseconds) before and one sign after the S or PU marker proved to be more accurate.

The annotation guidelines presented below are the final version we applied to all our data. In comparison with the first extensive guidelines, it represents a reduction of 66% of the time needed for annotation (from 150 min to 50 min for a 30-second video clip). The reduction might be due to transcribers getting used to the task, but the most important impact is due to the smaller number of non-manual elements to look at and of values for each non-manual element.

4.1 Gaze

As indicated above, the gaze component was taken into account from one sign before to one sign after the manual marker (S or PU). The tag set used distinguishes three behaviours and is based on Meurant (2006)'s study on gaze in LSFB.

First possibility: the gaze is tagged as "addressed". This means that the gaze addresses a real or a fictive interlocutor, namely a discourse participant to whom the signer may say 'I' or 'you'. Second possibility: the gaze is tagged as "spatial". This means that the gaze installs or designates meaningful positions in space, other than the positions of the real or fictive interlocutors. Third possibility: the gaze is tagged as "other". This means that the gaze is not addressed nor related to meaningful positions in space. It can for example be oriented to the floor, to the side or in the air, or be shifty.

When a change of gaze occurs and is accompanied by a blink, the blink is considered as the beginning of the new gaze behavior.

In a previous version of the guidelines, the "spatial" tag was split into "spatial – out of a role" and "spatial – within a role". The former covered the gaze that installs or designates positions in the frontal space (Meurant 2006, pp. 407-408) without any relationship to the actualization of a character in a role-taking form. The latter covered the gaze that installs or designates positions in the space surrounding the signer (Meurant 2006, p. 409) in relation to the actualization of a character in a role taking form. We kept records of this previous tagging. The analyses of the data (section 5) suggest that the distinction between "out of a role" and "within a role" is relevant, especially within the PU and the S1:end categories. This means that the four-tag set (addressed / spatial – out of a role / spatial – within a role / other) will be re-introduced in our next guidelines.

4.2 Eyes

As for the gaze, the eye component was taken into account from one sign before to one sign after the manual marker (S or PU). The tag set includes six features: "closed", "blink", "eyelid down", "wide open", "squint" and "other". The interval of a blink begins the frame before the closing position and ends at the opening of the eyes; the mean length of a blink is 5 frames as a whole. If the eyes are maintained in the closed position more than one image, they are considered as closed (Chételat-Pelé and Braffort 2010).

4.3 Eyebrows

Only two tags are used to describe the eyebrow movements: "raised" and "frown". To avoid noisy information, they are used strictly within the interval of the manual marker: the movement can appear after the beginning of the S or the PU, but it never goes beyond the end of the S or the PU. The eyebrows movement is coded from one frame before the beginning of the raising or frowning movement to one frame after the peak. The movement after the peak is not coded because it is often hard to see.

4.4 Head

Coding the head components proved to be quite difficult. We came to the conclusion that the more consistent principle (in order to avoid to code movements that are not related to S or PU, but rather to the surrounding context) was to code only the changes that occur during the manual marker. Moreover, we excluded from these the changes that overlap with the manual marker but that are due to the next context (e.g. a negation after the S or the PU that produces a head turn before the very end of the S or the PU). We used seven tags for the description of the head: "nod", "shake", "turn", "tilt", "chin up", "chin down" and "other". Sometimes it is hard to distinguish the turn from the tilt. We tagged "turn" if the chin goes on one side and the face is no longer facing the interlocutor. We tagged "tilt" if the top of the head moves without a change in the direction of the face. The idea is to annotate the most salient feature. For example when a turn occurs, it is only coded as "turn" and not for the movement of the chin that is unavoidably linked to the turn.

4.5 Mouth

In a first step, we used the tags described in the sign language transcription conventions for the ECHO Project (Nonhebel et al. 2004). After having coded the mouth components in detail (open/closed, corner of the lips, tongue, teeth, etc.) for 92 Ss or PUs, we substantially cut down the number of features because the data would have been too heterogeneous to analyse in combination with non-manual tags. The seven remaining tags are the following: "closed", "closed with lip movement", "closed with air (breathe out)", "open", "open with lip movement", "open with air (breathe in)", and "mouthing". We have limited the coding to the strict interval of the S or PU. A mouth movement that is similar to the 'erm' in spoken language has been coded as "open with air" and not as "mouthing" because the mouthing is not always clear enough.

Table 4 shows an overview of the final and complete tag set used for this pilot study to describe nonmanuals.

Gaze (G:)	Eyes (E:)	Eyebrows (B:)	Head (H:)	Mouth (M:)
addressed	closed	Raised	nod	closed
spatial	blink	frown	shake	closed-lip mov.
other	eyelid down		turn	closed with air
	wide open		tilt	open
	squint		chin up	open-lip mov.
	other		chin down	open with air
			other	mouthing

Table 4: Tag set for nonmanuals

5. Results

After having annotated the small-scale corpus presented in section 2 according to the guidelines presented in section 3 for the manual elements and in section 4 for nonmanuals, we were able to start a multi-layer search in ELAN in order to extract for each occurrence of a pause or a palm-up sign its overlapping non-manual components.

Our first question aims to investigate the type of information nonmanuals give about pauses and palm-up signs. We tried to see whether some nonmanuals or some combinations of nonmanuals behave regularly when a pause or a palm-up appears, and whether these regularities can help distinguish consistent categories of pauses and palm-ups.

In practice, we started the analyses with a spreadsheet containing all the Ss (113) and PUs (80) occurrences (total: 193), each one being associated with its respective tags on non-manual components. Within these data, we investigated each pause (sub-)group and each palm-up group by filtering the data by non-manual tags. These filtering operations resulted in successive occurrences sets that we systematically examined in terms of consistency. Repetitively, the question was "is there any apparent coherence between the groups resulting from this filter (or combination of filters)?". The consistency was approached in terms of position (within the turn or within the semantic unit, if the turn was made up of several ideas) and in terms of functions (in a broad sense and out of any theoretical typology of functions).

The results of this investigation are presented below, showing the more consistent categories of manual markers arising from the regularities observed in their co-occurring nonmanuals.

5.1 Palm-up signs and nonmanuals

All categories of palm-up signs (PU, PU-R, PU-L, PU-L(I)) are clearly divided into two main categories by the criterion of gaze (see Table 5). A PU with a gaze tagged as "spatial" (more precisely a spatial gaze within a context of role taking[2]) fulfills the function of a modality marker (PU-Mod): It conveys a subjective comment or evaluation from the point of view of the role-played character or the signer himself/herself on what is being said (disagreement, feeling of inability, pleasure, etc.)[3]. All the other gazes ("addressed", "spatial out of a role" and "other") indiscriminately cover the uses of PU as lexical units (THAT-IS or NOW) and fillers (PU-Lex/Fill), whatever the position of the PU is: at the starting, during or at the end of the semantic unit. See Examples 1-3 with Figures 1-4 to have an illustration of each category.

The PU-Lex/Fill are often accompanied by other potential (dis)fluency markers, such as pauses (S1 or S2), false starts, connecting particles, etc. Within the two categories, no other consistent sub-category seems to be related to any other nonmanual.

PU	Defining nonmanual	Tag	Number of occurrences
Modality marker	**G:spatial (within a role)**	PU-Mod	21
Lexical units and fillers	G:all the other tags	PU-Lex/Fill	59

Table 5: Palm up categories

Ex. 1 BEFORE FG:E Grid $_{\text{PU-Mod}}$ FG:E GIVEN UP GRID GRID CALCULATION GRID $_{\text{PU-Lex/Fill S2:body-BoE}}$
Before, the sign for Excel was with the letter E. ***It is not good****. We gave up the letter E and we kept only the sign for grid.* ***Here it is****.*

Figure 1: PU-R-Mod on the left, PU-Lex/Fill on the right

Ex. 2 $_{\text{PU-R-Lex/Fill}}$ I SIGN $_{\text{PU-Lex/Fill}}$ YES $_{\text{S2:crossed-BoM}}$ ERM I $_{\text{PU-L(I)-Lex:Fill}}$ DEAF WORLD DAY TRUE DEAF WORLD DAY WHY?
Here it is*. I sign* ***now****. [/] Yes, erm, according to me,* ***well****, what is the point of the Deaf World Day?*

[2] As previously mentioned (section 4.1), the distinction between "spatial out of a role" and "spatial within a role" made in a pre-final step of the annotation guidelines should be re-introduced in the tag set.

[3] This is in line with van der Kooij et al. (2006).

Figure 2: PU-R Lex/Fill (HERE-IT-IS) on the left, PU Lex/Fill (NOW) on the right

Figure 3: S2:crossed-BoM on the left, PU-L-(I)-Lex/Fill on the right

Ex. 3 DEAF $_{\text{PU-L (I)-Mod}}$ NOT ENOUGH IN MORE WORLD DEAF
*Deaf people, **Oh!** they **really** are not involved enough in the deaf world.*

Figure 4: PU-Mod

5.2 S1:end and nonmanuals

In a similar way as for the PUs, the S1:end markers are firstly sub-categorized by the opposition between the gaze-tag "spatial within a role" and the other gaze-tags ("addressed, "spatial out of a role" and "other"). This first distinction identifies a group of S1:end functioning as a modality marker (S1:end-Mod) in the same way as the PU-Mod (see below Ex. 4).

Ex. 4 I WALK BEAUTIFUL DUCK MANY I LOOK S1:end-Mod I BECAUSE I DEAF HEARING MANY I ALONE S1:end-Mod DEAF
*I'm walking. There are many beautiful ducks. I **look** at them **for a long time**. There are many hearing people around me, but I am the **only** deaf person.*

As for the other cases, namely with a gaze which is not "spatial within a role", the presence of a head movement is relevant. When there is a head movement other than "nod", S1:end functions as a marker of stress (S1:end-Str) (see Ex. 5). When there is a head nod, it fulfills a phatic function, namely it shows that the signer makes sure he is well understood (S1:end-Pha) (see Ex. 6).

Ex. 5 BEFORE WIRE$^{\text{S1:end-Pha}}$ WIRE COMPUTER HOME WIRE $^{\text{S1:end-Pha}}$ $_{\text{PU-Lex/Fill}}$ WIRE NOTHING$^{\text{S1:end-Str (shake head)}}$
*Before, there was a wire, **ok**. At home, there was a wire line computer, **ok, well** there is **no** more wire.*

Ex. 6 YES FUTURE BETTER CHANGE FOR EXAMPLE $_{\text{S2:crossed-BoM}}$ TOO MUCH SPELLING$^{\text{S1:end-Pha}}$ FOR EXAMPLE USB$^{\text{S1:end-Pha}}$ BETTER KEY$^{\text{S1:end-Pha}}$
*Yes, it is better to change. For example **[/]**, there are too many signs with **acronyms, ok**. For example, for the sign **«USB», ok**, it is better to use the sign for **key, ok**.*

When S1:end is not accompanied by a head movement and the gaze is not the same as for the modality marker, it rather produces an effect of suspension within the discourse, a sort of blank in the communication (see below Ex. 7). Table 6 sums up these four categories.

Ex.7 DEAF WORLD DAY YES THERE PARIS PARIS$^{\text{S1:end-Sus}}$ ERM THREE FOUR YEAR PAST
*Yes, the Deaf World Day took place in Paris, **Paris**, erm, three or four years ago.*

S1:end	**Defining nonmanual(s)**	**Tag**	**Number of occurrences**
Modality marker	**G:spatial (within a role)**	S1:end-Mod	9
Others Stress	G:addressed or G:spatial out of a role **H:movement but not "nod"**	S1:end-Str	7
Phatic	G:addressed **H:nod**	S1:end-Pha	19
Suspension	G:addressed or G:other **H:/**	S1:end-Sus	12

Table 6: S1:end categories

5.3 S2:body/crossed and nonmanuals

S2:body and S2:crossed are both categorized in the same way by the nonmanual components. They all function as boundary markers (Bo). Once again, the gaze draws relevant boundaries between them. Combined with the

regularities in terms of position of the markers, the gaze distinguishes between three main S2:body/crossed categories. At the beginning or the end of a speech turn, a S2:body/crossed is perceived as a framing pause (S2:body/crossed-BoS and S2:body/crossed-BoE). In most cases (BoS and BoE) the gaze is addressed and may be highlighted by a head nod. But in some cases (only at the starting of a turn – BoS), the gaze is tagged as "other" and is layered by a turn. The S2:body/crossed markers that appear within a turn (S2:body/crossed-BoM) mark the end of a semantic unit. They may be accompanied by a turn. Table 7 provides an overview of these categories and Figures 3/5 illustrate the difference of gaze within these categories.
At the end of a turn, a S2:body/crossed with a nod fulfills a phatic function, in a similar way as S1:end-Pha.
The various S2:body/crossed often appear just after or before a PU.

S2:body S2:crossed	**Defining nonmanuals**	**Tag**	**Number of occurrences**
Boundary marker			
Framing pause - End of turn (phatic)	G:addressed H:nod	S2:body/crossed-BoE	10
	G:addressed H:/		7
Framing pause - Start of turn	G:addressed H:nod	S2:body/crossed-BoS	2
	G:addressed H:/		6
	G:other H: turn		4
Middel of turn, end of semantic unit	G:other H:turn or /	S2:body/crossed-BoM	9

Table 7: S2:crossed and body categories

Figure 5: S2:crossed-BoM on the left, S2:body-BoE on the right

5.4 S2:neutral and nonmanuals

Three categories of S2:neutral have been found. These three categories are summed up in Table 8. All three are similar to the already established categories for the other kind of manual markers. The clue nonmanuals are the head (movement or not) and the gaze (spatial or not). The presence of a head movement characterizes the modality marker (S2:neutral-Mod, also recognizable by its usual "spatial – within a role" gaze) (see an illustration in Figure 6) and a boundary marker (with "addressed" or "other" gaze). As a boundary marker, S2:neutral specifically marks the transition between a concept and its explanation (S2:neutral-BoEx) as illustrated in the Example 8.

Ex. 8 SOCIETY STRONG DIFFERENT $_{\text{S2:nEUTRAL-BoEx}}$ POOR RICH WORLD ONE WORLD TWO
*The society is very different **[/]** there are two worlds: one for the poor and another one for the rich.*

The lack of head movement (whatever the gaze and the position of the marker is) produces an effect of suspension of the discourse (S2:neutral-Sus), in the same way as in S1:end-Sus (see Ex. 9). This third category often appears in the close context of another (dis)fluency marker, as for example S1 pauses, auto-contacts, "flying indexes", etc.

Ex. 9 INFORMATION DIFFERENT ASSOCIATION THERE-IS FOR FOCUS$^{\text{S1:end-Sus}}$ $_{\text{S2:neutral-Sus}}$ CULTURE DEAF
*There are different associations giving information in order to focus **[/]** on the deaf culture.*

S2:neutral	**Defining nonmanual(s)**	**Tag**	**Number of occurrences**
Modality marker	**H:movement** G:spatial (within a role)	S2:neutral-Mod	3
Boundary marker Explanation	H:movement G:addressed or other	S2:neutral-BoEx	4
Suspension	**H:no movement** G:addressed or other	S2:neutral-Sus	12

Table 8: S2:neutral categories

Figure 6: S2:neutral-Mod on the left, S2:neutral-Sus on the right

5.5 S1:start and nonmanuals

Our data only contained 9 occurrences of S1:start, so it is required to treat the remarks below with caution. We hypothesize that when a pause comes at the beginning of a sign, it can either produce an effect of hesitation similar to a false start, or mark a stress (see Table 9). Depending on our examples, the latter function is cued by a combination of five non-manual features: G:addressed or spatial (out of a role), E:wide, B:raised, H:chin up, M:closed. Figure 7 shows an illustration of the contrast between these two categories.

S1:start	Defining nonmanual	Tag	Number of occurrences
Hesitation	**G:spatial (within a role) or other or addressed**	S1:start -Hes	7
Stress	G:addressed or spatial (out of a role) E:wide B:raised H:chin up M:closed	S1:start -Str	2

Table 9: S1:start categories

Figure 7: $_{\text{S1:start-Str}}$WHOLE on the left, $_{\text{S1:start-Hes}}$WINDOW on the right

No cases of S1:middel were found in our data.

6. Discussion

The results presented in section 5 suggest that the non-manual components of LSFB make distinctions within pauses and palm-up signs consistently and contribute to the value of the manual marker. Each marker category was shown to cover various functions, such as modality or boundary or phatic markers. The distinction between the different functions can be linked to the non-manual information and even to a reduced set of non-manual features which may have a significant impact on the annotation work. In the same vein, the improvement of the guidelines we established (mainly the delimitation of the intervals to consider and of the features to examine) for the coding of the nonmanuals co-occurring with potential (dis)fluency markers such as pauses and palm-up signs, is in itself a considerable gain (66% of time saving) for the annotation efficiency.

This study and its results are limited by the shortcomings that are inherent to every pilot study: the reduced amount of data, of signers, of speech context variety, etc. The 193 occurrences of pauses and palm-ups we examined represent only a sample of 10 minutes of the productions of four signers. Despite the small-scale data, a qualitative study could be carried out that paves the way for the next – more extensive – steps of this research on (dis)fluency markers in LSFB. By using a broader corpus and quantitative analysis techniques (Chi2 and multivariate analysis for instance), we should be able to test the relevance of the nonmanuals combinations resulting from this first investigation on the sub-categorization of pauses and palm-ups.

With regard to the issue of nonmanuals and their relation to the two manual markers we have focused on, the preliminary findings can be summed up as follows.

[1] The fact that pauses and palm-up signs frequently appear with other probable (dis)fluency markers confirms that they deserve being taken into account in the combinatory study we pursue.

[2] The annotation guidelines presented in sections 4 and 5 seem to be appropriate and efficient for our subject. A small change will be done, within the gaze-tag set. Coming back to a previous choice, a four-tag set will be used for coding the gaze: addressed/spatial – out of a role/spatial – within a role/other.

[3] Two types of nonmanuals must be coded in order to describe pauses and palm-ups accurately, namely the gaze and the head. Together they form the defining cues for the sub-categories of all groups of markers: PU, S1:end, S2:body and S2:crossed, S2:neutral, S1:start. Moreover, depending on the marker, the annotator can know which nonmanual refines the information provided by the gaze and the head and which ones are not expected to provide regular information.

[4] One specific type of gaze (namely the "spatial – within a role" gaze) gives the same function to the PU, the S1:end and the S2:neutral markers. This function has been identified as the marking of modality.

[5] A particular behaviour of the head, namely the absence of movement of the head, layered with a pause or a palm-up and with a sort of fixity in all manual and nonmanual components, produces an effect of suspension that is common to S1:end and S2:neutral.

The presence of a nod, be it with S1:end or with S2:body or S2:crossed, gives to the marker a phatic function.

[6] These regularities among groups of markers can be seen as a signal of accuracy among the categories and features we found.

[7] Within the PUs, S1:end and S2:neutral, the opposition between "addressed" and "other" gaze surprisingly do not impact the function of the sign. The same can be seen with other markers in LSFB, like THAT-MEANS (see Figure 8), ALSO or the use of list buoys. This prompts us to investigate whether the gaze

would be independent, and whether it could be considered, in itself, as a (dis)fluency marker.

Figure 8: SAY WHAT THIS-IS
(*What does it mean? It is…*)

7. Conclusions

This study shows that the non-manual components of LSFB make distinctions within pauses and palm-up signs in a consistent way and contribute to the value of the manual marker. The relevant combinations of nonmanuals, in the context of pauses and palm-up signs, help speeding up the annotation process by reducing the number of nonmanuals that must be taken into account and by limiting the number of features to examine for each nonmanual. The gaze and the head appeared to be necessary and sufficient to describe pauses and palm-up signs accurately.

These findings are limited to the extent of this pilot study. But it will pave the way for the next steps of the broader research project on (dis)fluency markers in LSFB (Degand et al. 2012) this work is part of. The next two steps will be to test the validity of these results on a broader corpus and to extend the study to other potential (dis)fluency markers. We will have to make a selection between, among others, false starts, self-repairs, repetitions, "flying indexes", gestures/motions fillers, spatial discourse organization, constructed actions, connecting signs such as rhetorical questions, AND, ALSO, SAME, and finally maybe the eye gaze.

8. Acknowledgements

We would like to thank our deaf informants, as well as Simon Delauvaux, Calogero Notarrigo and Amandine Dumont for their collaboration. Our research is funded by a F.R.S-FNRS Research Fellow Grant FRESH FC 60970 and the A.R.C. n°12/17-044.

9. References

Atkinson, J. et al. (2002). When sign language breaks down: Deaf people's access to language therapy in UK. In Deaf worlds 18 (1) 9-10, Forest Bookshop, pp. 9-13.

Chételat-Pelé, E.; Braffort, A. (2010). Investigation et analyse des Gestes Non Manuels impliqués en LSF: Le cas des clignements. *Session posters at the Atelier TALS*. Montréal, Quebec.

Crystal, D. (1987). The Cambridge Encyclopedia of Language. Cambridge: Cambridge University Press.

Degand et al. (2012). Fluency and disfluency markers. A multimodal contrastive perspective, ARC: 12/17-044, University of Louvain-La-Neuve: Belgium.

Götz, S. (2013). Fluency in Native and Nonnative English Speech. *Studies in Corpus Linguistics*, 53. John Benjamins Publishing Company.

Herrmann, A. (2008). Sign Language corpora and the problems with ELAN and the ECHO annotation conventions. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood (Eds.), *Construction and Exploitation of Sign Language Corpora. [Proceedings of the 3rd Workshop and 6th LREC 2008]* Paris: ELRA, pp. 68-73.

Johnston, T. (2011). Auslan Corpus annotation guidelines.

Lehtonen, J. (1978). On the problems of measuring fluency. In M. Leiwo and A. Rasanen (Eds.), AFinLA Yearbook 1978. Jyväskylä: AFinLA, pp.53-68.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (ed.), *Perspectives on Fluency*. Ann Arbor, MI: The University of Michigan Press, pp. 25-42.

Marshall, R.C. (2000). Speech fluency and Aphasia. In H. Riggenbach (ed.) *Perspectives on fluency*, Ann Arbor: The University of Michigan Press, pp. 74-88.

Meurant, L. (2008). The speakers'eye gaze. Creating deictic, anaphoric and pseudo-deictic spaces of reference. In R. M. de Quadros (ed.) *Sign Languages: spinning and unraveling the past, present and future*. TILSR 9, Florianopolis, Brazil, pp. 403-414.

Millet, A. ; Estève, I. (2012). Segmenter et annoter le discours d'un locuteur de LSF : permanence formelle et variabilité fonctionnelle des unités. In A. Braffort, L. Boutorat et G. Sérasset (Ed.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, Atelier DEGELS 4-8 juin 2012*, France, pp. 57-73.

Neidle, C. (2002). Sign Stream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. Boston, MA: *American Sign Language Linguistic Research Project Report* No.11, Boston University.

Nonhebel, A.; Crasborn, O.; van der Kooij, E. (2004). Sign language transcription conventions for the ECHO Project. Version 9, SSL mouth annotations and BSL and NGT mouth annotations version 2. University of Nijmegen http://www.let.kun.nl/sign-lang/echo/docs/

Schembri, A.; Crasborn, O. (2010). Issues in creating annotation standards for sign language description. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz & A. Schembri (Eds.) *Corpora and Sign Language Technologies. [Proceedings of the 4th Workshop and the 7th LREC 2010]* Paris. ELRA, pp.210-216.

van der Kooij, E. et al. (2006). Manual prosodic cues: PALM-UP and pointing signs. Poster presented at the 9th TISLR. Florianopolis, Brazil.

van Loon, E. (2012). What's in the palm of your hands ? Discourse functions of Palm-up in Sign Language of the Netherlands, Master Thesis. University of Amsterdam, Netherlands.

Taking non-manuality into account in collecting and analyzing Finnish Sign Language video data

Anna Puupponen, Tommi Jantunen, Ritva Takkinen, Tuija Wainio, Outi Pippuri
Department of Languages (Sign Language Centre), University of Jyväskylä, Finland
P.O. Box 35, FI-40014 University of Jyväskylä, Finland
E-mail: {anna.puupponen, tommi.j.jantunen, ritva.a.takkinen, tuija.wainio}@jyu.fi & outi.k.pippuri@student.jyu.fi

Abstract

This paper describes our attention to research into non-manuals when collecting a large body of video data in Finnish Sign Language (FinSL). We will first of all give an overview of the data-collecting process and of the choices that we made in order for the data to be usable in research into non-manual activity (e.g. camera arrangement, video compression, and Kinect technology). Secondly, the paper will outline our plans for the analysis of the non-manual features of this data. We discuss the technological methods we plan to use in our investigation of non-manual features (i.e. computer-vision based methods) and give examples of the type of results that this kind of approach can provide us with.

Keywords: Finnish Sign Language, non-manuals, video data, Kinect, SLMotion, head movements

1. Introduction

This paper describes the process of collecting high quality video data in Finnish Sign Language (FinSL) and how, in the process, we took into account the investigation of non-manual elements (i.e. the movements and positions of the head and torso, eyes, eye brows, and mouth). The paper also describes how we plan to analyze the non-manual elements in the data. We present a technological method that has been specifically developed for such an analysis and, in addition, demonstrate how this method has already been used in the phonetic and linguistic analyses of FinSL head movements.

The data collection and the work with non-manuals are directly motivated by two research projects presently being carried out in the Sign Language Centre of the University of Jyväskylä, Finland. The first is the *FinSLs Corpus* project, which aims to build a high quality video corpus for the sign languages of Finland[1]. The second is the *ProGram* project, which aims to investigate the syntax and prosody of FinSL[2]. Both projects are closely linked to other current Finnish projects dealing with data collection and technological methods, most notably the *Corpus and Sign Wiki* project[3] and the *CoBaSiL* project.[4]

Large video corpora on sign languages have traditionally been collected and analyzed only in terms of manual activity (e.g. Crasborn & Zwitserlood 2008b; Johnston 2009; Wallin et al. 2010). No widely used standards for collecting and analyzing non-manual elements exist and, consequently, when non-manual elements have been investigated systematically, researchers have had to investigate them on the basis of video material that was not specifically recorded for the purpose. For the specific analysis of non-manual elements, technological methods have long depended on various utilizations of motion capture technology (e.g. Jantunen et al. 2012; Puupponen et al. 2013). However, recent developments in computer vision and image analysis techniques have also made it possible to deploy content-based video analysis methods for research into non-manuals (e.g. Karppa et al. 2011; Luzardo et al. 2013).

In the rest of this paper, Section 2 presents the collecting and processing of high definition (HD) video material, with particular emphasis on research into non-manuals in sign languages. Section 3 describes how the data can and has been used in the analysis of non-manuals. Section 4 offers a brief conclusion.

2. Video data on FinSL

2.1 Background

At the Sign Language Centre in the University of Jyväskylä, we aim to collect a corpus of FinSL and Finland-Swedish Sign Language (FinSSL). Currently, our data consists of 10 hours of multi-camera HD (1920x1080) 25-50 fps video material on FinSL, recorded in the Audio-visual Research Centre at the University of Jyväskylä. We have collected material from a total of seven pairs of informants (age 20 to 59 years) from different parts of Finland, who all performed a fixed series of seven tasks. The data includes both dialogue and monologue material.

The procedure for data collection mainly follows the conventions of earlier corpus projects of several other sign languages, e.g. German Sign Language (Hanke et al. 2010), the Sign Language of the Netherlands (Crasborn & Zwitserlood 2008a), and Swedish Sign Language (Mesch 2009). In the procedure, two signers take part in a conversation in which they first talk about themselves, their work, their hobbies or something they are interested in. The signers then take it in turns to sign from comics and tell a story from a picture book, and finally they discuss an issue that concerns the Deaf world or FinSL.

[1] http://viittomakielenkeskus.jyu.fi/projektit.html
[2] http://users.jyu.fi/~tojantun/ProGram/index.html
[3] http://www.kl-deaf.fi/fi-FI/Korpus-SignWiki/
[4] http://research.ics.aalto.fi/cbir/cobasil/

The material will be annotated in ELAN (Crasborn & Sloetjes 2008). At the time of writing, the annotation of manual activity has just begun and the annotation of non-manuals is scheduled to begin in the autumn of 2014. Also metadata is being gathered and indexed according to the IMDI standard. The metadata consists of age, sex, place of residence, school, education, age of sign language learning, languages used at home, work, languages used at work etc.

The video material, annotations and metadata are stored in Jyväskylä University's quota in the IDA storage service provided by the CSC – IT Center for Science[5]. The service will make it possible, for example, to publish the data for the use of the general public in the future.

2.2 Recording and processing the data

For the current data, the camera set-up consisted of seven Panasonic HD video cameras, illustrated in Figure 1. Of the cameras, six (Cams. 1-6) were directed towards the informants and one recorded the person giving the instructions (Cam. 7). Cam. 1 recorded an image of both of the informants (Signers A & B) facing each other, whereas Cam. 2 and Cam. 3 recorded an image of each of the signers from approximately a 45-degree angle.

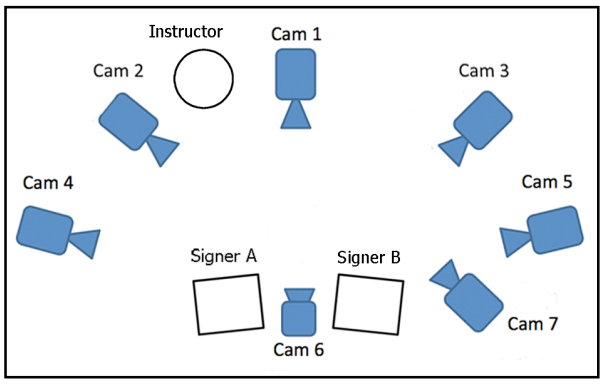


Figure 1: Camera arrangement in the recording of FinSLs corpus material.

In order to collect high quality material for research into non-manual activity, i.e. to observe more closely the movements and positions of the torso, head, and face, we had extra cameras recording close-up views of the upper body of both informants (Cams. 4 and 5). For these close-ups the cameras were positioned directly in front of the informants, so that they recorded a nearly direct image of the signers. The direct image footage with a front view of the signers makes possible computer-vision based analysis of non-manuals (see Section 3 of this paper).

To aid the analysis of the informants' signing in the dimension of depth, we had one camera recording the informants from above (Cam. 6). This camera was attached to the ceiling of the studio. Example frames of the video material from four different camera angles are presented in Figure 2.

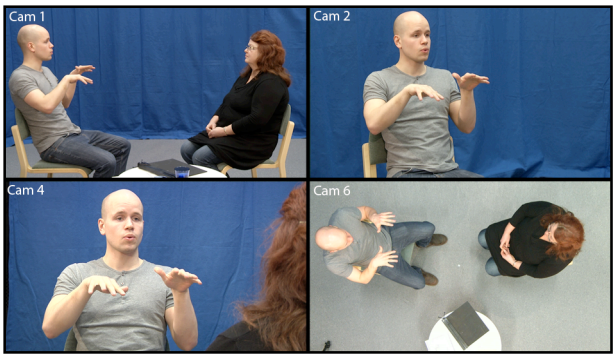


Figure 2: Screenshots of one frame in the video material from the recordings of Cameras 1, 2, 4, and 6.

The recorded video material was subsequently edited in *Adobe Premiere Pro CS6 6.0.5*. To help the editing process of multi-angle video clips, a clapperboard was used at the beginning and end of each of the seven tasks in all of the dialogues. The video material was edited so that the footage from six different camera angles (Cams. 1–6) resulted in six separate synchronised video clips for each task. The recordings of the person giving the instructions (Cam. 7) did not contain the clapperboard signals, and was therefore edited and compressed separately into one continuous clip, containing all the instructions for the different tasks in each dialogue.

All the recordings have been stored in Material eXhange Format (MXF) and compressed using H.264 in an MP4 container. The MXF container format contains time-code and metadata support and is not specific to a compression scheme. It is being used for the storage of the material to avoid restrictions in future compression. H.264 compression was used to ensure usability and compatibility between different operating systems when annotating the material in ELAN. The video material was compressed so that the annotation and analysis of both manual and non-manual activity could be done on the basis of HD material and with a reasonable file size.

2.2 Additional Kinect data

In addition to the HD video, our current material also includes data recorded with one *Kinect* motion sensing input device. In the studio, the device was stationed next to Cam. 2, where it always recorded the activity of one of each pair of informants (Signer B in Figure 1). The device was connected to an Apple MacBook Pro 15" laptop (2,6 GHz Intel Core i7) and controlled with specifically coded *NiRecorder* software, based on *OpenNi*[6] and *SensorKinect*[7] technologies. All the recordings have been stored on the hard drive of the laptop.

[5] http://www.csc.fi/english

[6] http://www.openni.org

[7] http://github.com/avin2/SensorKinect

The purpose of recording *Kinect* data was to complement the main HD video data, especially with quantitative information about depth, a dimension not inherently present in traditional video recordings. In practice, the *Kinect* data consist of a low-quality RGB video, augmented with a 16 Hz infrared video, and an automatically calculated skeleton model of the signer. Of these, the infrared video, shown in Figure 3, allows one to investigate the activity of signers in the dimension of depth to the precision of one millimetre. From the point of view of non-manuals, such data will be particularly useful in the analysis of the depth of head and body movements and postures which will be carried out in the *ProGram* project. More generally, when combined with data recorded with Cam. 6 the data make possible a very precise analysis of, for example, the spatial relationship of the hand and the rest of the body.

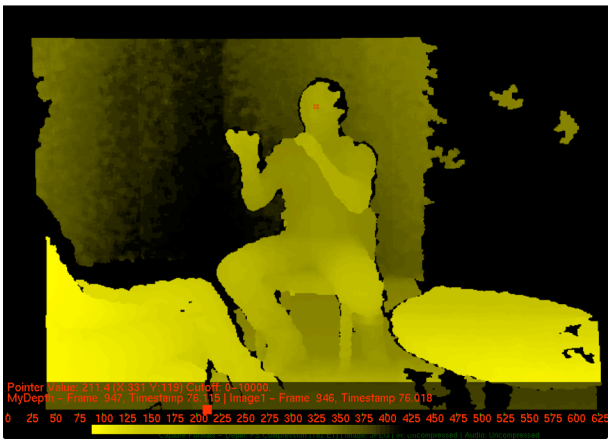


Figure 3: A screenshot showing a frame from the infrared data. The pointer value indicates the distance of the signer's forehead from the *Kinect* sensor.

The skeleton data adds further value to the analysis of the signers' movements as it provides data analogous to that collected with traditional motion capture (mocap) equipment (Chen & Koskela 2013). The skeleton figure, illustrated in Figure 4, is extracted on the basis of the depth data in real time during the recording. In practice, the skeleton figure gives a three-dimensional model of the global movements of the arms, legs, torso, and head of the signer.

The extraction of the skeleton figure is based on an algorithm that classifies a large three-dimensional point cloud into approximately a dozen human skeleton joint coordinates (Chen & Koskela 2013). This data is stored as a Comma Separated Value (CSV) file that can be easily imported to common mathematical software, such as *Matlab*, for further analysis. In terms of non-manuality, the skeleton data allows one to analyze such matters as the kinematic properties of global movements of the head and torso with a methodology developed for mocap studies (see Jantunen et al. 2012). Again, such work is planned to be carried out in the *ProGram* project.

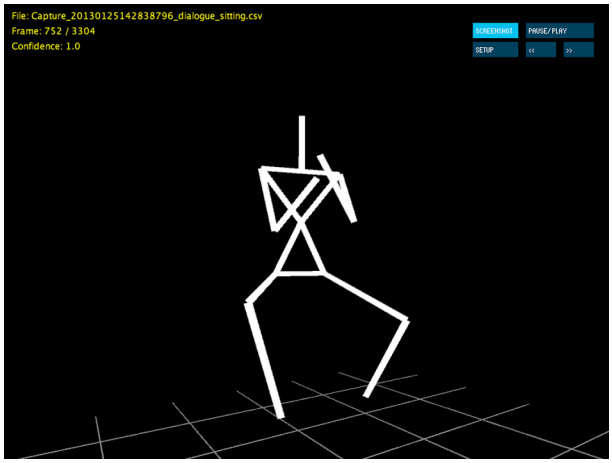


Figure 4: A three-dimensional skeleton model of a signer. The model can be viewed from different angles and from different distances.

3. Analyzing the data

3.1 SLMotion software

The high quality multi-camera video data will allow us to investigate non-manuality not only using traditional observation methods but also with various computer-vision based image analysis technologies. The main technologies that we are going to use are included in the *SLMotion* software (Karppa et al. 2014).[8] SLMotion is a tool for a near mocap-quality motion analysis of various articulators of signers visible in videos containing sign language. The first development versions of the tool focused on the hands and the head, and measured the motion of these articulators by first detecting parts of the person's bare skin on a video, then characterizing the shapes of the hands and the head with a point distribution model, and finally tracking their motion separately by the Kanade-Lucas-Tomasi algorithm and active shape models (Karppa et al. 2011). Recent development work has added to the tool the functionality to track and measure the movements and positions of the eyes, eye brows, and mouth (Luzardo et al 2013, 2014). A useful feature of SLMotion is that the quantitative results produced with it can be imported into ELAN for visualization and further analysis. This is illustrated in Figure 5 for the present data.

All the basic functions of SLMotion will be utilized in the *ProGram* project from 2015 onwards. Concerning non-manual activity, the tool will be used both as an aid to annotation and for the quantitative analyses of movements produced by the torso and the head. With respect to annotation, the ability of the tool to detect and classify, for example, eye blinks can be used to automate the manually time-consuming annotation process. Concerning the activity of the torso and head, the project will

[8] http://users.ics.aalto.fi/jmkarppa/slmotion/

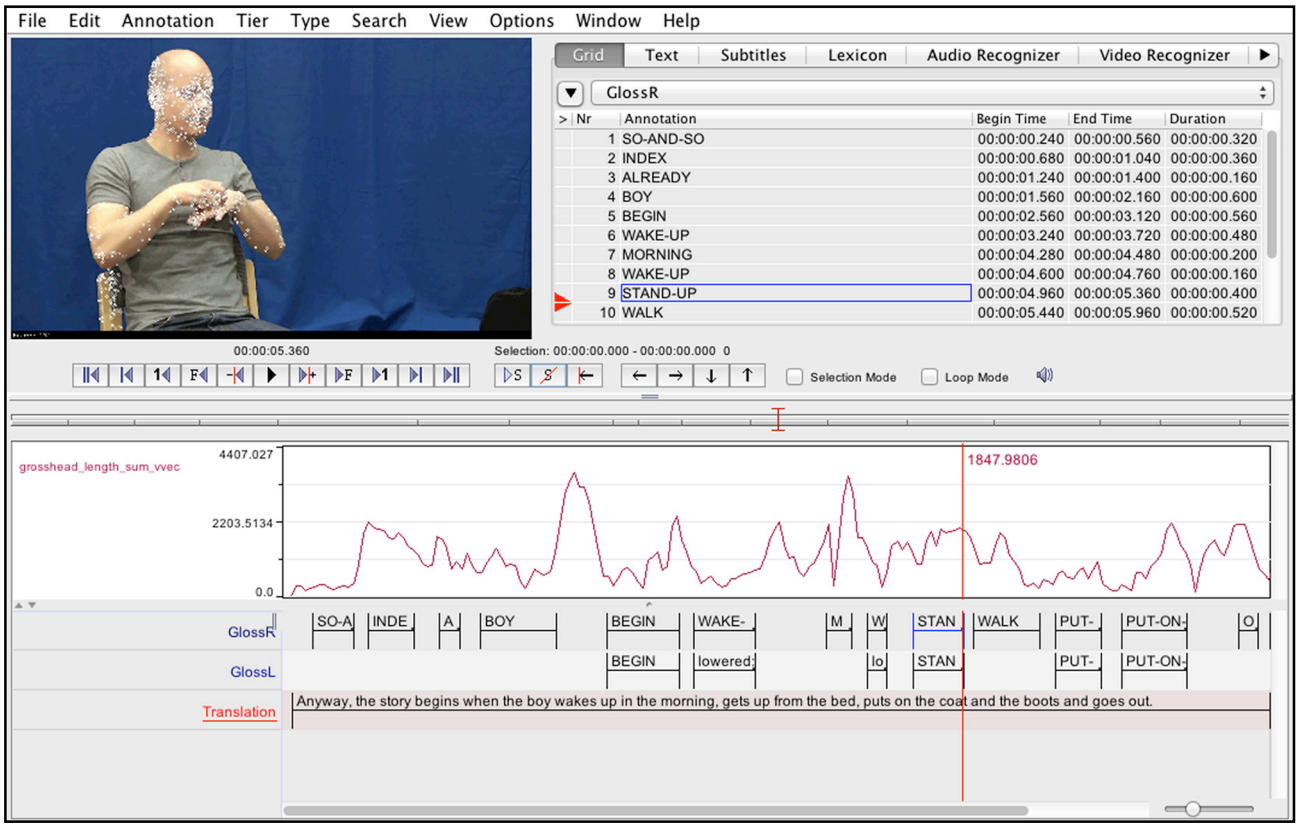


Figure 5: A screenshot from ELAN showing visualized SLMotion speed data of the movement of the head

focus on investigation of the signers' sentence-internal body and head movements and analyse them with various correlation functions for rhythm (see Jantunen et al. 2012). The quantitative data for this will be provided by the latest functions of SLMotion (see Luzardo et al. 2013).

Although non-manual elements have not yet been systematically investigated in the present data, we have already used the earlier development stages of SLMotion in the analysis of non-manual activity in FinSL (e.g. Jantunen et al. 2010; Puupponen 2012). In the following, we give examples of the results that this kind of technologically oriented analysis of sign language can produce. In particular, we will focus on head movements in horizontal and vertical dimensions.

3.2 On the head movements in FinSL

In our study of horizontal and vertical head movements in FinSL, we used traditional observation methods and SLMotion-based analysis to examine the phonetic forms and linguistic functions of articulations produced with the head (Puupponen 2012). The visualized SLMotion measurement data was found useful, for example, in identifying a particular head movement (e.g. a headshake) from the continuous stream of head movements, as well as in defining the starting and ending point of different head movements. Also a more detailed segmentation of head movements and an investigation of differences between head movements of a certain type were carried out on the basis of the numerical data.

In the data discussed in Puupponen (2012), seven different types of head movement were identified. Of these, five were included in our analysis: *nod*, *nodding* (a series of small repeated nods), *head turn*, *sideways tilt of the head*, and a *headshake*. Nodding movements and head shakes were repeated movements consisting of six to seven movement phases, whereas head nods, head turns and sideways tilts were non-repeated movements consisting of one to three movement phases. The two excluded movement types, *head thrust* and *backward pull of the head*, were produced in the dimension of depth. Because of the two-dimensionality of the video, the phonetic description and analysis of these movements was not possible with SLMotion at that time.

In general, the analysis showed that the head is very active during signing, as is demonstrated in Figure 6. It was argued in Puupponen (2012) that the continuous movement produced by the head has consequences for the annotation and analysis of head movements: identifying the head movements and distinguishing, for example, between linguistic and non-linguistic elements from the continuous stream of head movements is not always clear-cut (see also Puupponen et al. 2013).

Even though both articulators were moving continuously, the SLMotion measurements for the horizontal and vertical movements of the head did not correlate to those of the dominant hand ($r \leq 0{,}3$ in all cases). In addition, the head movements in the data were often not temporally aligned with the manually produced signing sequences (i.e. syntactic constituents).

Concerning the movement-internal features of different head movements, SLMotion analysis revealed that in most of the head movements involving repetition the amount of movement increased at the start and then decreased towards the end. This motion diminution is a feature associated with both horizontal headshakes and vertical nodding movements. The phenomenon is demonstrated in Figure 7.

The different types of head movements in the data signalled, for example, assertion and affirmation (nod, nodding), negation, semantic exclusion and hesitation (head turn, headshake, head tilt), and the end of a topic phrase (nod). The head movements also appeared at the beginning of text episodes (nod, nodding) and they made the signing textually and syntactically coherent by making meaningful use of the three-dimensional signing space (head tilt). Also in many cases two head movements occurred simultaneously. This was particularly a quality of head tilts, a fact possibly resulting from the long duration and textual-syntactic functions of head tilts, which allow the production of simultaneous head movements with, for example, emphatic functions.

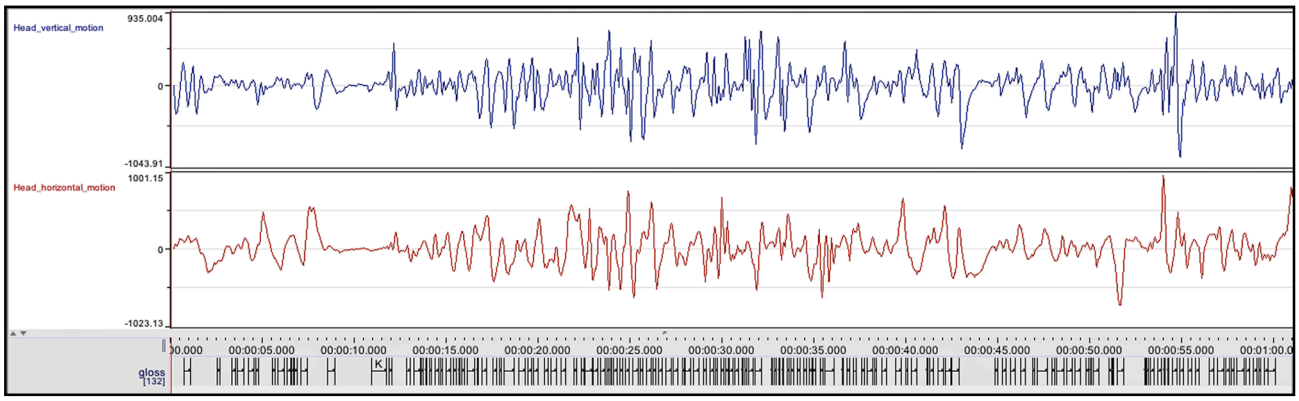


Figure 6: A screenshot from ELAN showing overall visualisations of the horizontal and vertical motion of the head in the data. The annotations of manual signs are shown in the annotation tier below the graphs.

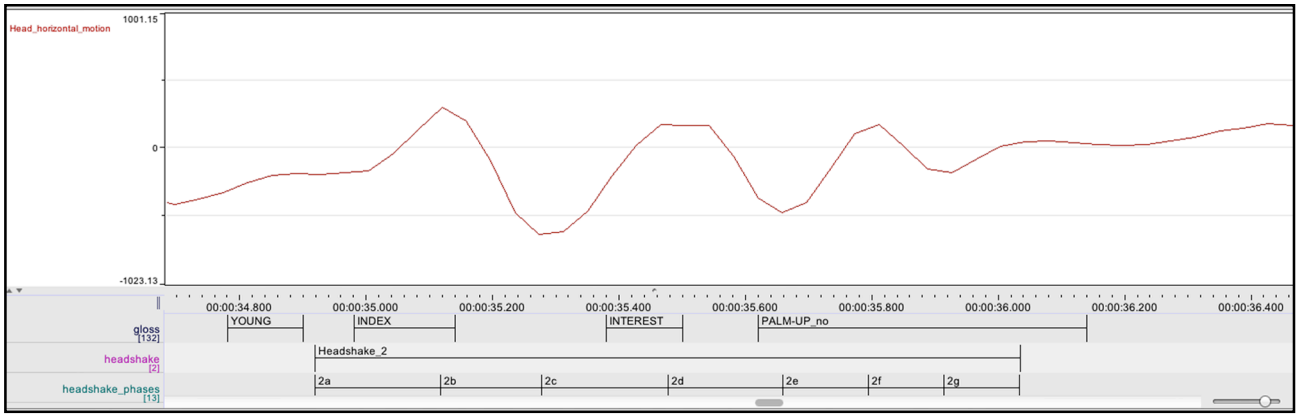


Figure 7: A screenshot from ELAN showing the visualised data of the horizontal motion of one headshake. The annotations for the manual signs, head movements, and movement-internal phases are shown on tiers below the graph.

4. Conclusion

In this paper we have discussed how we collected multi-camera HD video and Kinect material in FinSL, with particular reference to non-manuals. We have also introduced our already existing analyses of non-manuals in this type of data and presented our plans for the future. Although the analysis of non-manuals in the data that we have recently gathered for that purpose has not yet properly begun, we are convinced that knowledge of the processes we have described in this paper will also be of benefit to others working in the field.

5. Acknowledgements

The authors wish to thank Eleanor Underwood for checking the English of the paper. The financial support of the Academy of Finland under grants 269089 and 273408 is gratefully acknowledged.

References

Chen, X. & Koskela, M. (2013). Online RGB-D gesture recognition with extreme learning machines. In the *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI'13), Springer Verlag, pp. 467-474.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In the *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* [organized as a part of LREC 2008, Paris: ELRA, pp. 39-43.

Crasborn, O. & Zwitserlood I. (2008a). The Corpus NGT: an online corpus for professionals and laymen, In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, Paris: ELRA, pp 44-49.

Crasborn, O. & Zwitserlood, I. (2008b). *Annotation of the video data in the "Corpus NGT"*. Dept. of Linguistics & Centre for Language Studies, Radboud University Nijmegen, The Netherlands. Online publication http://hdl.handle.net/1839/00-0000-0000-000A-3F63-4 (accessed 15 January 2013).

Hanke, T., König, L., Wagner, S. & Matthes, S. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In the *Proceedings of the 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Paris: ELRA, pp. 106-109.

Jantunen, T., Koskela, M., Laaksonen, J. & Rainò, P. (2010). Towards the automated visualization and analysis of signed language motion: method and linguistic issues. In the *Proceedings of the 5th International Conference on Speech Prosody* (SP 2010), 100006:1-4.

Jantunen, T., Burger, B., De Weerdt, D., Seilola, I. & Wainio, T. (2012). Experiences collecting motion capture data on continuous signing. In the *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*, Paris: ELRA pp. 75-82.

Johnston, T. (2009). *Guidelines for annotation of the video data in the Auslan Corpus*. Dept. of Linguistics, Macquarie University, Sydney, Australia. Online publication http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf (accessed 15 January 2013).

Karppa, M., Jantunen, T., Koskela, M., Laaksonen, J. & Viitaniemi, V. (2011). Method for visualisation and analysis of hand and head movements in sign language video. In the *Proceedings of the 2nd Gesture and Speech in Interaction conference* (GESPIN 2011). [CD]

Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J. & Jantunen, T. (2014). SLMotion: An extensible sign language oriented video analysis tool. To appear in the *Proceedings of the 9th Language Resources and Evaluation Conference* (LREC 2014). Paris: ELRA.

Luzardo M; Karppa, M.; Laaksonen, J.; Jantunen, T. (2013). Head pose estimation for sign language video. In *Image Analysis Lecture Notes in Computer Science,* Vol. 7944, Springer Berlin Heidelberg, pp. 349-360.

Luzardo, M., Viitaniemi, V., Karppa, M., Laaksonen, J. & Jantunen, T. (2014). Estimating Head Pose and State of Facial Elements for Sign Language Video. To appear in the *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*. Paris: ELRA.

Mesch, J. (2009). Project Planning: the Swedish Sign Language Corpus. Presentation at the *Sign Linguistics Corpora Network Workshop 1: Introduction and Data Collection* in London, England, July 26-27, 2009.

Puupponen, A. (2012*). Horisontaaliset ja vertikaaliset päänliikkeet suomalaisessa viittomakielessä* [Horizontal and vertical head movements in FinSL], MA thesis, University of Jyväskylä, Jyväskylä, Finland. [http://urn.fi/URN:NBN:fi:jyu-201207242120]

Puupponen, A.; Jantunen, T.; Wainio, T. & Burger, B. (2013). Messing with the head: on the form and function of head movements in Finnish Sign Language. Presentation at the 11th Theoretical Issues in Sign Language Research conference (TISLR 11), University College London, July 10-13, 2013.

Wallin, L., Mesch, J. & Nilsson, A.-L. (2010). *Transcription guidelines for Swedish Sign Language discourse* (Version 1). Dept. of Linguistics, University of Stockholm, Sweden.

Visualizing the Spatial Working Memory in Mathematical Discourse in Finnish Sign Language

Päivi Rainò*, Marja Huovila[ó], Irja Seilola[†]
*Corpus and SignWiki Project,
Humak University of Applied Sciences & the Finnish Association of the Deaf,
Helsinki, Finland
[ó]Keskuspuisto Vocational College,
Helsinki, Finland
[†]Sign Language Centre, Department of Languages,
University of Jyväskylä, Finland
E-mail: paivi.raino@humak.fi, marja.huovila@keskuspuisto.fi, irja.r.seilola@jyu.fi

Abstract

In this paper, we will present problems that arise when trying to render legible signed texts containing mathematical discourse in Finnish Sign Language.
Calculation processes in sign language are carried out using fingers, both hands and the three-dimensional neutral space in front of the signer. Specific hand movements and especially the space in front of the body function like a working memory where fingers, hands and space are used as buoys in a regular and syntactically well-defined manner when retrieving, for example, subtotals.
As these calculation processes are performed in fragments of seconds with both hands that act individually, simultaniousity and multidimensionality create problems for traditional coding and notation systems used in sign language research. Conversion to glosses or translations to spoken or written text (e.g. in Finnish or English) has proven challenging and what is most important, none of these ways gives justice to this unique concept mapping and mathematical thinking in signed language. Our proposal is an intermediary solution, a simple numeric animation while looking for a more developed, possibly a three-dimensional representation to visualise the calculation processes in signed languages.

Keywords: Finnish Sign Language, mathematical discourse, visual working memory

1. Introduction

Mathematical problem solving discourse in Finnish Sign Language (FinSL) is carried out using fingers, both hands and the three-dimensional, neutral space situated in front of the signer – all of which are used as a kind of visual abacus or a visual working memory during the counting process. The mathematical discourse described here is part of everyday language worth of bringing forward in the corpora and descriptions of signed languages, despite the fact that signed calculations produced by sign language users are still incorrectly interpreted as merely "finger counting".

This paper deals with the a corpus that consists of seven monologues and six dialogues, in which native users of FinSL solve basic mathematics questions. Excluding literal translations and vocabularies translated from spoken language to sign language, mathematical discourse in idiomatic sign language use has not, to our knowledge, been highlighted in the descriptions of other sign languages than FinSL (e.g. Huovila & Rainò & Seilola, 2010; Rainò & Seilola, 2008).

One of the explanations lies in the fact that the calculation processes are complicated to transliterate (e.g. using glosses) or to translate legibly to spoken languages.[1]

As calculations are performed in sign language, specific handshapes denote numeric entities and moving hand constellations represent constantly varying relationships between those entities. The actual calculations are performed mentally using the visual working memory created in space in front of the signer where fingers, hands and non-manual spatial layers are used as buoys (c.f. Liddell, 2003) with which, for example, subtotals are retrieved in a regular and syntactically well-defined manner.

The use of space in arithmetic (as well as geometric) calculations in FinSL is parallel with the normal use of three-dimensional space in signed discourse where any concrete and abstract entities may be placed in front of the signing person or on her body. After reserving that location, its meaning can be activated by, for example, pointing with an index finger or even a glance until a new referent is introduced. The neutral space in front of the signer is utilized throughout the grammar in all (studied) sign languages, among other things, for pronominalisation, verb agreement and for textual grouping and semantic categorizations where e.g. paratactic items may be grouped horizontally or vertically.

[1] A vast amount of research has been conducted, however, on deaf students' learning difficulties in mathematics (cf. Bull, 2008; Hyde & Zevenbergen & Des Power, 2003; Kelly et al., 2002; Kelly & Lang & Pagliaro, 2003). Only a few studies mention the fact (for example, Foisack, 2003) that mathematics could be taught in sign language and students' thinking in sign language and visual problem-solving process could be at least as valid as operating in spoken language and using the terminology of that language.

(On the use of space, see Sandler & Lillo-Martin, 2006; Liddell, 2003; Neidle & al., 2001; Taub, 2001; c.f. FinSL Lukasczyk, 2008; Jantunen, 2003.)

2. Transliteration of signed mathematical discourse

In FinSL, cardinal and other sequential numbers are one-hand signs produced with the dominant hand. When signing the first nine cardinal numbers (1–9), palm orientation is towards the signer with fingers pointing straight up (cf. '1' in Figure 1a). Tens are signed with the palm to the side of the signer and with a slight movement downwards (cf. Figure 2), whereas 'hundreds' contain a straight movement to the side with fingertips pointing towards the centre line (Figure 1b).

Figure 1a: Cardinal numer '1' in FinSL
(Suvi, s.v. Numeraalit [Numerals])

Figure 1b: Cardinal numer '100' in FinSL
(Suvi, s.v. Numeraalit [Numerals])

Corresponding ordinal numbers – taking only one example of the vast semantic sphere of applicable morphemes for numerals – are produced by varying the palm and finger orientation and the position of the hand in the space. When signing calculations, then, the orientation of palm and fingers follow roughly those of cardinal numbers but hands are kept lower than normally and tilted slightly away from the signer (Figure 2). When a signer performs or illustrates calculations, he/she may watch his/her fingers, which is never the case in normal discourse unless the signer is recalling something and repeating his/her words sign by sign.sign by sign.

Figure 2: Numbers 8 (in the right hand) and 10 (in the left hand) used in calculations

To discuss the denotation of the visual working memory, we present its manifestation in simple tasks like additions and multiplications e.g. 3 x 8. In the example presented here (Figure 3), the calculation is first split into subcalculations: (8 + 8) + 8 where a group of two eights is placed in the index and middle finger of the non-dominant hand. Here, as in normal sign language discourse, hands may represent different entities: e.g. in multiplications the role of multiplicand is represented by the dominant hand and multiplier by the non-dominant hand (Figure 3a-c). In the latter the fingers act as so-called buoys, which represent discourse entities and the spatial relationships between them (cf. Liddell, 2003). In this example, the entity of multiplicand 8 touches the entity of multiplier 3 twice (Fig. 3a), and the intermediary sum 16 is produced with the right hand (Fig. 3b-c). Subsequently, the signer transfers the number 16 to memory in the intermediate space with a small inward movement until a third 8 is added producing the final sum, 24.

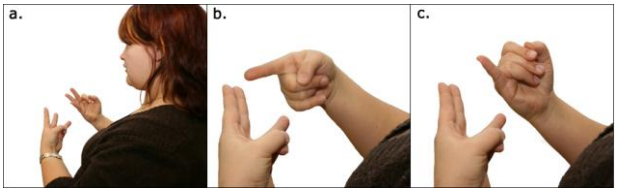


Figure 3: The process of calculating *3 x 8* in FinSL: 'The entity of *multiplicand 8* touches the entity of *multiplier 3* => 8 + 8 = 16 [+ 8 = 24]

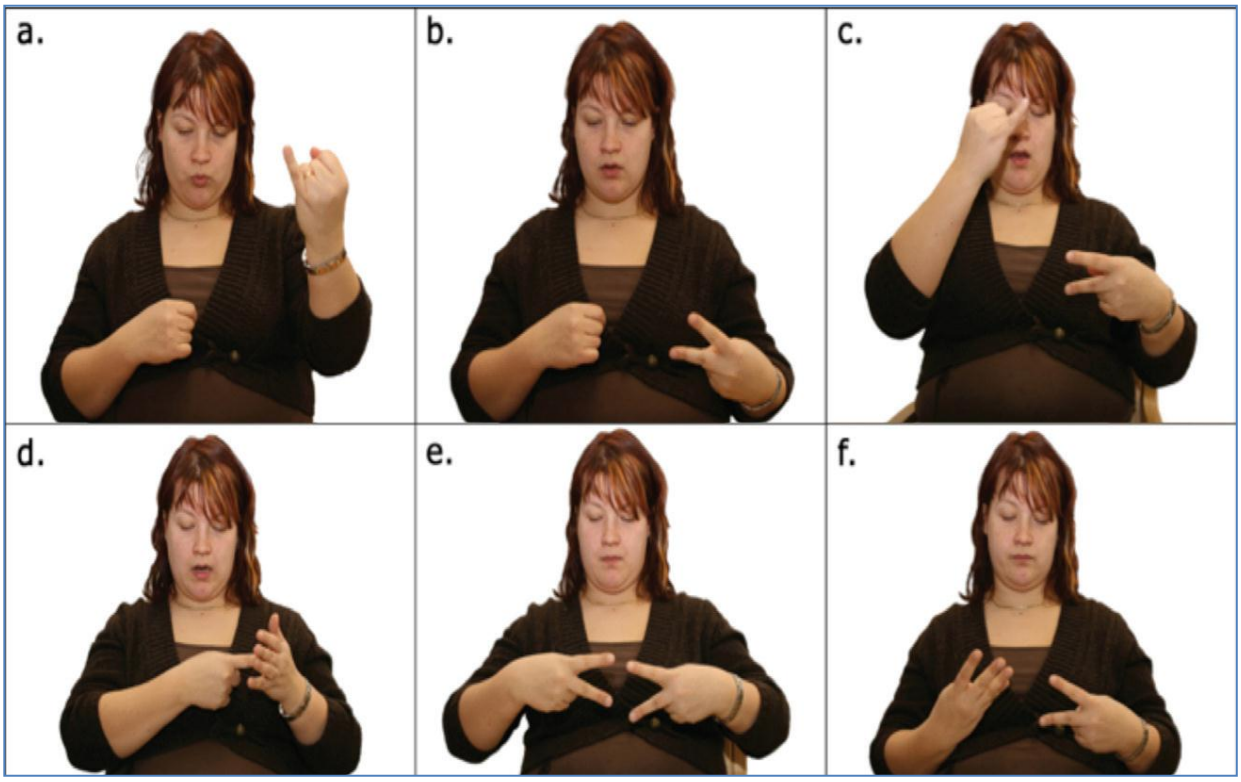


Figure 4: Calculating (3 + 3) x 2 + (3 + 9).

Besides transferring the intermediate sums with a slight movement inwards towards the signer, they can be kept in mind holding the sum in the non-active hand or positioning the sums higher in space as it were a scratch pad as can be seen in Figure 4: The first sum (3 + 3 = 6) is placed up on the left-hand side (Fig. 4a). The sum of the second calculation in brackets (3 + 9 => 12 is being signed (Fig. 4b) and kept in the intermediate memory in the signer's left hand while the 6 in memory is multiplied by 2 (Fig. 4c-d). Then the first sum (12) is taken in the right hand visualised next to the buoy '12' in the left hand (Fig. 4e). Finally, the (two) tens are moved into the non-dominant left hand and the ones into the dominant right hand (Fig. 4f). – The final sum (24), is signed using the normal orientation for cardinal numbersand using the dominant hand.

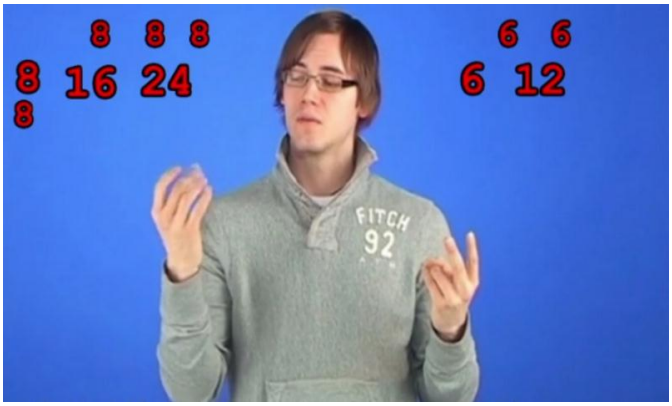


Figure 5: Visualising the process of calculating (8+8+8) + (6+6). (Animation by Mikko Palo)

3. Conclusion

When mathematical reasoning in sign is rendered in a textual representation of a spoken language (compare to the captioning of Figure 4 above), it transforms the calculation
process and the function of the hands and spatial layers in the mental scratch pad unintelligibile for the reader (or listener of the interpretation).
This is why we propose an intermediary solution – a simple numeric animation added as a layer on the video – while looking for a more developed, possibly a three-dimensional representation for the calculation processes in signed languages (cf. Figures 5 & 6).

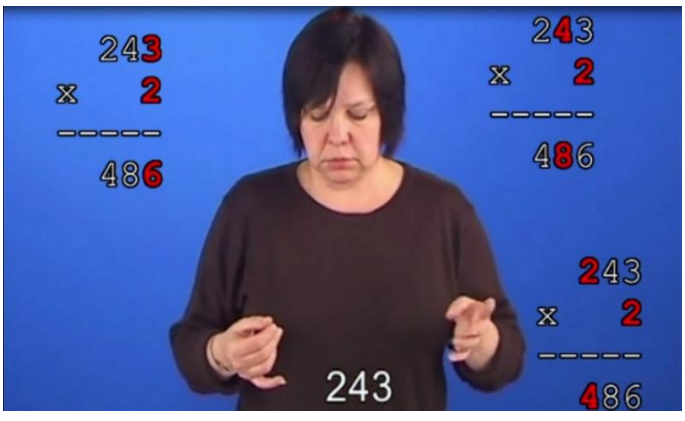


Figure 6: Visualising intermediary phases of the task 2 x 243. (Animation by Mikko Palo)

In our proposal the active/non-active state of the numeric entities in working memory is highlighted by visualising the referents in varying colours and sizes in the background of the video screen.

The corpus of mathematical discourse will be placed in Finnish SignWiki, a multifaceted open access dictionary of FinSL that uses crowdsourcing for collecting information. We hope that this non-language-dependent solution could be a way to encourage discussion and comparison of the calculation processes between users of other sign languages than FinSL, and promoting the multidimensional mathematical thinking of the Deaf people.

4. Acknowledgements

We thank Kone Foundation (Finland) for funding the Corpus and SignWiki Project 2013–2015.

We also thank Mr. Mikko Palo (Mediapalo Company) for the numeric animation of the data.

5. References

Bull, R. (2008). Deafness, Numerical Cognition, and Mathematics. In M. Marschark, P.C. Hauser (Eds.), *Deaf Cognition*. *Foundations and Outcomes*. Oxford: Oxford University Press, pp. 170--200.

Huovila, M., Seilola, I., Rainò, P. (2009). Visuaalista matematiikkaa. [Examples of Visual mathematics.] http://www.viivi.fi/osata_matematiikka/matematiikka.html. Retrieved 9.2.2014.

Hyde, M., Zevenbergen, R., Power, D. (2003). Deaf and hard of hearing students' performance on arithmetic word problems. *American Annals of the Deaf* 148 (1), pp. 56--64.

Foisack, E. (2003). *Döva barns begreppsbildning i matematik* [Mathematical conceptualisation by deaf children]. Malmö studies in educational sciences No. 7. Doctoral dissertation. School of Education, Malmö University.

Huovila, M., Rainò, P., Seilola, I. (2010). Visual calculation processes in Finnish Sign Language. Deafvoc. Conference on deaf education with a special focus on vocational education. Klagenfurt, Austria19.11.2010. http://www.deafvoc2.eu/materials/05_Visual_CalculationProcesses.ppt . Retrieved 9.2.2014.

Jantunen, T. (2003). *Johdatus suomalaisen viittomakielen rakenteeseen* [Introduction to the structure of Finnish sign language]. Helsinki: Finn Lectura.

Kelly, R.R., Lang, H.G.., Mousley, K., Stacey, M.D. (2002). Deaf college students' comprehension of relational language in arithmetic compare problems. *Journal of Deaf Studies and Deaf Education* 8 (2), pp. 120--132.

Kelly, R.R., Lang, H.G., Pagliaro, C.M. (2003). Mathematics word problem solving for deaf students: A survey of practices in grades 6–12. *Journal of Deaf Studies and Deaf Education* 8 (2), pp. 104--119.

Liddell, S. K. (2003). *Grammar, gesture, and meaning in American Sign Language*. Cambridge: University Press.

Lukasczyk, U. (2008). *Sanottua, ajateltua, tehtyä*. *Referointi kolmessa suomalaisella viittomakielellä tuotetussa fiktiivisessä kertomuksessa* [Said, thought and done. References in three fictive stories produced in sign language]. MA Thesis in Pedagogics. Department of Teacher Education, University of Jyväskylä.

Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B., Lee, R.G. (2001). *The syntax of American Sign Language*. Massachusetts: MIT Press.

Rainò, P., Seilola, I. (2008). Matemaattisen diskurssin kielioppia suomalaisessa viittomakielessä [Grammar of mathematical discourse in Finnish Sign Language]. In J. Keski-Levijoki (Ed.), Opettajankoulutus yhteisön luovana voimana: näkökulmia suomalaisesta viittomakielisestä ja viittomakielisten koulutuksesta [Teacher training as a creative force in community: Views of Finnish sign language training and education for sign language users]. *Journal of Teacher Researcher* 6, pp. 59--65.

Suvi — Suomalaisen viittomakielen verkkosanakirja 2003. [The on-line dictionary of Finnish Sign Language]. Helsinki: Finnish Association of the Deaf. http://suvi.viittomat.net. Retrieved 10.2.2014.

Sandler, W., Lillo-Martin, D. (2006). *Sign Language and linguistic universals*. Cambridge: Cambridge University Press.

Taub, S. F. (2001). *Language from the body. Iconicity and metaphor in American Sign Language*. Cambridge: Cambridge University Press.

Use of Nonmanuals by Adult L2 Signers in Swedish Sign Language – Annotating the Nonmanuals

Krister Schönström, Johanna Mesch
Department of Linguistics, Stockholm University
SE-106 91 Stockholm, Sweden
E-mail: schonstrom@ling.su.se, Johanna.mesch@ling.su.se

Abstract

Nonmanuals serve as important grammatical markers for different syntactic constructions, e.g. marking clause types. To account for the acquisition of syntax by L2 SSL learners, therefore, we need to have the ability to annotate and analyze nonmanual signals. Despite their significance, however, these signals have yet to be the topic of research in the area of SSL as an L2. In this paper, we will provide suggestions for annotating the nonmanuals in L2 SSL learners. Data is based on a new SSL as L2 corpus from our ongoing project entitled "L2 Corpus in Swedish Sign Language." In this paper, the combination of our work in grammatical analysis and in the creation of annotating standards for L2 nonmanuals, as well as preliminary results from the project, will be presented.

Keywords: Swedish Sign Language, L2 signers, nonmanuals

1. Introduction

In SSL, nonmanuals serve as important grammatical markers for different constructions, in particular with respect to the syntax required to mark negation and distinguish between different clause types (e.g. wh-questions and relative clauses). General L2 theories, such as Processability Theory (Pienemann, 1998), normally count syntactic structures as one of the most difficult grammatical stages to acquire for L2 learners of any language. We assume that SSL provides no great exception to this. To account for the acquisition of syntax by L2 SSL learners, therefore, we need to have the ability to annotate and analyze both the nonmanual signals and the manual ones within different syntactic constructions.

For signed languages, use of the nonmanuals by L2 signers has, to some extent, previously been studied in ASL (McIntire & Reilly, 1988; Emmorey, Thompson, and Colvin, 2009). However, these signals have never been the topic of research in the area of SSL as an L2. Nor has research been based on data from any L2 sign language corpus. Thus, a suitable method of annotating nonmanual signals used by adult L2 learners of Swedish Sign Language (SSL) is needed. A first step toward annotating and analyzing some aspects of grammatical errors in SSL as an L2 provides annotation suggestions for other L2 corpora.

In our first study of the data from the "L2 Corpus in Swedish Sign Language" (Mesch & Schönström, forthcoming), we propose to analyze the use of syntactic constructions. The analysis therefore includes the analysis of nonmanuals. So far, longitudinal data from four informants totaling 91 minutes have been analyzed at this stage. This paper presents some suggestions on how to annotate L2 outcomes and on how to combine these with L2 analysis, i.e. grammatical analysis with a focus on nonmanuals.

2. Building L2 Corpus in Swedish Sign Language

The first part of the L2 corpus - dataset collections 1 and 2 - consists of video recordings from 18 (14 female and 4 male) non-native signers, ranging from 18 to 40 years of age (Table 1). These L2 signers are from the central part of Sweden (11), the southern part of Sweden (3), and other countries (4). Of these, ten studied earlier at the university level, while eight had not studied at any university or college before enrolling our sign language and interpretation B.A. program. With respect to their linguistic background at the onset of the project, 11 had studied SSL for only three or four weeks, four had studied for four years, two for two years and one for five years. Only four of the students reported having a deaf friend or family member.

Age group	
18-20	5
21-25	7
26-40	6
Total	18

Table 1: Informants from the two first data recordings of "L2 Corpus in Swedish Sign Language"

We based the starting point for data collection on earlier experiences creating the SSL Corpus (Mesch et al. 2012; Mesch & Wallin, under review). The recording studio at the Department of Linguistics, Stockholm University is already equipped adequately for the SSL Corpus project. Each participant was filmed using five cameras (three cameras on floor and two cameras for a bird's eye view) in Figure 1. We created cut-outs of the face view for analysis of non-manuals and face gestures. However, we adjusted our elicitation method according to the L2 context. The data collected so far consists of:

1) Dialogues through interviews in specific target domains (e.g. family, local environment and interests) linking to appropriate L2 stages according to the Council of Europe's Common European Framework of Reference for Languages (CEFR).

2) Picture descriptions, including a single picture from the story "Frog, Where are you?" and selected pictures from the Volterra picture elicitation task (Volterra et al., 1984).

3) The retelling of a short movie clip from "The Plank".

We later added an imitation task. We propose to continue our data collection with the current group, and to collect additional data as new group of students enroll in our SSL programs.

Figure 1: The five camera views used at the recordings

All of the L2 corpus material from dataset collection 1 (53 video files) and dataset collection 2 (74 video files) has been edited and will be partly annotated using ELAN software (Crasborn & Sloetjes, 2008). A portion of this work will be made accessible online to researchers in the near future. Some video clips have been selected from the corpus as a pilot study for annotation and analysis of nonmanuals. Similar L2 corpus projects with parallel data collection are being conducted on Irish Sign Language at Trinity College Dublin, Ireland, and on American Sign Language in University of Illinois at Urbana-Champaign, USA. Thus the SSL L2 corpus can be used not only for the analysis of Swedish Sign Language, but also for comparison between L2 learners across unrelated signed languages.

3. Combining L2 analysis with annotation of the use of nonmanuals

The analysis components were twofold and linked to the annotation methods. An interlanguage analysis was carried out in combination with an error analysis of nonmanuals. Here we focused on the grammatical use of nonmanuals. In this analysis, we then focused on eyebrow movement and mouthing particularly. In an L2 research context, an analysis of language production is an important tool to have in order to account for the interlanguage of L2 learners. In our study, we therefore adopted an analysis based on an interlanguage perspective of L2 structures (see e.g. Selinker, 1972) along with error analysis. In our interlanguage analysis we marked the use of the nonmanual markers regardless of whether they were target-like or non-target-like, i.e. correct or error. Then, we used the standard tiers for eyebrow movement and mouthing. These were accompanied by an error tier in which we marked whether errors occurred (in the form tier and type tier, respectively), i.e. we presented the results of error analysis.

In this way we can account for which syntactic structures the learners have acquired and which they have not. From a longitudinal point of view, we are then able to find L2 developmental pattern in later recordings.

4. Building an annotation tool for L2 analysis

During the first analysis of the data, we have been working with the issue of how to annotate nonmanual markers in SSL, i.e. mouthing and eyebrow, gaze and head movements. We have attempted to find methods for annotating the nonmanuals, annotating L2 errors, and annotating both of these together. This will be an important issue with respect to future collaboration, i.e. sharing our L2 corpus with other researchers for cross-linguistic comparisons.

Of crucial importance are an appropriate analysis tool and an annotation standard that enable the sharing, comparing and understanding of data. In our work, we have focused on creating a working standard for annotating these L2 nonmanual markers. However, they need to be linked to the manual ones. We decide to create tiers exclusively for L2 issues for manuals and nonmanuals. (Figure 2)

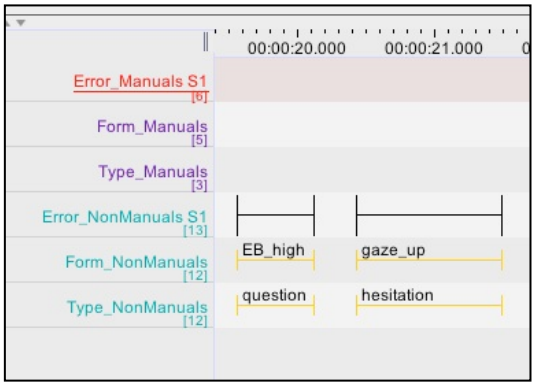


Figure 2: The error tiers for manuals and non-manuals

Each tier (manual and nonmanual) has child tiers in which there are two subcategories: one related to error forms, and another one related to error types. (Figure 3)

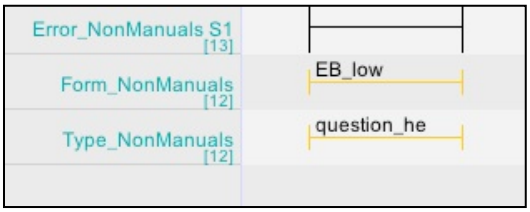


Figure 3: The nonmanual error tier

5. Creating tiers and error annotations

The analysis of the nonmanual markers is divided into two parts. The first part is connected to the use of nonmanuals generally. Here we focus on interlanguage analysis, in which we mark all the grammatical use of nonmanuals whether they are correct or not. At this stage, we have focused on use of eyebrows and mouthing. The second part deals with error analysis, in which we mark errors and, at the analyzing stage, possible errors.

5.1 Interlanguage analysis: The eyebrow and mouth tiers

Here the use of eyebrows by learners is annotated. Eyebrows play an important role in the syntactic structure of SSL. Raised eyebrows have several functions, marking, e.g. topic, y/n questions and relative clauses, whereas lowered eyebrows indicate wh-questions (Bergman, 1984).

With respect to mouthing, we decided to not include these movements in the error-form tier. Usually mouth actions are annotated as mouthings (Swedish-borrowed), mouth gestures or other mouth actions (see Crasborn et al. 2008). We expect L1 transfer among L2 learners using Swedish mouthing to a greater extent. But due to the great variability that is possible for different mouth actions, it is in some cases difficult to identify a mouth error on the basis of a single use, except for the most deviated ones, which are mouth gesture errors. These are marked as *mouth_g* in the error-form tier. Principally, this analysis follows the same standards for tiers and annotations that are implemented in the Swedish Sign Language Corpus (Wallin & Mesch, 2014).

5.2 The error analysis

In the area of general L2, error analysis is a commonly used method. At the same time, it has been subject to criticism. Our view is that this analysis provides an understanding of what errors are common among L2 learners, which can contribute to an overall understanding of the L2 learning process, along with the interlanguage analysis. This fits with our aims related to the SSL as L2 corpus project.

At this stage, while annotating the errors or the entities considered to be errors, we use a relatively broad definition of error, i.e. non-target-like constructions that differ from those in the target language. Deviations and errors, including potential errors, were marked in the analysis. These will be subject to future analytical work aimed at refining and differentiating these marked errors depending on the goals and purposes of the user.

5.2.1. Error forms

Error forms refer to L2 errors made by the learner. Here we focus on form, i.e. what is wrong? We mark forms that are errors, for instance, eye gaze, as well as eyebrow, mouthing and head movement, including a marker for non-use of nonmanuals that indicates omission.

5.2.2. Error types

Error types deal with the type of error being made. Here we use terms from the area of L2 acquisition, i.e. those related to L2 strategies, for instance, overgeneralization, overuse, simplifiers, and omissions.

6. Preliminary results

As this project is ongoing, no striking results have been found yet regarding the use of nonmanuals among L2 signers. However, several observations have been made. First, the grammatical use of nonmanuals, i.e. marking syntactic structures, is relatively limited among L2 learners at this stage. Second, most of the nonmanual behavior is related to universal human expressions.

Regarding gaze fixing, our data shows that L2 SSL learners are likely to frequently shift the gaze away from the addressee, as were observed in Emmorey, Thompson and Colvin (2009). Also we found that universal facial expressions are used to a greater extent among L2 learners, as been observed in previous research (McIntire & Reilly, 1998).

Moreover, we observed omissions of raised eyebrow in wh-questions in our data. There were examples in our data in which our subjects (L2 learners) did not raise the eyebrow in order to indicate, e.g. wh-questions non-manually (while using wh-adverbials manually). In the target language, SSL, raised eyebrow movement is required to mark wh-questions together with the use of a wh-adverb.

In our analysis, non-linguistic behavior such as hesitations and focusing are also annotated, in particular when they affect linguistic outcomes. L2 learners largely rely on focusing on how to pronounce some signs or constructions while turning their gaze away from the addressee. Another common behavior includes hesitations performed by raising the eyebrow like a hesitated question, expressing "Am I signing this correctly?" in the middle of the task.

We assume this tier to be a flexible and open one depending on research questions and what one wants to analyze.

7. Discussion and conclusions

In an L2 analysis environment, one can expect greater variability than in L1 texts. This is not only with regard to linguistic signals but also with respect to gestural ones. This pertains to human communication. An L2 learner who does not master the L2 fully produces hesitations, pauses and so on. Nonmanuals serve as channels for linguistic signals as well as gestural expressions. As a researcher, it is a challenge to keep these components apart. Sometimes these non-linguistic signals, in fact,

merely interrupt the flow of utterances, while others in fact contribute to linguistic errors. Over time, increased experience in annotating L2 data will lead to a better overall picture of how to treat these markers and the dynamic variability among L2 learners.

Future comparisons using our control group, which consists of native signers, could also contribute to a better understanding of how L2 learners use nonmanuals and how to annotate them.

In future work, we propose to describe the acquisition of syntactic structures. A description of the use of the nonmanuals, in particular eyebrow movement, is therefore determinant along with the appropriate method of how to segment text in macrosyntagms or an equivalent concept, i.e. t-units, and finally the manuals.

7.1 Limitations

With respect to the accounts regarding the use of nonmanuals by L2 signers, the amount of data analyzed in this study is still relatively small at this stage. More data is needed before substantial results can be presented as well as for the annotations to be standardized.

8. Acknowledgements

The authors would like to thank Lamont Antieau for editing the English version of this paper. This research was carried out at Stockholm University within the framework of the project L2 Corpus in Swedish Sign Language.

9. References

Bergman, B. (1984). Non-manual components of signed language. Some sentence types in Swedish Sign Language. In Loncke, F., Boyes-Braem, P. & Y. Lebrun (eds.). *Recent Research on European Sign Languages*. Lisse:Swets & Zeitlinger B.V., 49-59.

Crasborn, O., van der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behaviour of different types of mouth actions in three sign languages. *Sign Language & Linguistics, 11*(1), 45-67.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora [organized as a part of LREC 2008], Paris: ELRA, pp. 39-43.

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye gaze during comprehension of American Sign Language by native and beginning signers. *Journal of Deaf Studies and Deaf Education*, *14*(2), 237-243

McIntire, M. L., & Reilly, J. S. (1988). Nonmanual behaviors in L1 and L2 learners of American Sign Language. *Sign Language Studies*, *61*, 351-375.

Mesch, J. & Schönström, K. (forthcoming). Dataset. Corpus of Swedish Sign language as second language project 2013-2014 (version 1.) Sign Language Section, Department of Linguistics, Stockholm University. http://www.ling.su.se/teckensprak.

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012). Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University.

Mesch, J. & Wallin, L. (under review). Some gloss annotation issues in the Swedish Sign Language Corpus. *Journal of Corpus Linguistics*.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Studies in Bilingualism, Vol. 15. Amsterdam: John Benjamins Publishing.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, *10*(1-4), 209-232.

Wallin, L. and Mesch, J. (2014). Annotation guidelines for sign language texts. Version 5. [In Swedish]. Sign Language Section, Department of Linguistics, Stockholm University.

Volterra, V., Laudanna, A., Corazza, S., Radutzky, E., & Natale, F. (1984). Italian Sign Language: the order of elements in the declarative sentence. In F. Loncke, P. Boyes-Bream & Y. Lebrun (Eds.), *Recent Research on European Sign Languages*, Lisse: Swets and Zeitlinger.19-46