Measuring facial non-manual markers with a depth sensing camera: A case-study on polar questions in NGT

Lyke Esselink, Marloes Oomen, Floris Roelofsen

Aim. This paper aims to make a methodological contribution to experimental research on non-manual markers (NMMs) in sign languages. Namely, we explore the use of a *depth sensing camera* to track features of the face of a signer (e.g., brow raise, eye squint, mouth shape). As a case study, we use this method in a study of polar questions in Sign Language of the Netherlands (NGT). The method allows us to obtain fine-grained, quantitative representations of facial expressions used to express polar questions in NGT, and is straightforwardly applicable to other empirical domains and other sign languages as well.

Traditional methods. Research on NMMs in sign languages is generally based on video data. Such data, however, is two-dimensional and therefore never fully captures the actual physical reality that it represents, which is three-dimensional. Furthermore, important details are sometimes not visible on video footage because of a limited frame rate, limited resolution, motion blur, or occlusion (e.g. a hand in front of the face). Ideally, researchers would be able to base their analysis on data that captures the poses and movements of a signer, in particular NMMs, in a format that stays closer to the original, with less inherent transformation (3D to 2D), compression (frame rate, resolution), and noise (blur, occlusion).

Analysis of video data starts with annotation. This process is notoriously laborious, especially when NMMs are concerned. Even when done with great care, manual annotation has some inescapable limitations. It is inherently *subjective* (two annotators may disagree as to whether an eyebrow is raised or neutral), *not robustly reproducible* (a single annotator may label an eyebrow as raised one day, and the same eyebrow as neutral six months later), and inherently *categorical* (an eyebrow can be labeled as raised or neutral, perhaps 'half raised', but not 'raised to degree 0.35') while in reality eyebrow raise and other facial features are quantitative/continuous variables, not categorical ones—so in the annotation phase the data is further 'compressed', losing part of the original information. Ideally, researchers would have a method to annotate NMMs that is less laborious, not subjective, reproducible, and quantitative rather than categorical (meaningful categories may be identified in a later stage of analysis, but should not be imposed on us from the start).

Recent advances. Recent work by Kimmelman et al. (2020) and Kuznetsova et al. (2021, 2022) addresses the limitations of manual annotation of NMMs, building on initial proposals by Metaxas et al. (2012), Liu et al. (2014), and Puupponen et al. (2015). They use face recognition software (OpenFace) to automatically detect a signer's eyebrows and eyecorners, and compute a degree of eyebrow raise/lowering in terms of the distance between these. This method to extract degrees of eyebrow raise/lowering from video data is automatic, objective, and quantitative. However, there are still some limitations. First, measurements of relevant facial features like brow raise are *indirect* and *not robustly reproducible*. OpenFace detects facial landmarks. Features like brow raise have to be derived from distances between landmarks, but this cannot be straightforwardly done in a reliable way because these distances partly depend on the distance and angle between the camera and the signer's face (as discussed by Kuznetsova et al., 2021), which are impossible to keep constant across and even within recordings. Second, the proposed method still takes 2D *video data* as its starting point. This is what OpenFace takes as its input. So, while this body of work makes an

important first step in addressing the limitations of manual annotations, it does not address the issues of inherent transformation, compression and noise associated with video data.

Proposal: using a depth sensing camera. We explore a way to overcome these issues, at least partly, by using a depth sensing camera in addition to ordinary video cameras for data collection. Specifically, we make use of a TrueDepth camera built into an iPhone 13 in combination with the free Live Link Face application by Epic Games. This hardware/software combination can be used to measure 61 facial features, called ARKit *blendshapes*. Not all 61 ARKit blendshapes (click here for a full list) are relevant for the study of NMMs in sign languages. For our purposes, we selected 9 relevant blendshapes (motivation for this choice will be provided in the paper): BrowInnerUp, BrowOuterUp, BrowDown, EyeWide, EyeSquint, CheekSquint, NoseSneer, MouthShrug, and MouthFrown. Blendshape coefficients are values between 0 to 1, indicating the degree of engagement for each feature. Blendshape coefficients are measured at a frame rate of 60 fps.

Unlike OpenFace, which performs landmark detection based on video input, this method thus bypasses the main issues associated with video data, and moreover directly measures facial features that are of interest for sign language research as opposed to landmark coordinates, which first have to be translated into feature coefficients, something which, as mentioned above, cannot always be done in a straightforward way, if at all.

Case study: polar questions in NGT. As a concrete case study, we collected data on the use of facial NMMs in polar questions in NGT. Previous work in this empirical domain (Coerts, 1992; de Vos et al., 2009) mainly focused on eyebrow movement, and found much variation—in particular, both raised and lowered brows often occur. Our experimental design controlled for two contextual factors which may influence the way in which a polar question is expressed: prior speaker belief and immediate contextual evidence concerning the question radical. For instance, when prompted to ask *Is the zoo open?* a participant may be given prior information (through role play with a confederate) that the zoo is probably open, but be faced with immediate contextual evidence (through role play with another confederate) that the zoo is actually probably closed, and similarly for other combinations of prior belief and immediate contextual evidence. We recorded participants with a depth sensing camera as well as an ordinary video camera. This allowed us to gather fine-grained, quantitative data on the facial NMMs that are used in the expression of polar questions in NGT, across contexts and participants. The data is very rich and lends itself to various types of quantitative analyses. Space prevents us from discussing these in detail here. We highlight the fact that a clustering analysis yields three main clusters of facial expressions. Cluster A is characterized by high values of BrowInnerUp (0.75), BrowOuterUp (0.67), and EyeWide (0.82); cluster B by high values of BrowDown (0.60) and moderately high values for EyeSquint (0.39); and cluster C by low values ($\leq$0.20) for all blendshapes, thus containing relatively neutral facial expressions. Our contextual manipulations clearly affect which facial expressions are used in a polar question. For instance, while in general expressions from cluster A (brow raise, eyes wide open) are much less commonly used than ones from cluster B (brows furrowed, eyes squinted), 19% vs 39%, they are slightly *more* common when there is neutral prior belief and positive contextual evidence, 35% vs 33%. On the other hand, they are *never* used when there is positive prior belief and negative contextual evidence. Further analyses and results will be discussed in the full paper. All data and analyses scripts will be made freely accessible and reusable.

Selected clickable references Kimmelman et al. (2020) Eyebrow position [...] in KRSL. Kuznetsova et al. (2021) Using computer vision to analyse NMM [...] in KRSL. De Vos et al. (2009) [...] Questions in NGT.

References

Coerts, J. (1992). *Nonmanual Grammatical Markers: An Analysis of Interrogatives, Negations and Topicalisations in Sign Language of the Netherlands*. PhD thesis, Universiteit van Amsterdam.

de Vos, C., van der Kooij, E., and Crasborn, O. (2009). Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and speech*, 52(2-3):315–339.

Kimmelman, V., Imashev, A., Mukushev, M., and Sandygulova, A. (2020). Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PloS one*, 15(6).

Kuznetsova, A., Imashev, A., Mukushev, M., Sandygulova, A., and Kimmelman, V. (2021). Using computer vision to analyze non-manual marking of questions in KRSL. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 49–59.

Kuznetsova, A., Imashev, A., Mukushev, M., Sandygulova, A., and Kimmelman, V. (2022). Functional data analysis of non-manual marking of questions in Kazakh-Russian Sign Language. In *Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*.

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. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4512–4518.

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 Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2414–2420.

Puupponen, A., Wainio, T., Burger, B., and Jantunen, T. (2015). Head movements in Finnish Sign Language on the basis of motion capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls. *Sign Language and Linguistics*, 18(1):41–89.