MSc Thesis Artificial Intelligence

Radboud University Nijmegen

Computer Vision and Machine Learning for the Analysis of Non-Manual Markers in Biased Polar Questions in Sign Language of the Netherlands

Author:
L.D. Esselink
s4686969
l.d.esselink@uva.nl

Internal Supervisor:
Prof. M.A Larson
m.larson@cs.ru.nl
Radboud University

External Supervisor:
Dr. F. Roelofsen
f.roelofsen@uva.nl
University of Amsterdam

Second reader:
Dr. L.F.M. ten Bosch
louis.tenbosch@ru.nl
Radboud University

December 18, 2022

Abstract

This research project carried out a methodological exploration of the application of Computer Vision (CV) technologies for gathering data for sign language research, and of the application of Machine Learning (ML) techniques to analyse such data. In order to explore these methods, we applied them to the specific domain of analysing prototypical facial expressions for question marking in Sign Language of the Netherlands (NGT), otherwise referred to as non-manual markers (NMMs). Typically, the literature on this subject describes only one way in which questions are marked. However, there is more variation in the NMM of polar questions than is generally acknowledged in the literature (de Vos et al., 2009; Klomp, 2021). This research project aims to investigate if bias can account for part of this variation in NMM of polar questions in Sign Language of the Netherlands (NGT).

We hypothesized that speaker bias could account for some of these variations in facial expressions, and that there are multiple variations with restrictions relating to applicability in specific contexts. Data was elicited in an experiment through role-play type conversations between participants and confederates. A 3D depth camera was used to measure engagement of 61 facial features, and the ML method *clustering* was used to find prototypical facial expressions in the data.

We found two overarching prototypical facial expressions for marking polar questions in NGT. One included raised eyebrows and eyes wide open, whereas the other included furrowed eyebrows and squinted eyes. Further, we found several variations within these two facial expressions, varying in feature intensity and in additional non-manual markers such as a mouth frown. We found two definite constraints of these facial expressions relating to their applicability. The first was that facial expressions with raised eyebrows do *not* occur in contexts in which a person holds a positive prior belief and is subsequently presented with negative contextual evidence. The second constraint was that facial expressions involving significantly frowning mouth do *not* occur in the beginning of a question, but only in the middle and end. Future work includes further examination of these clusters by crossing contexts and temporal windows to investigate the temporal development of these facial expressions and whether further constraints can be found.

Contents

Chapter 1

Introduction

This research project carries out a methodological exploration of the application of Computer Vision (CV) technologies for gathering data for sign language research, and of the application of Machine Learning (ML) techniques to analyse such data. In order to explore these methods, we apply them to the specific domain of analysing prototypical facial expressions for question marking in Sign Language of the Netherlands (NGT). Typically, the literature on this subject describes only one way in which questions are marked. However, it is known that there is more variation within these facial expressions than generally acknowledged in the literature (de Vos et al., 2009; Klomp, 2021). We hypothesize that speaker bias could account for some of these variations in facial expressions, and that there are multiple variations with restrictions relating to applicability in specific contexts.

The question of how bias influences non-manual marking in polar questions in NGT is part of a larger project at SignLab Amsterdam, where this research project is carried out. We will refer to this larger project as the 'Biased Questions in NGT' (BQ-NGT) project, to distinguish between that and the current research project. For the BQ-NGT project, an experiment was set up that specifically controlled for context. In this experiment, participants conversed with confederates (both deaf) in a role-play setting. The goal of the experiment was to elicit polar questions in various conditions. Confederates introduced combinations of positive/neutral/negative original speaker bias, and positive/neutral/negative contextual evidence, after which the participants would ask the target question.

Traditionally, sign language research is qualitative rather than quantitative. The CV and ML methods considered in this research project allow us to investigate whether there are quantifiable differences in prototypical facial expressions for question marking. As we do not know the extent to which these variations may exist, or what they might look like, we will apply a ML method called *clustering*, which is a technique that attempts to differentiate between groups (or 'clusters') of samples in the data. This machine learning technique is typically applied to unlabeled data in order to infer new information about it. The data in our project is unlabeled, meaning that we do not know beforehand what facial expressions may result from specific combinations of non-manual markers and their respective amplitudes. Moreover, our aim is to infer new information about the data, specifically what 'groups' of facial expressions can be found in it. Further, we aim to map these facial expressions to their contexts, to explore potential constraints on the applicability of these facial expressions.

This Chapter provides an introduction to the context of the research project and our research questions. We highlight some of the most important aspects of sign languages that are relevant to this project in Section 1.1. Next, we discuss the process of traditional sign language research in Section 1.2. We introduce the research questions of this project in Section 1.3. Section 1.4 provides an overview of relevant literature. We discuss the proposed methodological refinements in Section 1.5. This Chapter concludes with the hypothesis in Section 1.6.

1.1 Sign languages

In a recent report, the World Health Organization estimates that 20% of the world's population is affected by hearing loss, and around 5.5% of the world's population experiences disabling hearing loss (2021). It estimates that, in 2050, 1 in 4 people will experience hearing loss, and that nearly 1 in 14 will live with "moderate or higher levels of hearing loss in the better hearing ear" (World Health Organization, 2021). In the Netherlands, it is estimated that there are 795.000 people that are hard of hearing or deaf, including and estimate of around 15.000 people that rely on sign language as their primary form of communication (Volksgezondheid en Zorg, 2022; European Union of the Deaf, 2022).

We will begin by highlighting some of the most important aspects of sign languages. A complete and comprehensive overview is given in Baker et al. (2016). First, one untrue assumption most people make is that sign language is universal: like spoken languages, sign languages evolved in their communities around the world, differing from region to region and include their own dialects and grammatical constructs. As a result, there is not necessarily a correlation between the spoken language of a country and its sign language (e.g. two English-speaking countries can have different signs and word order for a sentence that is otherwise identical in the spoken language).

Second, an important aspect of sign languages is that the meaning of a sign does not only rely on the *manual* component of a sign (e.g., handshape, location, movement), but is also largely dependant on *non-manual* markers (NMMs – e.g., eyebrow movement, mouth gestures, movement of head and/or body) (Baker et al., 2016).[1]

Besides being part of many sign's lexical makeup, NMMs are often used to convey grammatical information as well – comparable to pitch or intonation in spoken languages (Crasborn, 2006). Sign language linguists use so-called *glosses* to represent sign language utterances. For instance, the gloss in (1) represents the Dutch Sign Language (NGT) translation of the question '*Is the entry to the park free?*'.

(1) $\overline{\text{ENTRY PARK FREE}}^{\text{q}}$

Lexical signs are written in small-caps, and consist of at least a manual component which is sometimes accompanied by non-manual components. The upper tier of a gloss shows its grammatical NMMs, of which the duration is indicated by the horizontal line. The NMM 'q' refers to a combination of features that indicate a polar question. In (1), we see that it occurs for the entire clause.[2] A more detailed overview of the use of NMMs is given in Section 1.4.1.

Third, sign languages are low-resource languages: they have not yet been studied as much as most written languages, and their data sets are meager compared to those for written languages (de Coster and Dambre, 2022). There is still much to be uncovered, but the process to do so is costly – in terms of both money and time.

1.2 Traditional sign language research

The traditional approach to collecting and annotating sign language data has a number of limitations: 1) it is a very laborious process that is 2) not objective, 3) not reproducible, and 4) categorical.

Video recordings are necessary to study sign languages, as they are articulated visually. These recordings need to be annotated manually, which is a tedious and time-intensive process that requires a high level of concentration. Annotations are categorical, meaning that a feature either falls into a category or it does not; there is nothing in between.[3] For instance, consider a video in which a participant slowly raises their eyebrows from a neutral expression, holds this position for a few seconds, and then transitions back into a neutral expression. An annotator has to decide at which frame the eyebrows are 'raised', and at which frame they are 'neutral'. The transition period is disregarded, as is the overall amplitude with which the NMM is expressed.

Because of this, one can imagine why researchers may not always agree on which specific label to assign to a feature. One researcher could even disagree with themselves: they might mark a movement as an eyebrow raise on one day, and hesitate to mark it as such on another. The same data set could therefore be annotated quite differently depending on the researcher and the day. Variation within and between annotators certainly does not invalidate human-annotated ground truth. However, these manual annotations are not *necessarily* an objective golden standard.

An iterative annotation method is often employed in order to ensure that annotations are as constant and accurate as possible. A set of annotation guidelines is established, which is then used by multiple researchers to independently annotate videos. Annotations are then compared, discussed, and adjusted until a satisfactory inter-annotator agreement is reached. This entire process is highly intensive and can take months – even for a small data set – but an annotated data set is strictly necessary *before* a topic can be studied.

1.3 Aim

The contribution of this research project is twofold, methodological and empirical. This Section will first discuss the methodological contribution and the accompanying research questions. To investigate these questions, the methods of interest will be applied in a specific domain that will contribute empirical

[1] The above paragraphs are adapted from the internship report.

[2] The above paragraph is adapted from earlier work in Roelofsen et al. (2021b).

[3] This is not the case for all human annotators across fields. In some fields, annotations are created with a slider that does provide an amplitude. However, in the field of sign language linguistics this is typically not the case.

knowledge about NMMs in NGT. The research questions and sub-questions of the empirical contribution steering this research project will be discussed in Section 1.3.2.

1.3.1 Methodological contribution

The application of advancements in Computer Vision (CV) and Machine Learning (ML) can support sign language linguists at various phases of research. For example, these techniques could speed up the annotation process and increase precision, or help infer information about the use of linguistic properties such as NMMs.

Compared to manually annotated data, computationally annotated data is not directly a golden standard either. However, machine learning annotations will always be reproducible: input x will always lead to the same output y. In that sense, machine learning annotations are more objective than human annotators. It is important to keep in mind, however, that machine learning annotations could simply be inaccurate, or perhaps they might be accurate but miss important nuances that a human annotator might not miss.

Moreover, features can be assigned a range of values indicating the amplitude with which they are expressed, rather than being assigned to a category in a black-or-white manner.

The application of CV and ML for sign language research is still in an exploratory phase, and has only been adopted in a few cases. These will be discussed in Section 1.4.3 below. The research questions for the methodological contribution of this paper are defined as:

RQ1. **How can we use Computer Vision technology to collect data on non-manual markers in sign languages?**

RQ2. **How can we use techniques from Machine Learning to analyse such data?**

The methodological contributions that will be investigated in RQ1 and RQ2 will not only pertain to the specific domain of biased polar questions in NGT. Rather, they will be relevant for the field of sign language research as a whole, irrespective of the language or context studied.

1.3.2 Empirical contribution

The concrete empirical domain that we will focus on is that of non-manual marking in *biased polar* questions in varying situations. A *polar* question is one in which the speaker expects a yes/no answer from the addressee. In some situations, the speaker has previous beliefs, or *bias*, on what the answer to their question will be. In the case of (1), for instance, the speaker's belief could be that the entrance to the park is indeed free. This is referred to as a *biased* polar question.

The empirical aim of this project differs in an important way from earlier related work. As will be discussed in the literature review in Section 1.4 below, most linguists focus on entire classes when studying NMMs, considering all different types of polar questions to fit in one category. However, more variation exists within these classes than is typically acknowledged (de Vos et al., 2009; Klomp, 2021). For instance, consider the following different ways to pose the same question in (spoken) English: "Did you invite Alice?"; "Did you invite Alice, or not?"; "You invited Alice, right?"; "Didn't you already invite Alice?". Some of these indicate a certain expectation of what the answer will be, or are the result of a clash between prior belief (bias) and contextual evidence.

For sign languages, while variations in non-manual markers can be expected, previous researchers have not controlled for context to take them into account. Moreover, variations in non-manual markers have sometimes been reported – albeit in a limited manner – but it has not been previously investigated how these variations map to their contexts. The empirical aim of this research project is to specifically investigate these points. The accompanying research question and its sub-questions are defined as:

RQ3. **How are different types of biased polar questions marked in Sign Language of the Netherlands?**

(a) Based on a literature review, what are the results we expect to see?

(b) What are prototypical facial expressions for marking polar questions?

(c) How do variations of these facial expressions map to their respective contexts?

(d) What is the temporal progression of relevant non-manual markers during biased polar questions?

1.4 Literature review

This Section first provides a brief overview of how NMMs are used in general, then discusses the literature on NMMs in question marking cross-linguistically, and finally zooms in on question marking in NGT specifically. After this, the interaction between affective (emotional) and linguistic functions of the eyebrows are considered for questions in NGT. The Section concludes with an overview of previous work in which CV and/or ML techniques have been used.

1.4.1 Non-manual markers

General use cases Non-manual markers have use cases at various levels, such as phonological, morphological, and syntactical. At the phonological level, NMMs can be an essential part of a sign's lexical make-up (in the same way that handshape, location, and movement are) (Pfau and Quer, 2010). At the morphological level, NMMs can modify the meaning of nouns without the need for a manual adjective (e.g. in German Sign Language, puffed cheeks are used to express that an object is larger than one would expect); Moreover, NMMs can be used to modify verbs as an indication of how an event took place (e.g. frantically executing a sign to express that it happened in a rush) (Pfau and Quer, 2010; Klomp, 2021). At the syntactic level, NMMs can be used to negate or affirm a phrase. For instance, a signer could respond to the question posed in (1) with:

(2) ENTRY PARK $\overset{\text{hs}}{\overline{\text{FREE}}}$

While the manual signs in (2) remain identical to those used in (1), the NMM is now a 'head shake' that only lasts for the word FREE instead of the entire duration of the phrase. This strictly non-manual difference changes the meaning of the utterance to '*The entry to the park is not free*'.[4] There are many other ways in which NMMs are used in sign languages. For more information, see Pfau and Quer (2010); Cecchetto (2012); Crasborn (2006); Benitez-Quiroz et al. (2014).

Question marking In addition to the above, NMMs are used for question marking. For most sign languages, they are even the only discriminatory feature between a declarative statement and a polar question, as the manual signs do not change (Pfau and Quer, 2010). Two types of questions are usually studied by sign language linguists: polar questions and wh-questions. The exact definition of non-manual marking in polar questions ('q') varies slightly between researchers both cross-linguistically and language internally, but typically involves a combination of 'raised eyebrows', 'head tilted forward', 'body tilted forward', 'eyes wide', and 'continuous eye contact with the addressee' (Cecchetto, 2012; Pfau and Quer, 2010; Zeshan, 2004; Coerts, 1992). The former three features are almost always included in 'q', while the latter two are sometimes omitted (Cecchetto, 2012; Coerts, 1992). While some languages also have a manual sign that signifies a question, they are not sufficient to mark a polar question by themselves (Coerts, 1992). In most sign languages, 'q' lasts for the entire duration of a clause; in fact, "in the majority of utterances, ... the brow raise in 'q' has begun and reached apex by the time the first manual sign begins, and it continues past the articulation of the last sign" (Coerts, 1992).

In contrast to polar questions, wh-questions are usually marked with 'furrowed eyebrows', often combined with 'head tilted backwards' (Pfau and Quer, 2010). In some cases, they may be marked solely manually (Cecchetto, 2012). Cross-linguistically, wh-question markers ('wh') mostly differ by which part of the clause the NMMs are active on (Cecchetto, 2012; Pfau and Quer, 2010). However, Cecchetto (2012) stresses that, "one should be very cautious when drawing a generalization from these data, since the set of sign languages for which the relevant information is available is still very restricted, not to mention the fact that much controversy remains even for better studied sign languages". Moreover, it is known that, language internally, a lot of variability can exist *within* these classes (Klomp, 2021). For a nuanced and detailed overview of research on NMMs for polar and wh-questions between and within sign languages, see Cecchetto (2012); Pfau and Quer (2010); Coerts (1992).

Sign Language of the Netherlands Non-manual question marking in NGT generally aligns with the definitions provided above. Coerts (1992) defines the NMM of polar questions in NGT as 'raised eyebrows' and 'head tilted forward'. In some cases, 'eyes wide' and 'body tilted forward' were observed, but no evidence was found that these are strictly caused by regional or age-related variance (Coerts, 1992). The duration of relevant features of 'q' varied, but Coerts (1992) found that "'q' must be present at least during part of each sign that falls under its scope". A manual marker, PALMS-UP, may be added, but does not replace 'q'. On the other hand, 'wh' is marked by 'furrowed eyebrows' and 'head tilted backwards' (Coerts, 1992). In most cases, 'furrowed eyebrows' preceded 'chin up', with an onset preceding or happening simultaneously with the onset of the manual signs.

The literature discussed in this Section provides background information about the use cases of non-manual markers, and the combinations of non-manual markers we may expect to find in prototypical facial expressions for polar questions.

1.4.2 Affective and linguistic functions

de Vos et al. (2009) studied the interaction between the linguistic and affective (emotional) functions of the eyebrows for question marking in NGT. They compare facial expressions in sign languages to intonation in spoken languages, which can convey both "paralinguistic information" and "grammatical functions". de Vos et al. (2009) focused on the function of

[4] The above example is adapted from my earlier work in Roelofsen et al. (2021b); Esselink et al. (2022b)

the eyebrows in expressions of surprise (raised) and of anger (furrowed). They used the Facial Action Coding System (FACS) developed by Ekman et al. (2002), which groups facial expressions into 'Action Units' (AUs) based on muscle activity (de Vos et al., 2009). The relevant AUs for this study were "AU 1, the Inner Brow Raiser; AU 2, the Outer Brow Raiser; and AU 4, the Brow Lowerer", which could be used "individually or simultaneously in varying combinations" (de Vos et al., 2009). In these terms, 'q' of a neutral polar question (without affective expression) would be defined as the combination AU 1+2, and 'wh' of a neutral wh-question would be defined as AU 4. However, in contrast to previous research, de Vos et al. (2009) found that 'q' was not always characterized by AU 1+2. Rather, in one-third of the cases, AU 1+2+4 was used. Similarly, 'wh' was expressed with AU 1+2 instead of the expected AU 4 in slightly more than one-third of the cases. Moreover, "combinations of AU 1+2 and 4 occurring simultaneously and/or sequentially were frequently found" (de Vos et al., 2009). They attribute this variation in NMMs to the more fine-grained coding system.

In addition to the above findings, de Vos et al. (2009) found significant differences between the neutral versions of 'q' and 'wh' as opposed to the surprised and angry versions. The angry polar questions were marked solely by AU 4 78% of the time; the rest with a sequential and/or simultaneous combinations of AU 1+2+4. In this case, the affective function of AU 4 appeared to override the linguistic function of AU 1+2. Similarly, the angry wh-questions were marked solely by AU 4 72% of the time; a sequential and/or simultaneous combinations of AU 1+2+4 23% of the time; AU 1+2 3% of the time; and a neutral brow position 2% of the time (de Vos et al., 2009). Although the affective and linguistic functions both solely expect AU 4 to be used in this situation, AU 1+2 appeared in more than one-fourth of the cases. Additionally, the intensity levels of AU 4 were higher in the case of angry wh-questions than in the case of neutral wh-questions (de Vos et al., 2009).

On the other hand, surprised polar questions were marked with AU 1+2 only 58% of the time; the rest with a sequential and/or simultaneous combinations of AU 1+2+4 (de Vos et al., 2009). Similar to angry wh-questions, although both affective and linguistic functions solely expect AU 1+2 to be used in this situation, AU 4 appeared in more than 40% of the cases. However, the intensity levels of AU 1+2 were again higher in the case of surprised polar questions than in the case of neutral polar questions – albeit to a lesser extent than the case of AU 4 above (de Vos et al., 2009). Surprised wh-questions were marked by a sequential and/or simultaneous combinations of AU 1+2+4 68% of the time; AU 4 16% of the time; and AU 1+2 16% of the time (de Vos et al., 2009). Although the affective function of AU 1+2 did not override that of AU 4, it has a clear influence on the NMMs in these questions.

The authors conclude that, when the affective and linguistic functions overlap (i.e. surprised polar questions and angry wh-questions), intensity levels of the affective markers are raised in comparison to their neutral counterparts (de Vos et al., 2009). They attribute a "phonetic strength" to AU 4, stating that it is stronger than AUs 1 and 2, independent of the original (linguistic or affective) function. Finally, de Vos et al. (2009) conclude that in NGT, an interaction between the linguistic and affective functions of the eyebrows may affect the position of the signer's eyebrows.

The work discussed in this Section illustrates some variations within NMMs in questions that contrast with typical definitions of how questions are marked in NGT. We discuss this work to a high level of detail as it provides interesting insights about the interaction of context and the position of the signer's. Especially the "phonetic strength" attributed to AU 4 could be of interest to the present study, as we might thus expect to see 'furrowed eyebrows' more often than 'raised eyebrows'.

1.4.3 Computational research

This Section provides an overview of previous work that used CV and ML techniques to perform computational research on the use of NMMs in sign languages. As applications of NMMs have been discussed in Section 1.4.1, this Section focuses on the methodologies of these studies rather than their results.

Metaxas et al. (2012) found that most research on this subject was sensitive to occlusions of the face, and were limited to recognising *low-level* features (e.g. head gestures, eyebrow movements). To overcome these limitations, they proposed a new framework that uses 2D to 3D mapping to extract geometric and appearance features. First, landmark detection was used to track the location of facial landmarks (such as the position of the eyebrows, eyes, and nose) from a 2D video. These facial features were then mapped to a 3D face model, and transformed to a frontal view of the face. With the use of a Hidden Markov Support Vector Machine, NMMs were able to be detected and classified.

Liu et al. (2014) built on the framework proposed by Metaxas et al. (2012). Again, first tracking face landmarks and mapping 2D features to a 3D model. They extracted low-level features on the basis of individual frames, and *high-level* features (combinations of gestures) on the basis of multiple frames. A model of Conditional Random Fields trained on a combination of low- and high level features then recognised and identified non-manual grammatical markers with increased accuracy.

Kuznetsova et al. (2021); Kimmelman et al. (2020) performed a quantitative study on NMMs (specifically eyebrow movement and head tilting) during questions in Kazakh-Russian Sign Language (KRSL). Previous projects used manual annotations, which are "extremely time-consuming [to create] and potentially unreliable". Another method for conducting a quantitative analysis, through the use of motion tracking equipment, is both very costly and does not produce naturalistic data due to the trackers that signers have to wear. Therefore, previous analyses of sign languages are mostly qualitative. Kuznetsova et al. (2021) propose that the application of CV techniques can "facilitate reliable quantitative analysis and enable quantitative cross-linguistic comparison in future". They collected a data set with video recordings of 9 native signers (5 deaf and 4 interpreters who were hearing children of deaf adults (CODAs)). Participants signed 10 sentences in 3 forms: as a statement, a polar question, and a wh-question. The final data set of 270 videos was manually annotated. OpenFace software was used to extract 3D landmark information about the average position of a signer's eyebrows from the 2D videos in the data set. They found that OpenFace returned slightly biased data when the signer's head was tilted. To account for this, they trained a linear regression model to learn the neutral position of the eyebrows. The resulting data was analysed using a mixed-effects multivariate linear regression model.

Later, Kuznetsova et al. (2022) improved on their previous research using the same data set and landmark detection technique through OpenFace as described above. This time however, they improved their technique for dealing with OpenFace's bias through a multilayer perceptron. Moreover, to analyse the movement of the non-manuals as continuous data rather than discrete, Kuznetsova et al. (2022) explored the application of functional Principal Component Analysis (fPCA), a statistical tool that is able to account for data that is made up of temporal sequences, and therefore often used in speech research. They found that the principal components were interpretable and easy to explore through visualisations.

The related work discussed in this Section provides some insights in the methodologies of previous work in which CV and/or ML techniques have been applied, and their limitations.

1.5 Proposed methodological refinements

As discussed in Section 1.4.3, sign language linguists have previously applied CV and ML techniques to analyse the use of NMMs in other sign languages. For NGT, however, this has not yet been attempted. This research project aims to explore the computational analysis of NMMs in NGT, while overcoming some of the methodological limitations of previous projects.

First, previous methods relied on landmark detection from 2D videos, which has shown to produce bias when the face is tilted or rotated. Although researchers have found methods to account for this bias, it is unknown whether these methods suffice. In this project, data will be gathered through the use of a 3D depth camera, which measures 61 facial features of interest, called *blendshapes*. Blendshape coefficients are expressed on a scale of 0 to 1, indicating the amplitude of engagement for each feature.[5] The blendshape data are more reliable and precise than coordinates obtained through landmark detection. Moreover, they are directly expressed in the measurement of interest as opposed to landmark coordinates, which first need to be translated into a measurement of engagement.

Second, the research conducted by Kuznetsova et al. (2021) and Kuznetsova et al. (2022) has a small data set of videos of both deaf and hearing signers. In the latter project, they found that there was a significant difference in how deaf and hearing signers of KRSL used NMMs. The data set for the present project is based solely on deaf signers of NGT, as this is the group of interest.

Third, the data sets used in previous work on NMMs in questions are minimally structured: lumping all polar questions together in one bag, and all wh-questions in another, while it is known that a lot of variability can exist within these classes (Klomp, 2021). As this project aims to learn more about polar questions in depth, the data set focuses only on this class, and distinguishes between different contexts in which polar questions may be asked, involving different expectations on the speaker's part as to which of the two possible answers is most likely to be true. We expect that controlling for this contextual factor may shed new light on the large amount of variation found in previous work on NMMs in polar questions.

Finally, in previous studies, data was mostly elicited through written prompts, or through videos of a person signing the target data. For instance, in Kuznetsova et al. (2021); Kimmelman et al. (2020), hearing signers were presented with the stimuli in written Russian, while the deaf signers were presented with the stimuli as videos in Kazakh-Russian Sign Language. These videos may have influenced the signs produced by participants. In contrast, our experiment is designed to minimally influence participants in the signs used to produce the target questions, and the order in which they are signed.

[5] e.g. a value of 1 for EyeWide indicates that the eyes are fully opened, whereas a value of 1 for EyeSquint indicates that the eyes are fully squinted.

1.6 Hypothesis

Based on the literature review, the hypothesis is that NMMs for polar question marking in NGT will include mainly 'raised inner/outer eyebrows' and 'head tilted forward'. Additionally, 'eyes wide' may be observed. Although literature also often defines 'body tilted forward' as a characteristic feature of question marking, this will not be measurable through the methods applied in the present research project. Although the literature mainly defines NMMs for polar questions to include 'raised eyebrows', we expect that in some cases the movement of the eyebrows may be characterized by a sequential or simultaneous combination of 'raised inner and outer eyebrows' and 'furrowed eyebrows'. In the case of 'furrowed eyebrows', we expect to see 'eye squint' to some degree as well. Further, due to the "phonetic strength" attributed to 'furrowed eyebrows' over 'raised eyebrows', we hypothesize that the former feature may be observed more often than the latter in situations where contextual evidence contradicts with speaker belief. Finally, we hypothesize that there might be other non-manual markers – that have not been previously regarded by the literature as such – that serve as characteristic components of prototypical facial expressions used in question marking in NGT.

We will proceed as follows. In Chapter 2, we discuss the design, collection, and preparation of the data set. In Chapter 3, we discuss which features will be considered in our ML application, and the reasons for which they are (or are not) selected. In Chapter 4, we focus on the methodological application of ML techniques. In Chapter 5, we visualise the resulting prototypical facial expressions, and map them to their respective contexts. Finally, in Chapter 6, we discuss the explored methodologies and the results found in the present research project. We also discuss limitations of this project and propose subsequent avenues for further research.

Chapter 2

Data set design, collection, and preparation

This Chapter discusses the design, collection, and preparation of the data set. First, Section 2.1 introduces the requirements for the data set, and how the experiment through which data was gathered was subsequently designed. As mentioned in the introduction, the question of how bias influences non-manual marking in polar questions in NGT, and the experiment through which this research question is investigated are both part of a larger project at SignLab Amsterdam (referred to as the BQ-NGT project). The experiment was designed for the purpose of the BQ-NGT project, and the experimental design was completed before the present research project was introduced as an addition. The setup, content, and selection of participants for the experiment through which data was gathered for the current project are therefore not contributions of this research project. There is currently no publication about the BQ-NGT project in its entirety, as it is still in progress. However, a preliminary report on the materials and procedure of this experiment is provided in Oomen and Roelofsen (2022a). Section 2.2 discusses the data collection, which took place at the beginning of this research project. During this stage, personal contributions include taking the role of second experimenter and assisting the lead experimenter of the BQ-NGT project in their data collection, as well as collecting the 3D data relevant to the present research project. Finally, Section 2.3 discusses how data was pre-processed to prepare it for the next steps of this project, and the structure of the resulting data.

2.1 Design

2.1.1 Overall data set design

In order to study whether bias accounts for variation in non-manual marking of polar questions in NGT, an experiment needed to be set up that specifically controlled for context. As discussed in Section 1.4, previous research on NMM has not controlled for context in this way, and has regarded all types of polar question as one category. As mentioned above, the data set and experimental design that are discussed in this and the following Section were completed before this particular research project was introduced to the BQ-NGT project. Unless stated otherwise, the information reported in these sections has been gathered through personal communication with the researchers of the BQ-NGT project.

No detailed information is currently available about the type of variations one could expect for non-manual marking of polar questions in NGT. In order to analyse more natural variations in facial expressions, the researchers decided to place the participant in different contexts through role-play type conversations with two confederates, who "signed pre-scripted utterances in response to participant productions prompted by stimulus materials projected on a laptop screen" (Oomen and Roelofsen, 2022a). This way, there would be full control in the introduction of different combinations of original speaker bias (introduced by confederate A) and contextual evidence (introduced by confederate B) (Oomen and Roelofsen, 2022a). The productions were elicited from all participants in the same way, using the same instructions and experimental stimuli. Moreover, responses elicited in a role-play setting would likely be more natural and spontaneous than simply asking signers how they would convey questions in different conditions. In that case, it would also be more difficult to control the thought process of the participants, as the hypothetical situations could be quite complex.

Another important design decision was that the role-play production task allowed for all communication to occur in NGT, either live (scripted or unscripted) or through pre-recorded videos. As discussed in Section 1.4, previous research mostly used

Contextual evidence	**Original speaker bias**		
	Negative	Neutral	Positive
Positive			
Neutral			
Negative			

Table 2.1: Experimental conditions included in the study of the Biased Polar Questions in NGT project. Adapted from Oomen and Roelofsen (2022a)

textual representations to elicit data. The setup for this experiment allowed for a more 'free' interpretation of what responses could look like. The only requirements given to participants regarding their productions were "that they always had to ask *questions* to confederates, but that there were no restrictions on sign order or use of facial expressions. They were instructed to keep their productions brief, preferably a single sentence, and to sign them as naturally as possible" (Oomen and Roelofsen, 2022a).

The experimenters constructed six different situations, or *scenarios*, "designed to elicit polar questions from participants with different kinds of bias. The situations were loosely based on selected scenarios from Domaneschi et al. (2017) study on bias in polar questions in spoken German and English" (Oomen and Roelofsen, 2022a). For every scenario, seven different conditions were tested that consisted of different combinations of original speaker bias and contextual evidence, which are shown in Table 2.1 (Oomen and Roelofsen, 2022a). The combinations in conditions Positive-Positive and Negative-Negative were not considered for this study, as in these cases it would not be logical for the participant to repeat the target question to confederate B (Oomen and Roelofsen, 2022a). For every scenario, the target question was played out and recorded for every condition, as well as baseline questions (without speaker bias), and declarative statements.

The requirements of the data set regarding the participants were defined with the following considerations. The signing community in the Netherlands is not very large, which makes recruiting participants difficult. Additionally, data annotation and analysis is very time-consuming, so not too many people could participate. Moreover, the approach of the BQ-NGT project is more qualitative than quantitative, so there was no need to recruit a large number of participants required to perform large-scale quantitative analysis. However, in order to get an impression of individual signer variation in NGT, the researchers aimed to recruit at least five participants for the study. There was no hard limit on the maximum number of participants. However, participants needed to be signers of NGT, who use NGT in their daily life (Oomen and Roelofsen, 2022a). Moreover, in order to simulate conversations in a way that would be as natural as possible, the two confederates with which participants interacted were both deaf, and "early acquirers of NGT" (Oomen and Roelofsen, 2022a).

Further requirements of the data set were based on a few restrictions, the most important being time. The experiment could not last too long, as the task would by nature be repetitive, complex, and intensive, and participants should not lose their focus during it. Every scenario would comprise seven trials (condition within a scenario), in addition to recording a matching declarative statement. Considering the above, the researchers limited the experiment to five scenarios, and one 'practice' scenario.

The scenarios and target questions had a few constraints. Scenarios had to be fictional because participants should not have a personal prior belief about them, ensuring that each condition could be properly tested. Scenarios had to be designed in such a way that the intended target question would be a natural – or at the very least possible – production in all experimental conditions. Further, the target questions of every scenario needed to be as short as possible, as the participants should be easily able to remember the target question, and the target question needed to be comprehensibly visualised in a picture prompt. This picture prompt was used as a trigger for the target question in favor of video recordings of signs for the following reasons: 1) a picture prompt would allow the participant to directly respond to confederate B by asking the target question without the interruption of a video, 2) a video "could influence participants in terms of e.g. lexical choices but also sign order" (Oomen and Roelofsen, 2022a). In order to prevent influencing the sign order in which participants would pose the target question, the "pictorial representations of concepts [were placed] on top of rather than next to each other ... [and as] this vertical alignment could still have an effect on constituent order ... two sets of picture prompts [were created], where the top and bottom images on the left of the prompts were reversed" (Oomen and Roelofsen, 2022a).

Moreover, the scenarios were constructed in such a way that there would likely be little to no occlusions of the face, and that there would not be a difference in authority between the two confederates (Oomen and Roelofsen, 2022a). Besides these considerations, the further content of the scenarios was

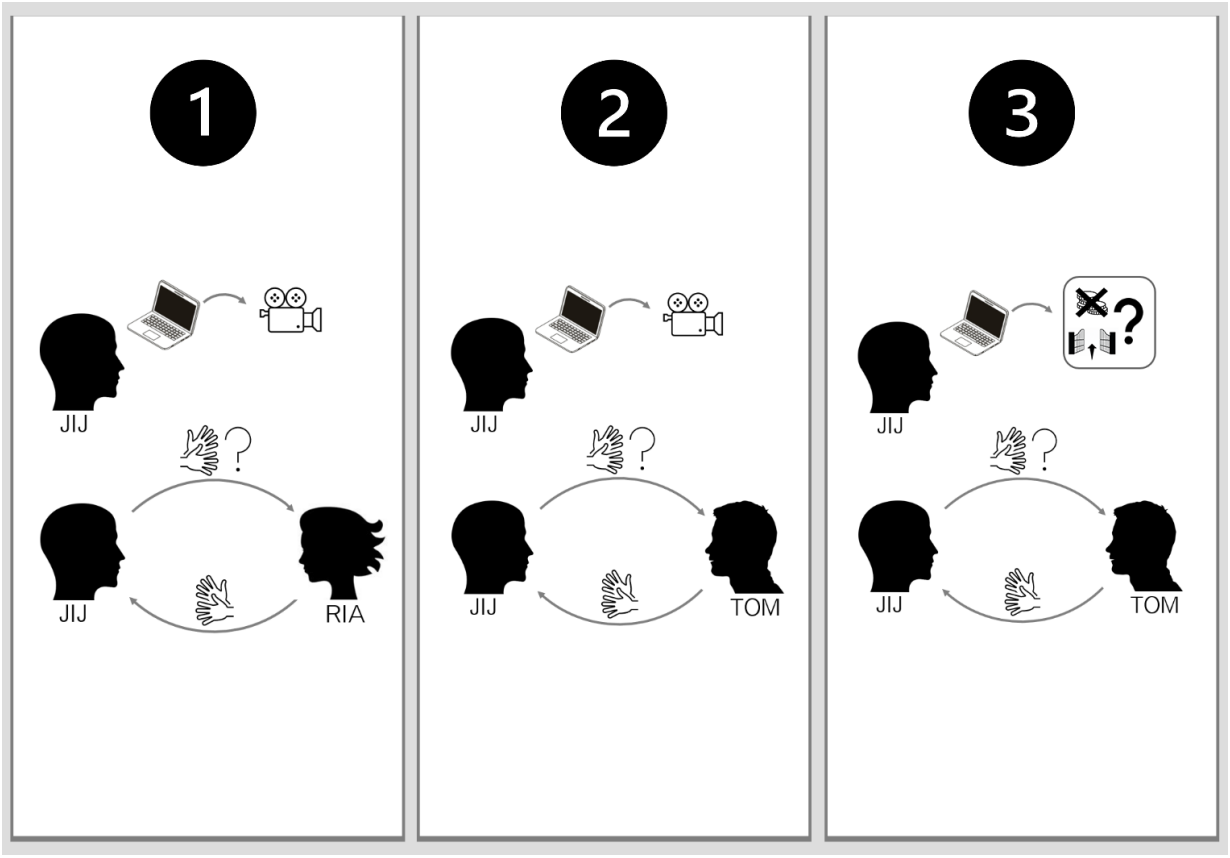


Figure 2.1: Instructions about the structure of a trial for the participants (jij is Dutch for 'you'). In the first stage they watch a context-providing video and interact with confederate A (Ria). In the second stage they watch a context-providing video and interact with confederate B (Tom). In the third stage, they repeat the target question to confederate B (Oomen and Roelofsen, 2022a)

irrelevant to the topic of investigation. The final scenarios, including context, confederate responses, and the picture prompts, are fully available in written English in Appendix A in (Oomen and Roelofsen, 2022a). With permission of the original authors, a duplicate of this Appendix can be found in Appendix A.1

2.1.2 Experimental design

This Section explains the structure of the experimental design by going over the different stages in one trial. Figure 2.1 shows the structure of a role-play conversation with the confederates, which was identical for all trials (Oomen and Roelofsen, 2022a). In the first stage, the participant watched a pre-recorded video that provided the context of the first interaction, as in (3).[1]

(3) "You want to visit a national park tomorrow. You do not know whether the entry to the park is free. Ria is a volunteer at the park. Ask Ria."

This prompts the first interaction, between the participant and confederate A, "who subsequently responds with a statement introducing either positive, negative, or neutral original speaker bias for the target question at the end of the third interaction. Content-wise, the target question is always either identical or almost the same as the question the participant is instructed to ask in this first exchange" (Oomen and Roelofsen, 2022a). As instructed by the context-providing video, the target question in this example concerns whether or not the entry to the national park is free. In this first interaction, there is "crucially, [not] any bias yet involved" (Oomen and Roelofsen, 2022a). For each scenario, one of the recordings of this unbiased target question is taken as the baseline. The first interaction is shown in (4).

(4) **Par:** *Is the entry to the park free?*
Ria: *No, a ticket costs €10.*

The original speaker bias has now been introduced, concluding stage one. The participant believes that the entry to the park is not free, and costs €10. The second stage begins with another pre-recorded video, explaining the new context for the upcoming interaction with confederate B, as in (5).

[1]The context-providing videos are available in NGT at Oomen and Roelofsen (2022b).

	Blendshapes				
Upper Region	EyeLookDownLeft	EyeLookDownRight	EyeLookUpLeft	EyeLookUpRight	EyeBlinkLeft
	EyeLookInLeft	EyeLookOutRight	EyeLookOutLeft	EyeLookInRight	EyeBlinkRight
	EyeWideLeft	EyeWideRight	EyeSquintLeft	EyeSquintRight	BrowInnerUp
	BrowDownLeft	BrowDownRight	BrowOuterUpLeft	BrowOuterUpRight	
Lower region	MouthLowerDownLeft	MouthLowerDownRight	MouthUpperUpLeft	MouthUpperUpRight	CheekSquintLeft
	MouthSmileLeft	MouthSmileRight	MouthFrownLeft	MouthFrownRight	CheekSquintRight
	MouthPressLeft	MouthPressRight	MouthShrugLower	MouthShrugUpper	NoseSneerLeft
	MouthRollLower	MouthRollUpper	MouthDimpleLeft	MouthDimpleRight	NoseSneerRight
	MouthFunnel	MouthPucker	MouthStretchLeft	MouthStretchRight	CheekPuff
	MouthLeft	MouthRight	MouthClose	TongueOut	
	JawLeft	JawRight	JawForward	JawOpen	
Rotation	LeftEyeRoll	RightEyeRoll	LeftEyeYaw	RightEyeYaw	HeadPitch
	RightEyePitch	LeftEyePitch	HeadRoll	HeadYaw	

Table 2.2: Blendshapes captured by Live Link Face

(5) "The next day you arrive at the parking lot of the national park. You cannot find the entrance to the park. In the parking lot, you run into Tom, another visitor; ask him."

The second interaction, between the participant and confederate B, presents the participant with "positive, negative, or neutral contextual evidence for the target question" (Oomen and Roelofsen, 2022a). Although the question in this interaction is not directly related to the target question, it "provide[s] a hook for the confederate to provide comment on the status of [the entry to the park being free] or not", as shown in (6) (Oomen and Roelofsen, 2022a).

(6) **Par:** *Where is the entrance to the park?*
Tom: *The entrance is over there by the white flag, you do not need a ticket.*

In the third stage, the participant is prompted to repeat the target question, '*Is the entry to the park free?*', to confederate B. As this should be a direct response to the interaction in (6), the participant is shown a picture prompt rather than a video (Oomen and Roelofsen, 2022a). However, the (positive) contextual evidence in the second stage contradicts the (negative) original speaker bias established in the first stage, which is expected to be reflected in the type of NMM used. (Oomen and Roelofsen, 2022a). For instance, they might furrow their brows ('bf') to indicate confusion, resulting in the gloss in (7).

(7) $\overline{\text{ENTRY PARK FREE}}^{\text{bf}}$

2.1.3 3D data set

The addition of the present research project to the experiment from the BQ-NGT project meant that an extra data set needed to be considered. As discussed in Section 1.4.3, previous researchers made use of methods to extract 3D landmark information from 2D videos. Technologies such as OpenFace, are very complex and have been designed specifically for the purpose of facial behavior analysis (Baltrusaitis et al., 2018). However, previous research has show that this technique can produce bias (Kuznetsova et al., 2021, 2022). One study found that "estimating AUs [using the FACS framework] precisely is a computationally expensive and difficult task to do with off-the-shelf software such as OpenFace toolkit 2.0" (Miyawaki et al., 2022). Further, other researchers found in their tests that "using a recording based on Apple's *TrueDepth* Sensor turned out to work better than capturing the face using purely RGB-video-based solutions like OpenFace 2.0" (Ehret et al., 2021). Phones that are equipped with this sensor can download the free application, *Live Link Face*, to measure 61 standardised ARKit blendshapes on the face (Epic Games, 2022b; Apple, 2022). Table 2.2 lists the blendshapes tracked by Live Link Face. They are measured with values ranging between 0 to 1; with the exception of the blendshapes that measure rotation of the head and eyes, which range between -1 and 1.

Blendshapes are typically used in animation, and have been successfully applied in previous research on animating speech on avatars (Miyawaki et al., 2022; Ehret et al., 2021; Chen et al., 2022), and for animating an Augmented Reality avatar sign language interpreter (Luo et al., 2022). The data recorded by Live Link Face can used to render animations of facial expressions on extremely realistic avatars (called MetaHumans) in Unreal Engine, which was an additional driver for selecting these technologies for recording the data (Epic Games, 2022a). SignLab Amsterdam aims to eventually create a humanoid avatar that can realistically convey information in NGT, and has found through user evaluations that facial expressions and non-manual markers play a vital factor in this (Esselink et al., 2022b; Roelofsen et al., 2021a,b). Moreover, previous projects of collaborations between SignLab Amsterdam and the Visualisation Lab at the University of Amsterdam had recently explored the combination of these technologies for application in the field of sign language research, and had specialized compatible hardware available.

In order to improve the accuracy of blendshape measurements even further, the Live Link Face app

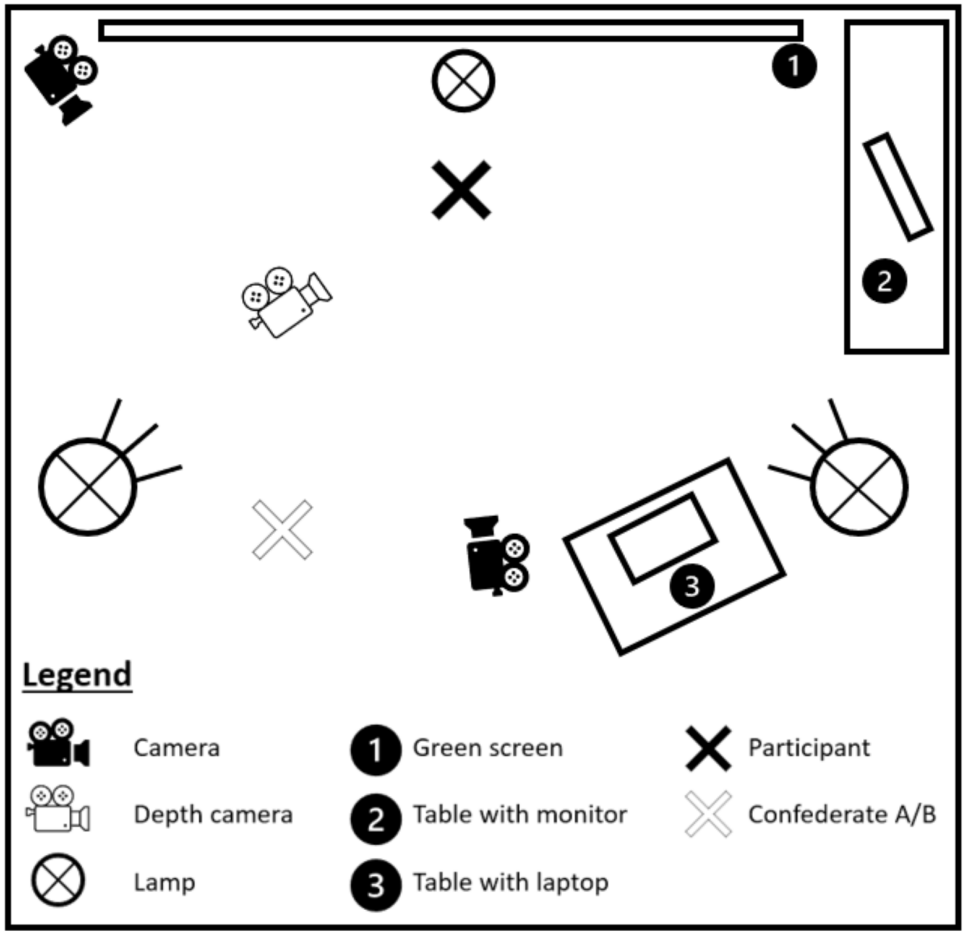


Figure 2.2: Overhead view of the University of Amsterdam's sign language recording studio, where the experiment took place (Oomen and Roelofsen, 2022a)

has the ability to calibrate its measurements against the neutral facial expression of the user, and automatically incorporates this in the resulting data. If for instance, a participant's eyebrows naturally seem 'raised' in the neutral position, they will not be recorded as such. Finally, the output of the Live Link Face app is not only useful for rendering avatar animations: it comprises video recordings as well as a CSV file with numeric measurements of all the blendshapes for every frame of the video. These CSV files are formatted comprehensibly, and require minimal pre-processing to be able to use them as an input for machine learning algorithms.

Taking the above into consideration, and for the purpose of exploring new methodologies for application in the field of sign language linguistics, we decided to make use of a 3D depth camera and the Live Link Face app to record facial features directly instead of extracting them from 2D videos after the experiments had concluded.

2.2 Collection

The experimental sessions took place over the course of eight days in June and July of 2022. Each session lasted between 3-4 hours (including breaks), and all communication took place exclusively in NGT. As mentioned at the beginning of this Chapter, personal contributions during data collection include taking the role of second experimenter and assisting the lead experimenter of the BQ-NGT project in their data collection, as well as collecting the 3D data relevant to this research project.

Figure 2.2 shows an overhead view of the University of Amsterdam's sign language recording studio, where the experiment took place (Oomen and Roelofsen, 2022a). The participant stood in front of a green screen (black 'X' mark), facing the interacting confederate (white 'X' mark). The positions at which the participant and confederate stood were marked on the ground with tape. Three video cameras were used to record the experiment: the first aimed at the participants; the second aimed at the confederates; and the third (3D depth camera) aimed at the participants. To properly collect the

3D data, the depth camera ideally needed to be positioned between the participant and the interacting confederate. It needed to be positioned so that the participant was facing the camera, but in such a way that it would not provide a distraction, and that the participant's view of the confederate was not obstructed. Live Link Face makes use of the front-facing camera of an iPhone, as the back-facing camera is not equipped with Apple's TrueDepth Sensor. The app was configured so that the screen was blacked out while recording, so that participants would not be distracted by seeing themselves on the phone screen.

The lead experimenter stood behind a high table (3) in Figure 2.2, and controlled a keyboard and mouse which were collected to a laptop facing the participant and to a monitor facing the confederates and experimenters (Oomen and Roelofsen, 2022a). The lead experimenter "guided each session and was responsible for providing instructions, projecting the stimuli ... on the laptop and monitor, as well as handling the participant camera" (Oomen and Roelofsen, 2022a). The second experimenter stood by the wall between the second camera and depth camera. The responsibilities of the second experimenter were "handling the confederate camera and depth camera ... [and indicating] the trial number on a poster on the wall as well as on a clapperboard, which was displayed in front of the cameras at the start of each trial" (Oomen and Roelofsen, 2022a).

Not all trials were recorded in one take. The participants first practiced each trial once or twice while viewing the context-providing videos, after which they were not repeated at the start of the first and second stage for every condition in a scenario. In some cases, participants did not clearly pose the target question as a question, but rather as a statement. In these situations the participants and confederates acted out the role-play again, starting at the first phase.

The participants consisted of eight deaf NGT signers. Two participants are excluded from the data of both projects. One of these participated in a test run of the experiment. The second participant "had indicated in the participant survey that he began learning Sign Supported Dutch from the age of four (primary school) and only started learning NGT in high school ... the influence from Sign Supported Dutch was obvious in his signing, and his productions clearly diverged from those by the other signers" (Oomen and Roelofsen, 2022a). Moreover, two participants wore glasses, which obstructed the camera view. In order to get the most accurate data, these participants are excluded solely from the data of this research project. The remaining four participants are all women, and use NGT on a daily basis (Oomen and Roelofsen, 2022a).

The seven experimental conditions over five scenarios resulted in 35 target questions per participant. Together with the additional recordings of the baseline questions and declarative statements, the data set comprised 45 recordings for each participant, or 180 recordings in total.

The complete data set for the six participants of the larger project at SignLab is available at Oomen and Roelofsen (2022c). The complete 3D data set for the four participants of the present research project is available at Esselink et al. (2022a).

2.3 Pre-processing

2.3.1 Selecting relevant data

Aligning camera recordings Not all of the recorded data is of interest. The recordings of the confederate camera are not used for the analysis in their entirety, and the recordings of the main camera are trimmed to only include the baseline, the target question, and the declarative statement for each condition. The lead experimenter of the BQ-NGT project therefore first selected the relevant videos for each condition (in the case that multiple takes were recorded), and created video compilations of every scenario for each condition, per participant.[2]

Initially, one goal was to be able to validate the 3D data through manual annotations.[3] However, the recordings of the main camera and the depth camera were not perfectly aligned, which is necessary for a frame-by-frame analysis. It was not feasible to automatically align them during post processing, as the internal timestamps of the cameras do not align, and the cameras started recording at different times. During the experiment, the cameras captured a different number of frames per second (fps): the main camera captured 50 fps and the depth camera 60 fps.[4] Therefore, the videos of the main camera were recompiled at a rate of 60 fps. The lead experimenter of the BQ-NGT project subsequently used the 'new' recordings of 60 fps for their manual annotations, so that these can later be compared to the measurements of the depth camera.

During the experiment, a clapperboard was used so that videos could be aligned on the basis of the sound waves. However, we found that aligning the videos on the basis of sound does not align the *visual* video recordings, which is the ultimate goal. The reason for this is not entirely clear, but it could be due to converting the framerate of the videos cap-

[2] A compilation can thus include the take recordings of the NegPos condition for scenarios 1-5, for participant 3.

[3] Although this was originally a goal for this research project, the manual annotations are not completed yet. However, this validation will take place after this research project concludes.

[4] The main camera was set to record at a setting which captured movement at the best possible setting, and did not allow recording at 60 fps as the depth camera did. The depth camera only had the options to record at 30 or 60 fps.

tured by the main camera, or perhaps a hardware issue with one of these cameras. Therefore, each video was aligned by selecting a uniquely identifying frame (for instance, the exact frame in which a participants fingers touched together). When the videos were exactly aligned, the videos recorded by the depth camera were cut to the exact beginning and end frame that the lead experimenter of the BQ-NGT project had selected. Subsequently, the CSV file of each video were trimmed to only include the frames that had been selected, so that only the relevant blendshape measurements will be taken into consideration.

3D data set The next step is to compile the CSV data for the selected videos into one data set. Each trimmed CSV file is loaded into RStudio, and every frame in that file is marked with a video ID, participant ID, scenario ID, condition ID, and a frame number ID (a unique ID for every frame in the video). In addition to measuring blendshapes, Live Link Face also records a time stamp and a count of how many blendshapes are recorded for every frame (61), which are discarded. Next, blendshape measurements are multiplied by a factor of 100 for readability. The processed files are then combined to form a data set with 66 columns for every frame – the IDs listed above and the 61 blendshape measurements. The total data set comprises 58.393 samples.

In some cases, some of the feature values were not recorded due to occlusions of the face. This was often only for a relatively small number of frames (less than a third of a second), so the missing data was interpolated from the surrounding frames that had captured these measurements.

Although the videos and CSV data of the depth camera have been aligned to the videos of the main camera in the previous stage, further trimming is necessary. Most videos are trimmed in such a way that they do not necessarily begin at the exact time a participant started signing the target question, but rather a second or two *before*. Likewise, they often end a few seconds *after* the participant had finished signing the target question. This means that the data still contains noise, as we are not interested in facial expressions before and after the target questions. Each video is therefore again examined for the true beginning and end of the target questions. The frame numbers at which this occurs are recorded for every video. Supplemental columns are added to the data set in RStudio: one indicating the frame ID at which the question begins, one indicating the frame ID at which the question is finished, one containing the frame's status for whether or not it will be used for machine learning (labeled as either '1': keep, or as '0': discard), and one indicating the total amount of frames that are marked as 'keep' for that video.

Modeling temporal development Previous research has modeled temporal development of NMM by normalising each video in such a way that the noun and verb began at identical frame numbers for every video in their data set (Kuznetsova et al., 2022). This was a feasible approach as the sentences in their data set were highly structured, containing the same three signs in the same order for every trial for every participant. However, it is not feasible for the current data set, due to the freedom that participants were (intentionally) given in the signs they used, and the order in which they signed them. This freedom resulted in a less structured data set, where participants sometimes added signs (for instance, '*The entry to the park is free, right?*') or switched around the sign order.

As the current data set cannot be split on constituents, a different method for modeling temporal development is employed. For each video, the frames are assigned to one of five *windows*, which can be used to place each frame in a point in time of the question. For instance, a frame belonging to the first window is recorded near the beginning of a question, while a frame belonging to the second window is recorded between the start and the midpoint of a question, and so forth. As every video has a different total amount of frames, the sizes of the windows also vary between videos. However, window size does not vary within videos: they are equally divided in 5 sections. Note that the allocation of frames into windows is only recorded for those frames that have been marked as taking place within the actual duration of the question, and not for those where the participant had not yet begun signing or has finished signing. This new information is again added to the data set in RStudio.

Final data set The combined data set in RStudio now comprises 71 columns and 58.393 samples. A new data set is created that only contains those samples that have been assigned the 'keep' status, and a check is carried out whether the correct frames have been discarded and whether the total amount of remaining frames is correct for each video. Next, the frame number IDs are reset to start the count from 1. Irrelevant columns – the start and end frame IDs of the question, the keep/discard status, and the total number of frames for each video – are discarded. The new trimmed data set comprises 67 columns and 30.590 samples.[5]

Both data sets are useful for different applications. The original untrimmed data set can be used to validate and compare the blendshape data against the manual annotations of the lead researcher of the BQ-NGT project. These manual annotations are currently still in in process of being finalised. On

[5]A complete overview of the recorded blendshapes and examples of what they look like can be found at https://arkit-face-blendshapes.com.

f_1	f_2	Correlation	$f_{combined}$
EyeBlinkLeft	EyeBlinkRight	0.9990	EyeBlink
EyeLookDownLeft	EyeLookDownRight	0.9999	EyeLookDown
EyeLookInLeft	EyeLookOutRight	0.9716	EyeLookRight
EyeLookOutLeft	EyeLookInRight	0.9711	EyeLookLeft
EyeLookUpLeft	EyeLookUpRight	0.9999	EyeLookUp
EyeSquintLeft	EyeSquintRight	0.9866	EyeSquint
EyeWideLeft	EyeWideRight	0.9999	EyeWide
JawLeft	JawRight	0.2508	–
MouthLeft	MouthRight	0.2322	–
MouthSmileLeft	MouthSmileRight	0.9950	MouthSmile
MouthFrownLeft	MouthFrownRight	0.9933	MouthFrown
MouthDimpleLeft	MouthDimpleRight	0.9860	MouthDimple
MouthRollLower	MouthRollUpper	0.6736	–
MouthStretchLeft	MouthStretchRight	0.9640	MouthStretch
MouthShrugLower	MouthShrugUpper	0.9596	MouthShrug
MouthPressLeft	MouthPressRight	0.9940	MouthPress
MouthLowerDownLeft	MouthLowerDownRight	0.9954	MouthLowerDown
MouthUpperUpLeft	MouthUpperUpRight	0.9951	MouthUpperUp
BrowDownLeft	BrowDownRight	1.0000	BrowDown
BrowInnerUp	BrowOuterUpLeft	0.9436	–
BrowInnerUp	BrowOuterUpRight	0.9445	–
BrowOuterUpLeft	BrowOuterUpRight	0.9992	BrowOuterUp
CheekSquintLeft	CheekSquintRight	0.9805	CheekSquint
NoseSneerLeft	NoseSneerRight	0.9642	NoseSneer
LeftEyeYaw	RightEyeYaw	0.9998	LeftEyeYaw
LeftEyePitch	RightEyePitch	1.0000	LeftEyePitch
LeftEyeRoll	RightEyeRoll	0.9562	LeftEyeRoll

Table 2.3: Correlation of features that have both a left and right (or similar) measurement, and their combined feature name

the other hand, the trimmed dataset no longer contains the noise of irrelevant frames at the beginning and end of every video, includes additional information that will help model temporal development, and can be used for machine learning.

2.3.2 Transforming the data

Correlation Many features that share both a left and right (or similar) measurement are highly correlated, as shown in Table 2.3. These measurements are recorded individually for the purpose of creating more realistic and natural expressions during animation (Apple, 2022). However, the minimal differences between these measurements are not relevant to this research project, as we are interested in the complete facial expressions in this data set, and not in using this data set to create hyper-realistic animations. Therefore, we simplify the data set by merging the features with a correlation above 0.95 together, listing the mean of these features as the measured value for every sample. Note that this means that despite the high correlation, the feature BrowInnerUp stays separate from both BrowOuterUp features. This is intentional, as the manual annotations will differentiate between BrowInnerUp and BrowOuterUp, and BrowInnerUp has a higher average measurement than BrowOuterUp. Further, MouthRollLower, MouthRollUpper, JawLeft, JawRight, MouthLeft, and MouthRight are not highly correlated, as they do not measure typically symmetrical movements. Features with single measurements are: JawForward, JawOpen, MouthClose, MouthFunnel, MouthPucker, CheekPuff, TongueOut, HeadYaw, HeadPitch, HeadRoll

Normalisation On closer inspection of the data, we find that the range of values recorded for each participant differs. An example of this phenomenon is illustrated for two features, f_1 and f_2 for two participants $P1$ and $P2$ in Table 2.4a. For f_1, we see that measurements of $P1$ range between 0 and 80, while the measurements of $P2$ for this feature range between 20 and 60. Likewise, for f_2, we see that measurements of $P1$ range between 10 and 70, while those for $P2$ range between 30 and 90. In practice,

	f_1		f_2	
	P1	*P2*	*P1*	*P2*
Min	0	20	10	30
Max	80	60	70	90
Val_1	28	28	25	33
Val_2	40	44	34	54
Val_3	72	56	52	81

(a) Original

	f_1		f_2	
	P1	*P2*	*P1*	*P2*
Min'	0		0	
Max'	100		100	
Val_1'	35	20	25	5
Val_2'	50	60	40	40
Val_3'	90	90	70	85

(b) Normalised

	f_1		f_2	
	P1	*P2*	*P1*	*P2*
Min''	10		20	
Max''	70		80	
Val_1''	31	22	35	23
Val_2''	40	46	44	44
Val_3''	64	64	62	71

(c) Range

Table 2.4: Comparison of blendshape values after transformation to normalised and ranged values

this means that if both $P1$ and $P2$ express f_1 to their maximum extent, the recorded blendshape value for $P1$ will be 20 points lower than that of $P2$. Since we are not interested in whether one participant is generally able to (for instance) raise their eyebrows higher than another participant can, we want to normalise these values. This is done according to equation 2.1, which expresses a value as a percentage of it's old range.

$$x' = \frac{x - min}{max - min} \times 100 \tag{2.1}$$

Table 2.4b shows the outcome for three example values of these features. For Val_1 in f_1, $x = 28$ for both $P1$ and $P2$. However, because their min and max are different, the resulting equations – and thus the resulting normalised values – are different too, as seen in equations 2.2 for $P1$ and 2.3 for $P2$.

$$x' = \frac{28 - 0}{80 - 0} \times 100 = 35 \tag{2.2}$$

$$x' = \frac{28 - 20}{60 - 20} \times 100 = 20 \tag{2.3}$$

Likewise, we see that Val_2 is different for these participants for f_2: $x = 34$ for $P1$, and $x = 54$ for $P2$. However, due to the participants' min and max values for this feature, the normalised value is the same for both participants, as seen in equations 2.4 for $P1$ and 2.5 for $P2$.

$$x' = \frac{34 - 10}{70 - 10} \times 100 = 40 \tag{2.4}$$

$$x' = \frac{54 - 30}{90 - 30} \times 100 = 40 \tag{2.5}$$

This normalisation is calculated for each participant individually. Note that we apply this transformation to all the samples in the data set that originate from that participant at once, thus taking their *overall* minimum and maximum values. Normalising over each video individually is not representative, as the true minimal and maximal values for a feature are not necessarily reflected in each video. The normalised data set is now a fairer representation for our intended purpose, as each participant now has the same range of values for each blendshape.

Range While a normalised data set is a fair representation for the purpose of machine learning, it is not a fair representation of the *actual* amplitudes with which features are expressed. Take, for instance, a feature of which the original maximum value (over all participants) is 40. It would then be represented as a value of 100 in the normalised dataset. This would be good for machine learning, as we want to indicate that the feature is expressed to the full extent of it's maximum amplitude. However, it would (naturally) not be fair to subsequently conclude that a facial expression contains that feature at an amplitude of 100, as this does not occur in our data set.

Therefore, we want to express the data in an additional manner: as a range of mean values. We use the formula in equation 2.6.

$$x'' = \left(\frac{x'}{100} \times (max'' - min'')\right) + min'' \tag{2.6}$$

To calculate the new min and max value for a feature, we take the mean min and max values of all participants for that feature. Turning back to one of our example features, the calculations for min'' and max'' for f_2 of Table 2.4a are shown in equations 2.7 and 2.8

$$min'' = \frac{10 + 30}{2} = 20 \tag{2.7}$$

$$max'' = \frac{70 + 90}{2} = 80 \tag{2.8}$$

Now that the new minimum and maximum values are known, we can insert them into equation 2.6 together with the normalised values to calculate the ranged values. For Val_3', the process is shown in equations 2.9 for $P1$, and 2.10 for $P2$.

$$x'' = \left(\frac{70}{100} \times (80 - 20)\right) + 20 = 62 \tag{2.9}$$

$$x'' = \left(\frac{85}{100} \times (80 - 20)\right) + 20 = 71 \tag{2.10}$$

Further examples shown in Table 2.4 will not be discussed here, but illustrate extra conditions.

Chapter 3

Selecting features

Following the steps taken in Section 2.3, we are left with 39 features. This Chapter will discuss the process of feature selection, and argue for each feature why it is or is not selected for the machine learning application. Such a selection is necessary, as not all recorded features are relevant to question marking, and some features introduce noise to the data set. We will begin with discussing the features that are known to be linguistically relevant through the literature in Section 3.1. The remaining features are then discussed in Section 3.2.

For some features, the decision of whether or not they are selected is illustrated through plotting these features over the course of one recording. In this recording, the subject is stood as still as possible, keeping a neutral expression while mouthing the words of the target questions of the experiment. It therefore shows the effect that mouthing has on value measurements of the recorded features. In this video, the mouthings of a new scenario start roughly every 200 frames. We will refer to this video as M.

Moreover, some decisions will be motivated through visualisations of the mean value of that feature for each condition of the experiment. Specifically, these figures highlight the difference between the mean feature values during the declarative statement recorded for each scenario, and the experimental conditions.

3.1 Features from the literature

This section discusses the features that belong to definitions of 'q' as described in the literature review in Section 1.4. We will first discuss the features that are necessarily included in 'q', after which we discuss the features that are sometimes omitted from 'q' in the literature.

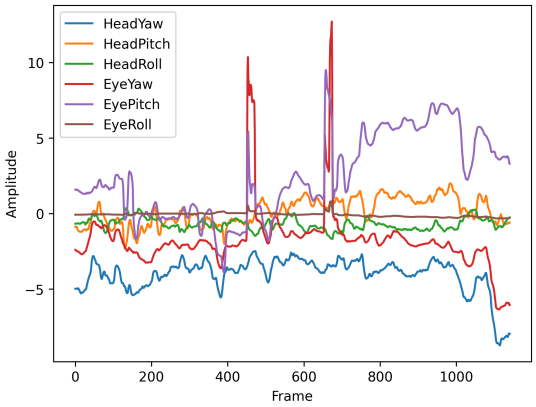
3.1.1 Eyebrows

Perhaps the most prominent markers for questions are the eyebrows, which are always included in definitions of 'q'. We have seen in Section 1.4 that 'raised eyebrows' is taken to be the most common non-manual marker for questions (Pfau and Quer, 2010; Cecchetto, 2012; Zeshan, 2004; Coerts, 1992). Additionally, 'furrowed eyebrows' may be used to mark questions (de Vos et al., 2009). The former has been measured with the features BrowInnerUp and BrowOuterUp, and the latter with BrowDown. Naturally, these features are selected for our task.

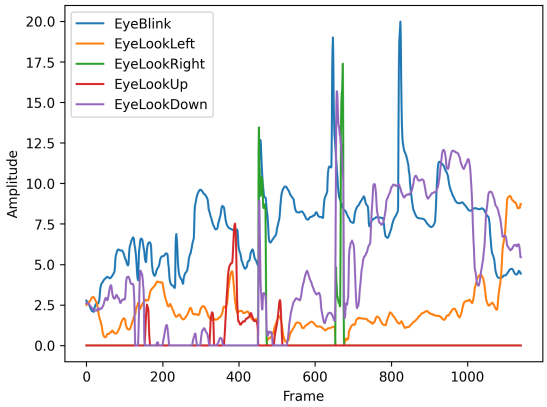
3.1.2 Head and body tilt

Like the eyebrows, we have seen that 'head tilted forward' and 'body tilted forward' are almost always included in definitions of 'q' (Cecchetto, 2012; Coerts, 1992). The latter concerns the position of the body, which is not measured by Live Link Face. This feature can therefore not be analysed using this specific method. Although 'head tilted forward' is measured by HeadPitch, we cannot automatically include this feature in our final selection.

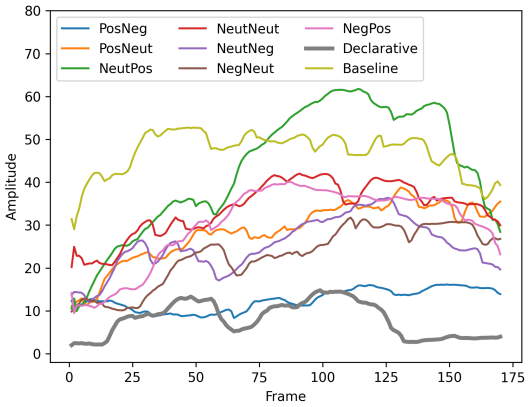
In Figure 3.1a, we see that the features measuring rotation are sensitive to noise, as they are affected by a participants relative position to the camera. During calibration, the participants explicitly faced the camera instead of the confederate, which is distinctly not the case during the rest of the experiment. Moreover, participants often moved around before, during, and after each trial. Despite the tape marking their position on the floor, it was often not the case that participants stood at that exact location. Especially after breaks a participant's location often differed from previous trials, and calibration was only executed once for each participant, at the beginning of the experiment. Although these factors did not seem to have a significant impact during the experiment itself, they caused too much variance in features measuring rotation. Therefore, the features not selected for our machine learning task were: HeadYaw, HeadPitch, HeadRoll, EyeYaw, EyePitch, and EyeRoll.



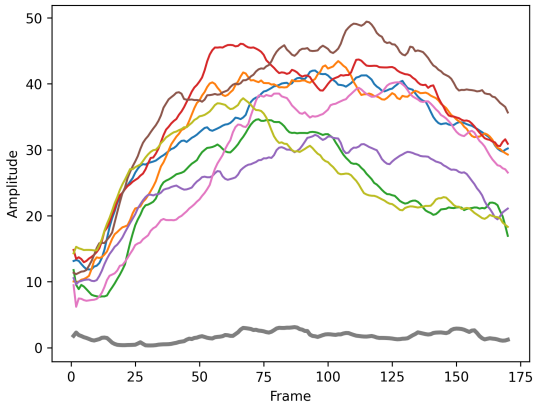
(a) Rotation



(b) Eye movement



(c) EyeWide



(d) EyeSquint

Figure 3.1: Feature measurements in M for features measuring rotation (a) and eye movement (b); mean movements per condition for EyeWide (c) and EyeSquint (d)

3.1.3 Eye gaze

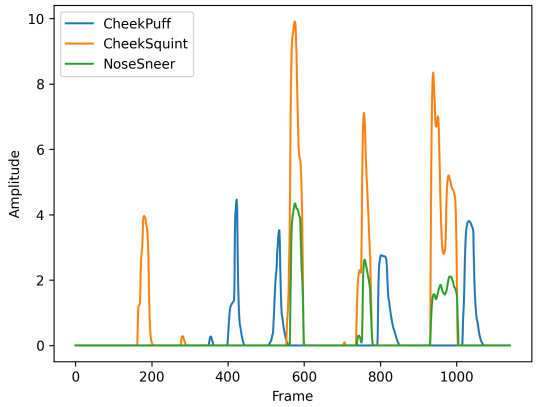
Although sometimes omitted, 'continuous eye contact with the addressee' may be included in some definitions of 'q' (Zeshan, 2004; Pfau and Quer, 2010). This non-manual marker can be extracted from EyeLookLeft, EyeLookRight, EyeLookUp, and EyeLookDown. Figures 3.1b shows the engagement of the remaining features measuring eye gaze (and EyeBlink, which will be discussed below). Again, we see that these features are sensitive to noise, and do not clearly show the subject looking in one direction. Further, as participants and confederates slightly shifted location surrounding the trials, each video would have to be manually inspected to determine the eye gaze, which is out of the scope for this research project. Therefore, these features are not selected.

3.1.4 Eye wide

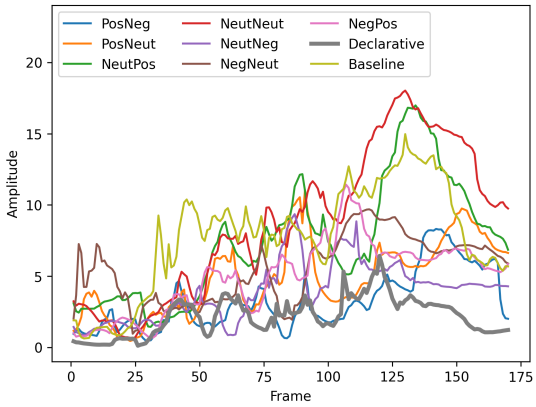
The final feature included in definitions of 'q' in the literature is the marker 'eye wide'. Like 'continuous eye contact with the addressee', this marker is not included in every definition of 'q' (Cecchetto, 2012; Coerts, 1992). In fact, it is not listed as part of *any* definition of 'q' for NGT (Coerts, 1992; Cecchetto, 2012; Klomp, 2021; Pfau and Quer, 2010; de Vos et al., 2009). Nevertheless, we select the feature EyeWide for two reasons: 1) it is listed as a defining non-manual marker for other sign languages, and 2) participants frequently had their eyes open wide while asking questions during the experiment. This is visualised in Figure 3.1c, where we see that the mean measurement for this feature is very low during the Declarative condition, while it is a lot higher for most other experimental conditions.

3.2 Remaining features

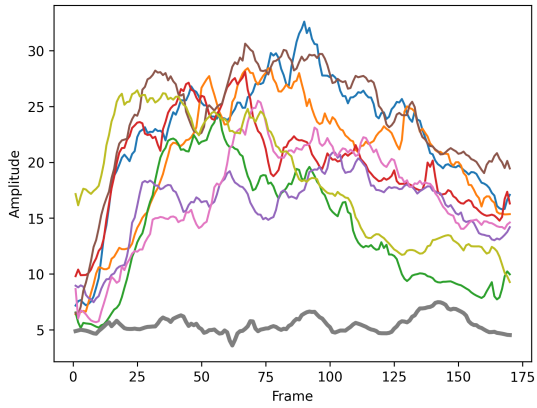
This section discusses the remaining 25 features that do not appear in the literature. Most of these features regard the lower facial region – specifically the jaw and mouth.



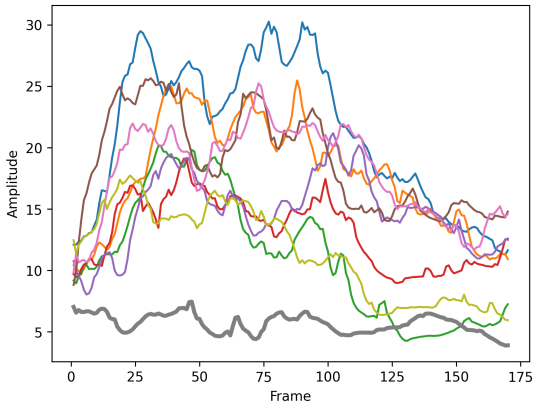
(a) Cheek and nose



(b) CheekPuff



(c) CheekSquint



(d) NoseSneer

Figure 3.2: Cheek and nose feature measurements in M (a); mean movements per condition for CheekPuff (b); CheekSquint (c); and NoseSneer (d)

3.2.1 Eyes

Two features measuring the eyes remain: EyeBlink and EyeSquint. Neither of these appears in any definition of 'q', although they are known NMMs outside of question marking (Pfau and Quer, 2010).

EyeBlink Figure 3.1b shows the measurement of EyeBlink over the duration of the example video. We see that it produces significant noise, as it is continuously active. This feature is therefore not selected.

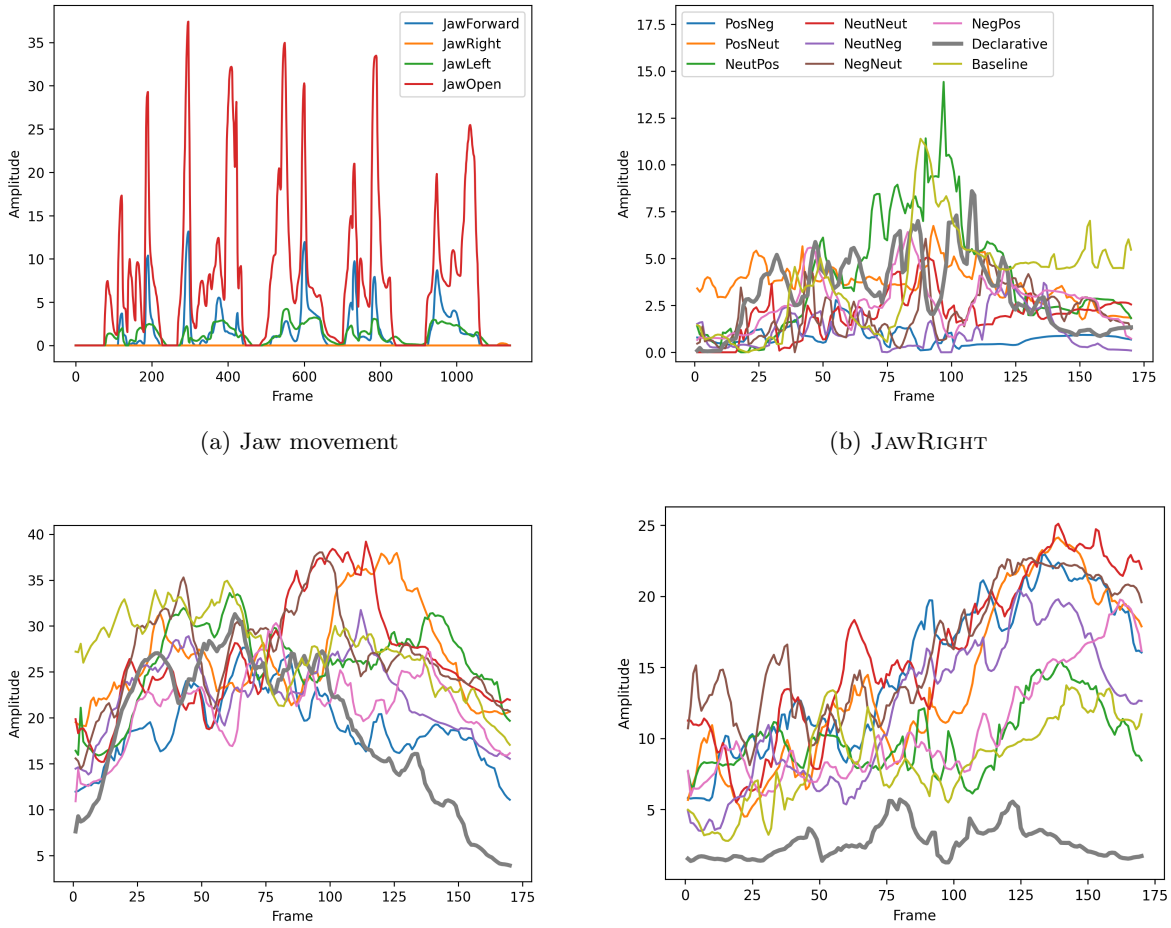
EyeSquint Figure 3.1d visualises the mean measurement of EyeSquint between conditions (the legend is identical to that in Figure 3.1c). Although it is not a component of previous definitions of 'q' from the literature, we hypothesize that EyeSquint could be a relevant NMM. During the experiment, this feature frequently occurred simultaneously with BrowDown. The contrast in engagement of this feature during the experimental conditions and the Declarative condition is clearly visible in the visualisation. For this reason, EyeSquint is selected for our machine learning data set.

3.2.2 Cheek and nose

The movements of the features measuring the cheek and nose, CheekPuff, CheekSquint, and NoseSneer, are visualised in Figure 3.2a.

CheekPuff We see that CheekPuff is minimally affected by the mouthings in M. However, Figure 3.2b shows that the mean measurements of this feature are only slightly lower in the Declarative condition than in the experimental conditions. Moreover, the maximum mean amplitude of this feature is not very high at 17. Therefore, CheekPuff is not selected.

CheekSquint Compared to the other features, CheekSquint is affected most by mouthings in M. However, the peaks (and noise) are minimal, at a maximum amplitude of 10. In Figure 3.2c we see that there is a definite contrast between the mean measurements of this feature during the Declarative condition and the experimental conditions. Moreover, the maximum mean amplitude is a lot higher



(a) Jaw movement

(b) JawRight

(c) JawForward

(d) MouthShrug

Figure 3.3: Feature measurements in M for jaw movement (a); mean movements per condition for JawRight (b); JawForward (c); and MouthShrug (d)

than CheekPuff. Further, this feature often co-occurs with EyeSquint and BrowDown (correlation is 0.72 and 0.73, respectively). Therefore, we hypothesize that this feature is a potential NMM of polar questions in NGT, and select it for our machine learning task.

NoseSneer Like CheekPuff, NoseSneer is minimally affected by the mouthings in M. However, like CheekSquint, we see a clear distinction between the mean measurements during the Declarative condition and the experimental conditions in Figure 3.2d, meaning that this feature too could be a potential marker of polar questions in NGT, and is also selected.

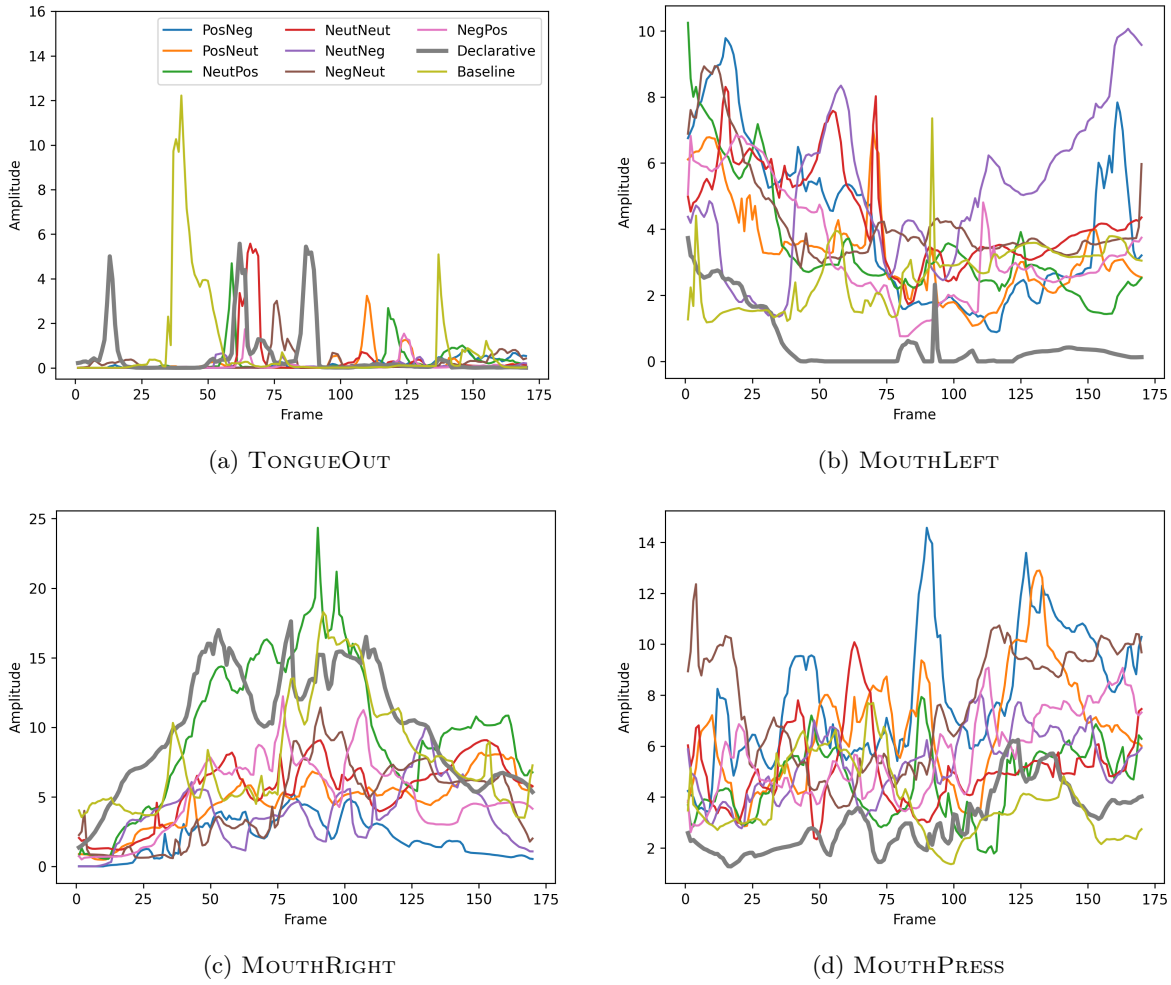
3.2.3 Jaw

The features concerning movements of the jaw are shown in Figure 3.3a. As expected, JawOpen is heavily influenced by the mouthings in this video. We further see that JawForward and JawLeft are also influenced by these mouthings, albeit to a lesser extent than JawOpen. In contrast to JawLeft, JawRight is not active in this video. This is likely subject-specific, and influenced by the angle of the camera.

Figures 3.3b and 3.3c, visualise how the measurements of JawRight and JawForward during the Declarative conditions are indistinguishable from those during the experimental conditions. The features related to movements of the jaw are all not present in the literature on question-specific NMMs, and we have no further reason to hypothesize that they are indicative markers of (polar) questions in NGT. As including these features in the data set would therefore only add noise, they are not selected.

3.2.4 Mouth

Many of the remaining features concern the mouth. These are: MouthClose, MouthFunnel, MouthPucker, MouthRight, MouthLeft, MouthSmile, MouthDimple, MouthStretch, MouthFrown, MouthRollLower, MouthRollUpper, MouthShrug, MouthPress, MouthLowerDown, MouthUpperUp, and TongueOut.



(a) TongueOut (b) MouthLeft

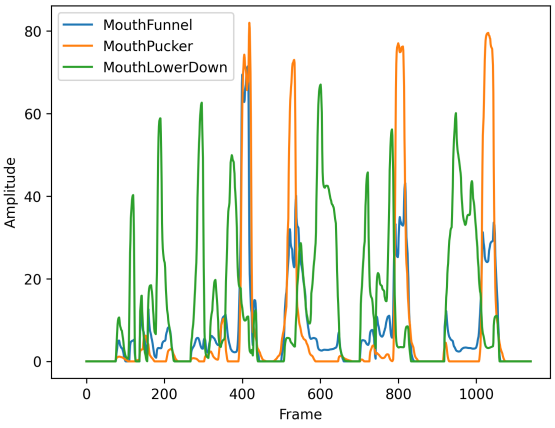
(c) MouthRight (d) MouthPress

Figure 3.4: Mean movements per condition for TongueOut (a); MouthLeft (b); MouthRight (c); and MouthPress (d)

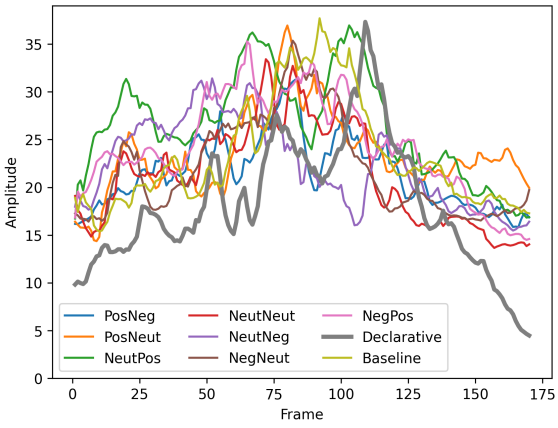
MouthShrug In Figure 3.3d, we see that the mean measurement of MouthShrug is decidedly lower during the Declarative condition than during the experimental conditions. This visualisation shows that this feature is likely affected by some of the mouthings, as the measurement in the Declarative condition does fluctuate. The effect is minimal however, with the mean amplitude barely passing 5. This is likely subject-specific, as this feature was not detected during M. Although the difference between measurements for these conditions is not very big at the beginning of the videos, we see that it increases as time passes. During the experiment, participants did display facial expressions with this feature to a noticeable extent while asking questions. We are therefore interested in exploring this feature further, and select it for our machine learning data set.

TongueOut In Figure 3.4a, we see that the feature TongueOut is not particularly engaged during any condition, and does not provide meaningful information about prototypical facial expressions marking polar questions. It is therefore not selected.

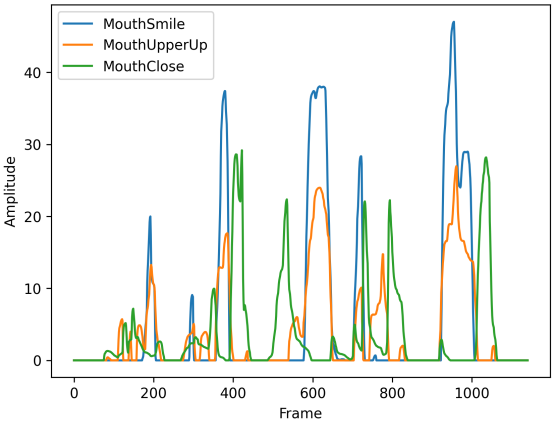
MouthLeft and MouthRight Figures 3.4b and 3.4c show the mean feature movement for MouthLeft and MouthRight over the experimental conditions. As these features measure a movement of both lips together towards a specific side of the face, they are inverse of each other. When MouthLeft is high, MouthRight is low and vice versa. The range of the mean amplitudes for MouthRight is a lot higher than those of MouthLeft. We see that the mean measurements during the Declarative conditions and the experimental conditions are not very different. For MouthLeft, the mean amplitude is slightly lower for the declarative condition than for the experimental conditions. The opposite is the case for MouthRight, where the mean amplitude is slightly higher for the declarative condition than for the experimental conditions. However, neither features are clearly distinguishable. As we have no further reason to hypothesize that these features are NMMs for polar questions in NGT, they are not selected.



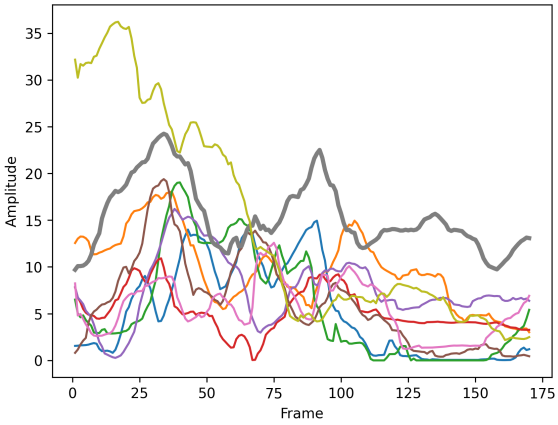
(a) Mouthings



(b) MouthFunnel



(c) Mouthings



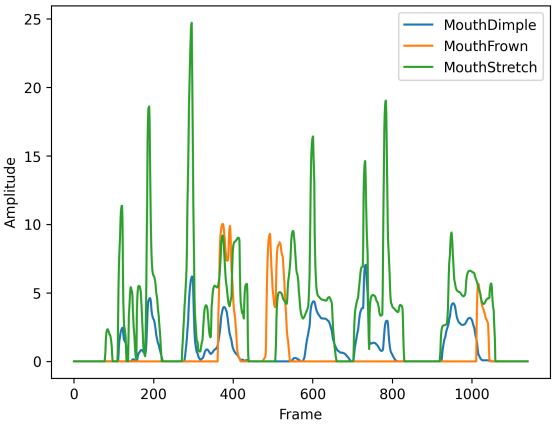
(d) MouthSmile

Figure 3.5: Feature measurements in M for MouthFunnel, MouthPucker, and MouthLowerDown (a); mean movements per condition for MouthFunnel (b); feature measurements in M for MouthSmile, MouthUpperUp, and MouthClose (c); and mean movements per condition for MouthSmile (d)

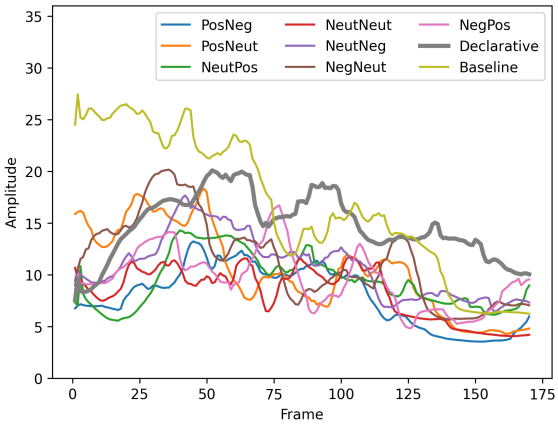
MouthPress Figure 3.4d, shows the low mean amplitude of MouthPress. Although the measurement for the Declarative condition is lower than for the other experimental conditions, they are not instantly distinguishable. Moreover, this feature is susceptible to noise added by mouthings, as the measurement for the Declarative condition intercepts the other conditions. In scenarios where the subject concerns a hypothetical person, 'Kim', this feature is measured at higher amplitudes than in other scenarios due to the lips pressing together at the 'm'. For these reasons, this feature is not selected.

MouthFunnel and MouthPucker In Figure 3.5a, we see that MouthFunnel is highly susceptible to mouthings in M, reaching amplitudes of 80 for the last four scenarios. This is due to specific mouthings in these scenarios (toegang (*entry*), open, thuis (*home*), and uur (*hour*). Further, figure 3.5b shows that the mean measurement of this feature in the Declarative condition is indistinguishable from the experimental conditions. MouthPucker is affected by the same mouthings as MouthFunnel in M, albeit to a lesser extent. Although it is not shown, like MouthFunnel, the measurement for the Declarative condition goes along with those of the experimental conditions. For these reasons, and in order to avoid noise in the data set, these features are not selected.

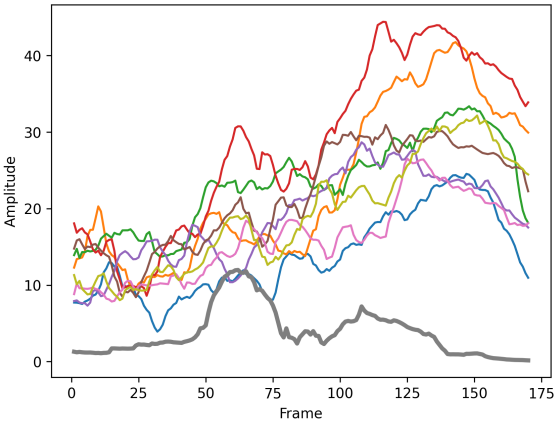
MouthLowerDown and MouthUpperUp Figure 3.5a shows that the mouthings in M have a significant effect on the feature MouthLowerDown, which measures a downward movement of the lower lip. Likewise, we see that the feature MouthUpperUp is also affected by the mouthings in M (Figure 3.5c). The measurements of this feature have a lower amplitude than MouthLowerDown, as this feature measures the upward movement of the lower lips, which naturally happens to a lesser extent while talking. These features are essentially the main features measuring mouthings, and would only add noise to the data set. Therefore, they are not selected for our machine learning task.



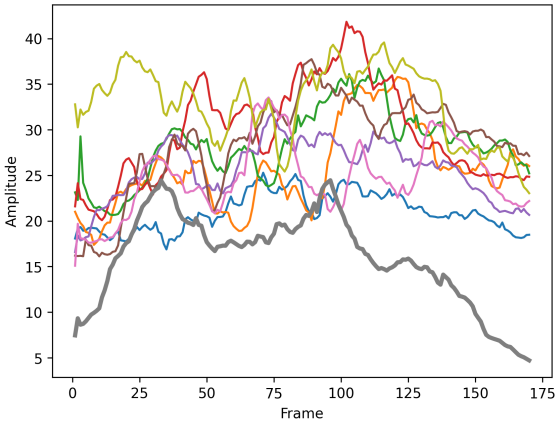
(a) Mouthings



(b) MouthDimple



(c) MouthFrown



(d) MouthStretch

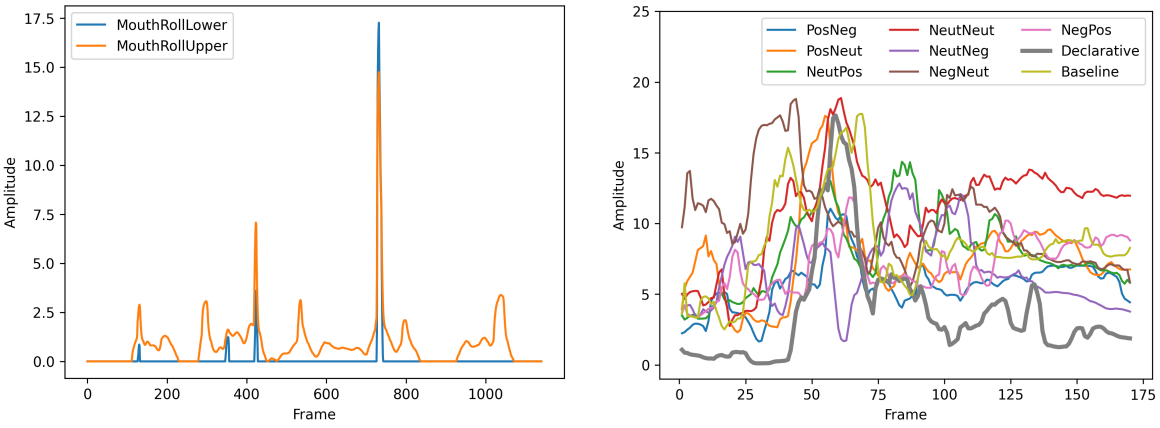
Figure 3.6: Feature measurements in M for MouthDimple, MouthFrown, and MouthStretch (a), mean movements per condition for MouthDimple (b), MouthFrown (c), and MouthStretch (d)

MouthSmile Figure 3.5c shows that the feature MouthSmile is greatly affected by the mouthings in M, reaching a maximum amplitude of 45. Moreover, in Figure 3.5d, we see that the mean measurement of this feature is generally higher in the Declarative condition than in the experimental conditions, but not by a large difference. As we do not hypothesize that this feature is indicative of polar questions in NGT, this feature is not selected.

MouthClose Like many features concerning the mouth above, MouthClose is influenced by the mouthings in M (Figure 3.5c). This feature measures how far the lips are closed, independent from JawOpen. In conjunction with a high value for JawOpen, a low value for MouthClose indicates that the subject has their jaw and their mouth wide open. On the other hand, a high value for JawOpen together with a high value for MouthClose indicates that the subject has their jaw lowered, but that their lips are closed together. This feature can therefore not be a NMM for polar questions in NGT, and is not selected for further investigation.

MouthDimple As the name suggests, the feature MouthDimple measures a backward movement of the mouth corners that typically causes dimples. Although slightly difficult to see due to the fluctuations of MouthStretch, Figure 3.6a shows that MouthDimple is affected by the mouthings in M. The amplitude of these not very high, reaching a peak around 8. Figure 3.6b, confirms that this feature is influenced by mouthings. We see that the mean measurements of this feature are generally higher than the experimental conditions, although they are close together. Since there are no reasons to hypothesize that this is a defining feature for marking polar questions in NGT, MouthDimple is not selected for our data set.

MouthFrown In Figure 3.6a, we see that the feature MouthFrown is affected by the mouthings of some scenarios in M. This is further illustrated in Figure 3.6c, where the mean measurement in the Declarative condition does show some fluctuations. However, in both Figures the amplitude of these fluctuations are relatively low, peaking around 10.



(a) Mouthings (b) MouthRollLower

Figure 3.7: Feature measurements in M for MouthRollLower and MouthRollUpper (a), and mean movements per condition for MouthRollLower (b)

Moreover, there is a clear difference between the mean measurements of MouthFrown during the Declarative condition and those during the experimental conditions. Like MouthShrug, we see an upwards trend over time. Further, the mean amplitude that is reached for this feature is quite high, reaching 45. To further investigate the role MouthFrown could play in marking polar questions in NGT, it is selected for our machine learning task.

MouthStretch In Figure 3.6a, we see that the feature MouthStretch is strongly affected by the mouthings in M. Although the amplitude is not extremely high with one peak at 25, this feature is constantly active throughout the duration of the video. Further, Figure 3.6d shows that the mean measurement during the Declarative condition is generally lower than the experimental conditions. However, the mean amplitude only truly diverges at the end of the videos. As it would introduce too much noise to the data set, MouthStretch is not selected.

MouthRollLower As seen in Figure 3.7a, the mouthings in M affect MouthRollLower less than they affect MouthRollUpper. However, there is one large peak around the mouthing of scenario 4, in which participants ask the confederate if the Efteling (a theme park in the Netherlands) is open this weekend. This peak is therefore likely caused by the lips rolling inwards during the mouthing of open. As further effect of the mouthings are minimal for MouthRollLower, we inspect its mean movements per condition in Figure 3.7b. This Figure shows that mouthings indeed affect this feature, as the mean movement for the Declarative condition does not differ much from the experimental conditions. We have no further reason to hypothesize that this feature could be a prototypical NMM for polar questions in NGT. Therefore, it is not selected.

MouthRollUpper As mentioned above, the feature MouthRollUpper is influenced by the mouthings in M. Although small, Figure 3.7a shows constant fluctuations. As this feature is essentially a more sensitive version of MouthRollLower, it is not selected for our machine learning task.

3.2.5 Summary

After our selection process, the final features consisted of: BrowInnerUp, BrowOuterUp, BrowDown, EyeWide, EyeSquint, CheekSquint, NoseSneer, MouthShrug, MouthFrown. The selection process was intentionally strict for the features that are not known by previous literature to be linguistically relevant, only selecting those where there was reason to assume that they might be relevant for question marking in NGT.

Chapter 4

Identifying most prototypical facial expressions

This chapter focuses on mainly on the methodological RQ2, how we can use techniques from machine learning to analyse the data. First, Section 4.1 will discuss the final steps in the preparation of the data for these specific research questions. Section 4.2 explains *clustering*, which is the ML technique applied to help analyse our data. This Section first gives an overview of K-means, which is one clustering algorithm that was initially attempted but rejected, and continues with an in-depth methodological description of the chosen algorithm, HDBScan.

4.1 Selecting samples

As discussed in Section 1.4, questions can be marked by a sequential combination of non-manual markers. Although transitions between facial expressions are interesting when examining their temporal progression over the duration of a question,[1] while identifying the most prototypical facial expressions we are not interested in the shift *between* expressions, but rather the expressions themselves. Therefore, the next step is selecting the relevant samples of each video of a question. A 'relevant sample' is defined as a sample in which 1) the features are relatively constant, and 2) at least one feature is engaged (i.e. the facial expression is not entirely neutral).

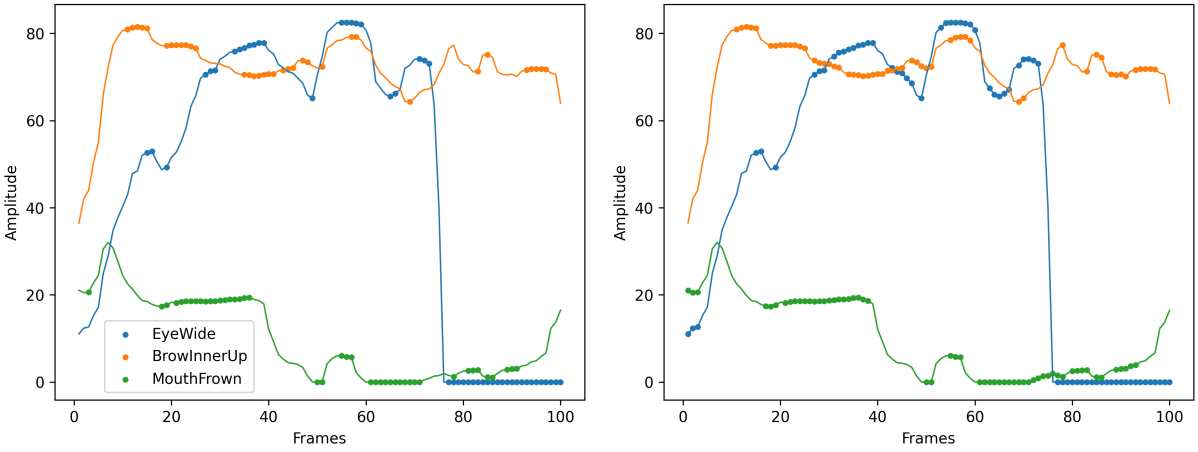
As these regions often move independently of each other, features are split up into two groups: the upper region and the lower region. The upper region comprises EyeWide, EyeSquint, BrowDown, BrowInnerUp, and BrowOuterUp. The lower region comprises MouthFrown, MouthShrug, CheekSquint, ans NoseSneer.

For this Chapter, videos of the DECLARATIVE condition are not considered, as they do not concern polar questions. This leaves 160 videos, amounting to a total of 27190 frames (samples). Further, each video is downsampled, discarding every third frame (the reason for this will be discussed in Section 4.2.2 below). This leaves a total number of 18127 samples. For every video, the selection of samples is made in two passes. We will refer to the total collection of samples in the data set as S_t.

First pass The first pass checks whether each individual feature is constant across samples. First, the rate of change is calculated between every sample and its predecessor. A sample is marked as 'qualifying' the first pass if 1) its rate of change is below the cutoff point x, and either 2.a) the rate of change for at least 2 neighboring samples are also below the cutoff point, or 2.b) the sample is at a local maximum or a local minimum. Figure 4.1 shows the qualifying samples of three features for different values of the cutoff point, x. We see that $x = 1$ and $x = 2$ are relatively strict, disqualifying a number of samples that are relatively similar to their neighbors. On the other hand, $x = 4$ and $x = 5$ are relatively lenient, qualifying a number of samples in which features are transitioning (this is especially prominent for EyeWide). A cutoff point of $x = 3$ is ideal; the qualifying samples are all relatively constant, and transitioning frames are disqualified.

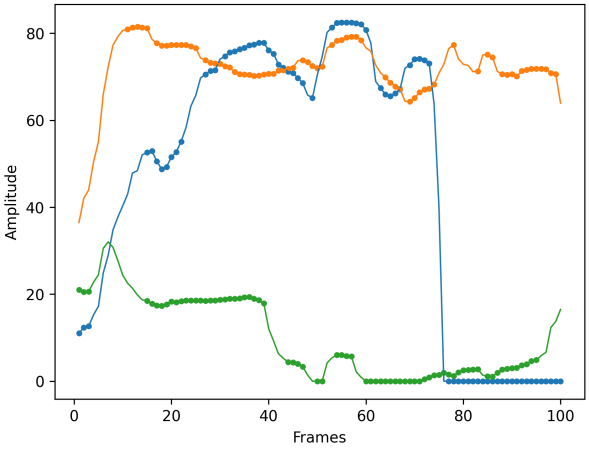
Second pass In the second pass, the decision is made on which samples make the final selection and which do not. Instead of considering each feature individually, the two groups of features are regarded as a whole. A sample qualifies if, for at least one of the groups, 1) all features in that group meet the requirements of being relatively constant, and 2) at least one feature is engaged. A feature is regarded as being engaged if the amplitude for that feature is above the threshold t. A value of $t = 7$ is used to ensure that at least one feature is *visibly* active, and to account for small variations above the neutral position. Figure 4.2a shows an example of the final

[1] This will be investigated further in future research.

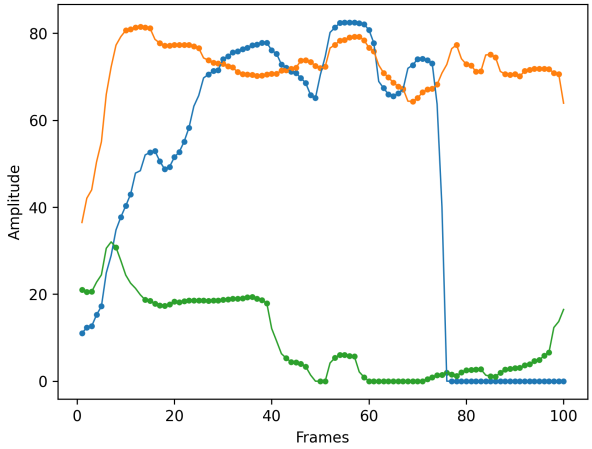


(a) $x = 1$

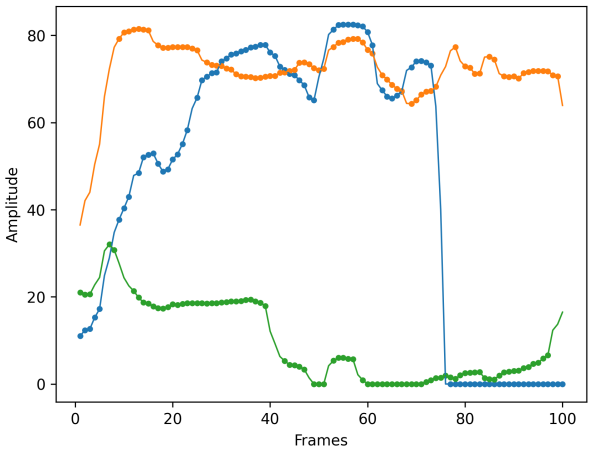
(b) $x = 2$



(c) $x = 3$

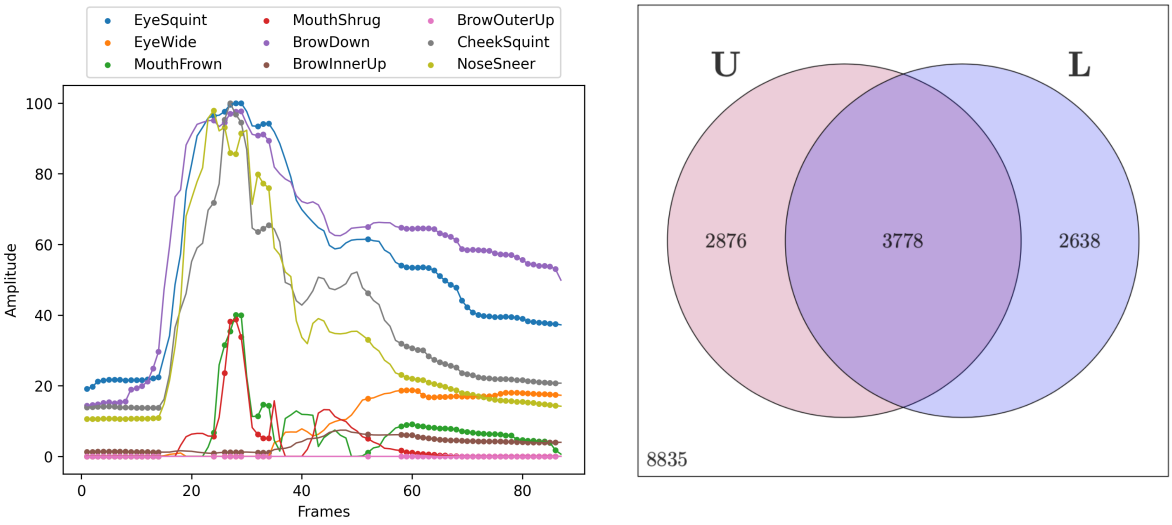


(d) $x = 4$



(e) $x = 5$

Figure 4.1: The effect on which frames are kept at different values of x, the accepted rate of change



(a) Selected frames

(b) Distribution of selected frames

Figure 4.2: An example of final selected frames for one video (a), and the distribution of selected frames over the two groups (b)

selected samples for one question. We see that especially samples at the beginning and the end of the question, and a few in the middle, are selected. Samples in which features are transitioning to other amplitudes are discarded. Figure 4.2b shows the final distribution of samples: 8835 samples are discarded; 2876 samples are kept of which only the features in the upper region qualify; 2638 samples are kept of which only the features in the lower region qualify; and 3778 samples are kept of which features in both the upper and lower regions qualify. S_t now contains 9292 samples.

4.2 Clustering

To find the most prototypical facial expressions, the resulting data set was used for *clustering*, which is an unsupervised machine learning technique. This technique is typically used on unlabeled data in order to obtain new insights from the data. In clustering, a model takes the input data set and assigns its data points into groups (or 'clusters'). There are several different algorithms that can be used to cluster a data set. Two of these algorithms were tested during this project: K-means and HDBScan. The latter of which was determined to be the best fit for this project. The following subsections provide an overview of how each algorithm works, and why they were considered. As K-means was not selected, Section 4.2.1 only provides a general overview on the algorithm, its implementation, and the reason why it was rejected. Section 4.2.2 will discuss the HDBScan algorithm and its implementation in depth. The results of this method are discussed in Chapter 5.

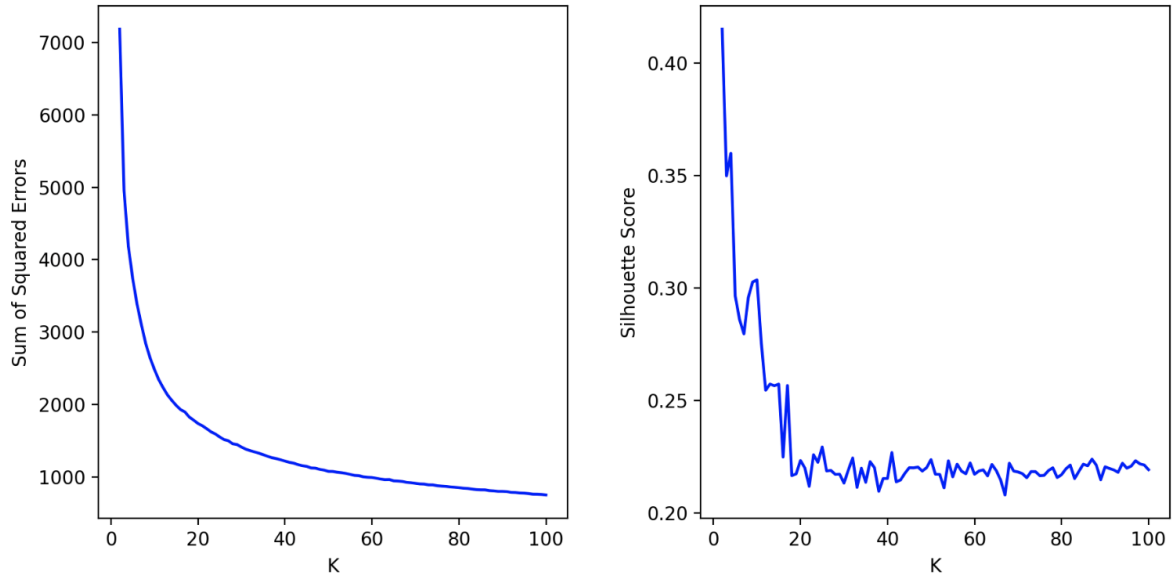
4.2.1 K-means

K-means is one of the most widely-used clustering algorithms. It clusters data points by separating them into a predetermined number of K classes which have equal within-class variance (MacQueen, 1967). This variance is the sum of squared errors (SSE) of the distance of each data point in a class to its centroid. This method was considered for the present project as it is relatively simple to implement and has been successfully used in different fields including sign language research (Zhang et al., 2011; Dardas and Georganas, 2011; Schmitt and McCoy, 2011).

The algorithm first randomly labels each data point by assigning it to a class, and computes the centroids of each class. It then updates the labels of each data point to the nearest centroid, after which the new centroids are calculated. This process continues until the model converges and no longer assigns new labels to the data points (Zhao, 2020).

K-means was implemented with the Scikit-Learn library (Pedregosa et al., 2011). Two approaches were taken to determine K: the elbow method and the silhouette score. For the elbow method, the SSE is calculated and plotted for a range of values of K. The optimal value of K is then at the 'elbow' of the plot: the point where the SSE no longer decreases by a significant amount for each additional class (Zhao, 2020). A large value for the SSE indicates that clusters are too big, and likely contain several smaller clusters. The silhouette score ss is calculated for each sample in a data set and is defined in equation 4.1.

$$ss = \frac{b - a}{max(a, b)} \tag{4.1}$$



(a) Elbow method results

(b) Silhouette score results

Figure 4.3: Results of the elbow method (a) and the Silhouette score (b): two methods to select the number of clusters, K, for K-means

Where: a is the mean distance between a sample and all the other points in its class and b is the mean distance between a sample and all the other points in the closest class (Loukas, 2020). The mean ss of S_t then indicates how well these clusters fit the data: 1 is the best score, -1 the worst, and 0 indicates overlapping clusters (Loukas, 2020). We want to select the lowest value for K that is near the elbow of the within-class variance plot and has a high silhouette score.

For both methods, scores were calculated for $K = 2$ to $K = 100$. For the elbow method, the 'elbow' in Figure 4.3a occurs around $K = 20$. However, Figure 4.3b shows that the Silhouette score around $K = 20$ is at it's lowest point. It decreases from its highest point at $K = 2, ss = 0.41$ – peaks around $K = 10, ss = 0.31$ – stabilizing at its lowest point around $K = 20, ss = 0.22$. While decreasing K results in a higher Silhouette score, it also leads to a (quickly significantly) higher SSE. Moreover, these silhouette scores are quite low, indicating that it is difficult to find clear boundaries between clusters in the data. It is therefore difficult to determine an optimal value for K for this data set.

Several values for K were attempted, but the resulting clusters were not very informative. Either K was too small, creating large clusters that lumped together discriminatory information; or K was too large, creating many clusters based on a small number of samples originating from single participants.

K-means essentially partitions the data, assigning every sample to a cluster. However, not all samples from this experiment necessarily belong to a prototypical facial expression for question marking. Thus, when these samples are assigned to the (limited number of) clusters, they typically merely add noise. Considering this fact, in addition to the limitations of the clusters produced by K-means, this algorithm was not selected for further investigation.

4.2.2 HDBScan

The final implemented algorithm was HDBScan: hierarchical density based clustering. This algorithm was considered for this research project as it, most importantly, allows clusters to have varying densities, whereas the most common clustering algorithms do not. Further, in contrast to K-means, HDBScan allows clusters to have variations in sizes, and not all data points are necessarily assigned to a cluster. Clusters are formed on the basis of dense regions in the data set, and samples that do not clearly belong to a cluster are classified as noise. This is especially relevant for the data in this project. Although the data set has been pre-processed to select the most relevant features and samples, not all samples necessarily belong to a prototypical facial expression. Moreover, because of the way the data is structured and pre-processed, it is highly likely that large variations exist in the number of samples that belong to each facial expression.

Algorithm HDBScan has two main parameters: MINIMAL_CLUSTER_SIZE, and MIN_SAMPLES. The first parameter regards the minimal number of data points that a cluster must contain. The second regards the minimal number of data points in the proximity of a point for it to be considered as a 'core

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
	1	1	–	–	–	100	**42**	**80**	–	**75**	**42**	**44**	**67**	18	–
	2	1	–	100	–	–	**49**	–	**58**	–	–	30	**53**	21	9
	3	2	–	100	–	–	**54**	–	**64**	–	–	**51**	**40**	23	9
1a	4	2	25	29	–	46	28	23	–	–	–	30	**74**	9	–
	5	13	59	27	8	6	–	**82**	–	**77**	**70**	–	21	–	–
	6	53	11	33	33	23	23	–	**44**	–	–	–	–	12	9
	N	28													
	1	12	61	28	5	6	–	**82**	–	**77**	**70**	–	20	–	–
1b	2	56	10	36	32	22	25	–	**46**	–	–	–	–	12	9
	N	32													

(a) Original data set

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
	1	2	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	3	–	100	–	–	**51**	–	**60**	–	–	**45**	**46**	21	9
	3	2	38	4	39	19	–	–	–	**81**	**76**	–	18	–	–
2a	4	2	23	21	–	56	24	16	–	–	–	23	**73**	8	–
	5	14	52	23	16	9	–	**80**	–	**76**	**69**	–	21	–	–
	6	55	12	37	31	20	26	–	**46**	–	–	–	–	13	9
	N	22													
	1	16	44	20	10	26	–	**82**	–	**75**	**66**	–	25	–	–
2b	2	67	13	35	30	22	26	–	**42**	–	–	–	8	13	9
	N	17													

(b) Downsampled data set

Table 4.1: Clusters (C) formed in implementations (I) of HDBScan for the original data set (a) and the downsampled data set (b). Samples assigned to a cluster (S_c) are shown as a percentage of S_t; their distribution over participants; and their median values. Values above 40 are marked in bold; values below 8 are not shown. Features: EyeSquint (*ES*), EyeWide (*EW*), BrowDown (*BD*), BrowInnerUp (*BIU*), BrowOuterUp (*BOU*), MouthShrug (*MS*), MouthFrown (*MF*), CheekSquint (*CS*), NoseSneer (*NS*)

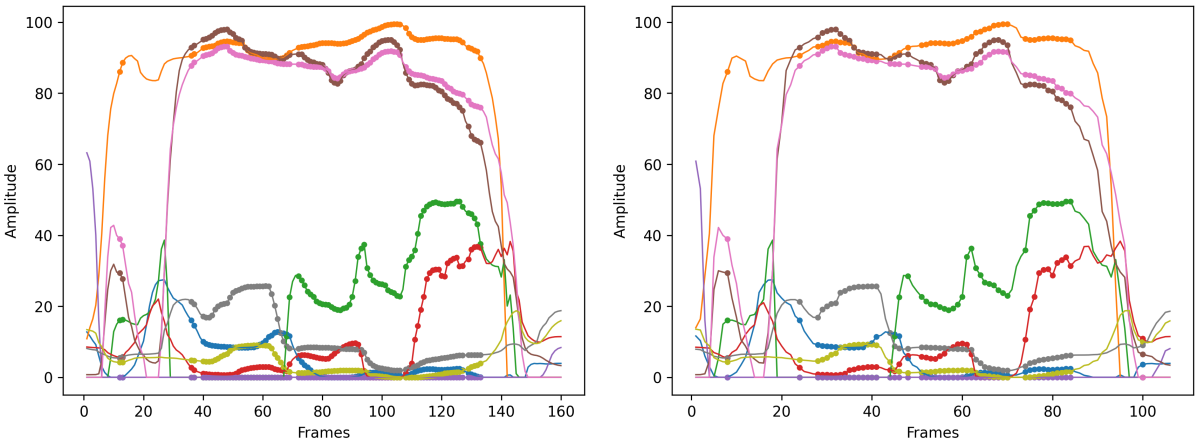
point' instead of noise; a lower value makes the clustering more lenient, while a larger value makes it more conservative (Campello et al., 2013).

HDBScan first transforms the space based on the levels of density and sparseness in the data. To prevent that two clusters are seen as one due to points of noise in between them, samples in dense areas are kept at the same distance to each other, while samples in sparse areas are pushed further away (Campello et al., 2013). It then creates a tree of samples, connecting each sample to the sample nearest to it. This tree is used to compute the cluster hierarchy. Every sample in the tree is first assigned to one cluster, which then iteratively splits up by removing samples from the root of the tree, creating new clusters (Campello et al., 2013). If these new clusters contain less samples than the predetermined minimal_cluster_size, the samples in the cluster are labeled as noise; otherwise, they are assigned to their new cluster (Campello et al., 2013). The most dense clusters in the tree are finally returned.

Implementation For a density based algorithm, the steps of normalising the data and selecting relevant samples are especially important. Otherwise, clusters will mostly be formed on the basis of neighbouring samples, as the values of features in these samples lie close together. Table 4.1a shows an example formation of 6 clusters (C), for each cluster listing: the percentage of S_t that it represents; how the samples assigned to that cluster, S_c, are distributed over participants; and the median values of its most engaged features.[2] While the model is applied to the *normalised* data set, the values of the

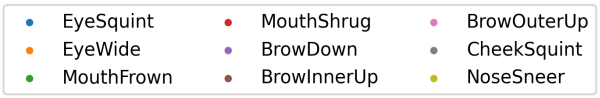
[2]For simplicity, feature values below 8 are not shown. The complete tables can be found in Appendix B.2.1.

[3]For an overview of the differences between these data sets and why they are used in this manner, see Section 2.3.2.



(a) Original data set

(b) Downsampled data set



(c) Legend

Figure 4.4: Final selected frames for the original data set (a), and the downsampled data set (b)

results are calculated according to the *ranged* data set.[3] Lastly, it is possible that a cluster's most engaged features include a combination of seemingly contradicting features (see Appendix B.1 for an example). Although a combination of these features during a singular frame seems unlikely, it occurs more often than one would expect. Their simultaneous presence in a cluster can be partly explained due to some samples in S_c containing a high value for EyeSquint, for example, while others contain a high value for EyeWide. If the median values of contrasting features are roughly the same, it can be an indication that the cluster is less 'pure', containing variations of facial expressions.

As seen in implementation 1a of Table 4.1a, despite normalising the data and selecting relevant samples, clusters are still formed around samples from single participants. 28% of the samples in the data is classified as noise (N). Clusters 1a.1-3 together represent only 4% of S_t, and $S_{1a.1} - S_{1a.3}$ are all completely made up of samples from one participant. Cluster 1a.4 represents the participants better, but still only accounts for 2% of S_t. Cluster 1a.5 represents a higher amount of S_t (13%), but $S_{1a.5}$ is still heavily skewed towards two participants. The distribution of $S_{1a.6}$ over participants is better for cluster 1a.6, which contains 53% of S_t. However, the median values of the most engaged features in this cluster are quite low compared to the other clusters.

Increasing the number of clusters leads to an increased amount of small clusters in which all samples in S_c are from one participant. Therefore, we will focus on a decreased number of clusters. As seen in implementation 1b in Table 4.1a, tweaking the parameters has little impact: the distribution over participants is mostly unchanged, leaving only the two larger clusters from implementation 1a (these will be referred to as 'super-clusters'). 32% of S_t are classified as noise. Cluster 1b.1 is mostly the same as in implementation 1a, representing 12% of S_t, and $S_{1b.1}$ again skewed towards two participants. The most engaged features are EyeWide, BrowInnerUp, BrowOuterUp, and to a lesser degree MouthFrown. Cluster 1b.2, represents 56% of S_t. The most engaged features are EyeSquint, BrowDown, and to a lesser degree CheekSquint, and NoseSneer. Again, the median values of these features are substantially lower than those in cluster 1b.1.

Downsampling It is not necessarily bad if the distribution of S_c is not completely even for all participants, as it is natural that some facial expressions may be favored by one or two participants but not by the others. Moreover, the total number of samples per participant varies slightly. It is therefore not a realistic benchmark for clusters to have an even distribution of samples. However, it is not possible to draw conclusions about variations in facial expressions if clusters are solely based on samples from one participant, and if S_c only represents a negligible percentage of S_t. In order to be able to do this, clusters should represent at least two participants, and at least 5% of S_t.

In order to improve these two factors, the density surrounding participants needs to be reduced. As mentioned in Section 4.1, every third sample in the data set is discarded. In addition to reducing density caused by neighbouring samples, downsampling leads to an improved selection in relevant sam-

		Bin									
		1	2	3	4	5	6	7	8	9	10
B10	Lower bound	0	10	20	30	40	50	60	70	80	90
	Upper bound	10	20	30	40	50	60	70	80	90	100
	Repl. value	0	15	25	35	45	55	65	75	85	95
B6	Lower bound	0	10	28	46	64	82	–	–	–	–
	Upper bound	10	28	46	64	82	100	–	–	–	–
	Repl. value	0	19	37	55	73	91	–	–	–	–
B4	Lower bound	0	10	40	70	–	–	–	–	–	–
	Upper bound	10	40	70	100	–	–	–	–	–	–
	Repl. value	0	25	55	85	–	–	–	–	–	–

Table 4.2: Categorisation of continuous values into bins of discrete values

ples. Figure 4.4 shows that the algorithm for selecting relevant samples is slightly more conservative for the downsampled data set (Figure 4.4b) as opposed to the original data set (Figure 4.4a). Although adjusted parameters were used to select the samples of the original data set,[4], we see that quite some selected samples are at points where features are transitioning (see EyeSquint, CheekSquint and NoseSneer circa frames 60-80 in Figure 4.4a). On the other hand, the downsampled data set retains the structural integrity of the data, while the selected samples are positioned better. The percentage of selected samples is slightly lower for the downsampled data set (51%) as opposed to the original data set (58%).

Table 4.1b shows that, although downsampling does have a positive influence on clustering, the effect is minimal – the distribution of samples over participants is improved marginally, and clusters represent a slightly larger portion of S_t. In implementation 2a, we see that most clusters are highly similar to those in implementation 1a. Clusters 2a.1; 2a.4; 2a.5; and 2a.6 are almost identical to clusters 1a.1; 1a.4; 1a.5; and 1a.6, respectively. Cluster 2a.2 shows some slight changes: it is now a combination of clusters 1a.2 and 1a.3. The biggest difference, however, is seen in cluster 2a.3: which is an entirely new cluster. It is still small, representing only 2% of S_t. The distribution of $S_{2a.3}$ over participants is fairly decent compared to other clusters, skewing towards participant 3 and 6, but representing participant 7 fairly well. Although participant 4 is not represented to a high degree in this cluster, some of their samples are present in $S_{2a.3}$. As with the original data set, increasing the number of clusters leads to more clusters in which S_c consists completely of samples from single participants. The two super-clusters in implementation 2b are again highly similar to those in implementation 1b, although the number of samples that are classified as noise is slightly lower.

Categorisation Although the above two implementations do result in some clusters with clear differences in facial expressions, they are not yet satisfactory. Whereas the two super-clusters are truly representative of the data, the elemental clusters generally do not allow for generalisation or further investigation. Nevertheless, they do suggest the existence of more variations in facial expressions. As the aim of this research project is to investigate these variations in facial expressions – and how they relate to their contexts – the data will be restructured. As mentioned in Section 1.2, one of the limitations of manual annotation is that features are labeled discretely: they are either engaged in a certain manner or they are not. On the other hand, Section 1.3.1 established that due to CV technology, features can be labeled with continuous values, denoting the amplitude of their engagement. However, this continuity of feature values has shown to lead to clusters that are based on neighbouring samples.

We categorise the data into 'bins', viewing feature values as discrete in lieu of continuous. This essentially forces the data away from forming dense areas predominantly with their immediate neighbouring samples, but rather towards forming dense areas with samples that have roughly the same values in common. In contrast to manual annotations however, feature values can be assigned to more fine-grained categories than 0 or 1. This sustains the benefit of more detailed empirical knowledge about features' levels of engagement.

This approach is implemented with bins of various sizes, producing 3 data sets: B10, B6, and B4. We will refer to the implementation of each data set as 'categorisation'. Table 4.2 shows the lower (l) and upper (u) bounds for each bin, and the subsequent replacement value (r) for normalised feature values (x) assigned to that bin. The first bin is identical for each categorisation: $l = 0$, $u = 10$, and $r = 0$. This allows us to distinguish between those feature values that are barely engaged (or fully disengaged),

[4] i.e. more neighbouring samples have to be relatively constant; lower tolerance for rate of change

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B10a	1	3	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	2	–	99	1	–	**49**	–	**58**	–	–	31	**54**	22	11
	3	2	20	25	–	55	24	25	–	–	–	32	**74**	9	–
	4	15	54	25	11	10	–	**81**	–	**77**	**69**	–	21	–	–
	5	62	13	33	32	22	24	–	**41**	–	–	–	–	12	9
	N	17													
B10b	1	17	51	24	11	14	–	**81**	–	**76**	**68**	–	21	–	–
	2	68	14	33	31	22	24	8	37	–	–	–	7	12	8
	N	16													

(a) B10

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B6a	1	3	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	1	–	99	1	–	**48**	–	**61**	–	–	34	**47**	21	12
	3	3	13	43	6	38	23	25	–	–	–	26	**74**	8	–
	4	15	53	24	10	13	–	**81**	–	**76**	**69**	–	21	–	–
	5	57	13	33	34	22	24	–	**42**	–	–	–	–	12	9
	N	21													
B6b	1	20	44	21	8	27	9	**81**	–	**73**	**66**	–	24	–	–
	2	69	14	35	30	21	27	7	42	–	–	–	7	13	9
	N	11													

(b) B6

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B4a	1	6	45	11	3	41	26	**81**	–	**74**	**64**	–	39	–	–
	2	7	5	71	1	23	**45**	–	**49**	–	–	33	**62**	22	7
	3	7	9	28	17	46	13	–	**44**	–	–	–	–	7	–
	4	7	35	12	20	33	10	37	–	–	–	–	12	–	–
	5	10	53	30	11	6	–	**81**	–	**77**	**69**	–	16	–	–
	6	13	1	45	40	14	**41**	–	**70**	–	–	–	–	17	13
	7	14	18	29	23	30	–	11	9	–	–	–	–	–	–
	8	15	6	29	55	10	**42**	11	**49**	–	–	–	8	19	12
	N	21													
B4b	1	19	47	20	7	26	11	**82**	–	**75**	**67**	–	25	–	–
	2	27	24	22	22	32	9	20	–	–	–	–	15	–	–
	3	39	8	38	36	18	39	–	**60**	–	–	–	–	17	12
	N	15													

(c) B4

Table 4.3: Clusters (C) formed in implementations (I) of HDBScan for B10 (a), B6 (b), and B4 (c). Samples assigned to a cluster (S_c) are shown as a percentage of S_t; their distribution over participants; and their median values. Values above 40 are marked in bold; values below 7 are not shown. Features: EyeSquint (*ES*), EyeWide (*EW*), BrowDown (*BD*), BrowInnerUp (*BIU*), BrowOuterUp (*BOU*), MouthShrug (*MS*), MouthFrown (*MF*), CheekSquint (*CS*), NoseSneer (*NS*)

and those that merely fall into a lower spectrum of engagement. Otherwise, if a data set is categorised into 3 bins, for instance, a feature with $x = 31$ will be assigned the same r as a feature that has been constant at $x = 0$. For the remaining bins, l, u, and r are equally spaced in the range of possible values.

In Table 4.3, we see that the clusters formed in implementation B10a and B10b are almost identical to those in implementation B6a and B6b, although clusters B6a.3 and B6b.1 are made up of a marginally more equal distribution of participant samples. Nevertheless, these data sets suffer from the same issue as previous implementations: additional clusters beyond the two super-clusters are often based on samples of only one participant. Using 10-6 bins is therefore still too fine-grained for our purposes.

A real difference is seen in the clusters formed in implementation B4a and B4b. The contents of Table 4.3c will be discussed in more detail in Chapter 5, but for the purpose of this section a few key points will be highlighted. First, we see that even for a larger number of clusters, the clusters formed in implementation B4a all represent at least 6% of S_t. Second, these clusters all represent every participant to at least a minimal degree – some distributions of S_c do skew towards a smaller number of participants, but as discussed above, this does not necessarily mean that these clusters are not representative of prototypical facial expressions or that they do not allow for generalisation. Third, a new super-cluster is added in implementation B4b, which is formed by samples with values that fall into bins 1 and 2, and were previously assigned to the other two super-clusters in other implementations, mostly to the counterpart of cluster B4b.3. Therefore, this additional super-cluster makes cluster B4b.3 more representative of the actual data set, as the samples with lower median values are now assigned to their own cluster, and no longer affect the median values of the other clusters. This was not possible in previous implementations, as adjusting the parameters to introduce a new cluster resulted in one that could not be regarded as a super-cluster (i.e. a cluster that is too skewed towards one or two participants, represents a too small portion of the data set).

Chapter 5

Results

The results in Tables 4.1, and 4.3 are example outcomes for only one combination of parameters. Adjusting the parameters does produce variations in results; however, there are no *optimal* parameters for this specific clustering exercise as there are multiple plausible results. Nevertheless, some outcomes are better than others – where clusters have better distribution of S_c over participants and represent more samples of S_t. The displayed results have been selected as a representative example of the outcomes of each implementation. Despite the fact that there are no optimal parameters, we have seen that adjusting the parameters influences how fine-grained the clusters are: increasing the minimal_cluster_size causes elemental clusters to converge into super-clusters, and vice-versa. We will first discuss the results of the rougher super-clusters and how they relate to the more fine-grained elemental clusters in Section 5.1. Next, we will compare the prototypical facial expressions through visualisations in Section 5.2. Finally, in Section 5.3 we map these facial expressions to contexts and temporal windows. All results that are discussed in this Section regard the density-based clustering implementation with the B4 data set, shown in Table 5.1.

5.1 Clusters

5.1.1 Super-clusters

Implementation B4b in Table 5.1 shows the three super-clusters. For this implementation, 15% of S_t is labeled as noise. Cluster B4b.1 represents 19% of S_t. 47% of $S_{B4b.1}$ originates from participant 3, 26% from participant 7, 20% from participant 5, and 7% from participant 6. The most engaged features (and their median values) in this cluster are EyeWide (82), BrowInnerUp (75), BrowOuterUp (67). Other features are MouthFrown (25) and EyeSquint (11). This cluster aligns with the 'raised eyebrows' + 'eyes wide' definition of 'q' in the literature, with the addition of a slight MouthFrown and EyeSquint.

Second, cluster B4b.2 represents a larger portion of S_t, 27%. $S_{B4b.2}$ is distributed over participants quite evenly, 32% originate from participant 7, 24% from participant 3, and 22% from both participant 5 and 6. Features in this cluster have a low median value, and are thus minimally engaged: EyeWide (20), MouthFrown (15), and EyeSquint (11). This cluster does not align with previous definitions of 'q'. It may well be that these relatively neutral frames are not really characteristic for question marking. We tried to filter out neutral frames in an earlier step, but it is possible that the samples in cluster B4b.2 are not really 'active' frames. Further samples in this cluster are likely those with feature values within the bounds of the first two bins, which are thus replaced by values of 0 or 25 (see Table 4.2).

The largest and final cluster B4b.3 represents 39% of S_t. $S_{B4b.3}$ mostly comprises samples originating from participant 5 (38%) and 6 (36%), while 18% originate from participant 7, and 8% from participant 3. As mentioned at the end of the previous Section, the median values of the most engaged features in this cluster are now more explicit compared to previous implementations. The most engaged features of this cluster are BrowDown (60) and EyeSquint (39). Other features are CheekSquint (17) and NoseSneer (12). This cluster aligns with the 'furrowed eyebrows' definition of 'q' found in the literature, with the addition of a prominent EyeSquint, and a slight CheekSquint and NoseSneer.

5.1.2 Elemental clusters

Implementation B4a in Table 5.1 shows the eight elemental clusters. 21% of S_t is labeled as noise, while the clusters do not have much variance in size.

First, cluster B4a.1 represents 6% of S_t. The majority of $S_{B4a.1}$ originates from participant 3 (45%) and 7 (41%); 11% originate from participant 5, and 3% from participant 3. The most engaged features of this cluster are EyeWide (82), BrowInnerUp (74), BrowOuterUp (64), MouthFrown (39);

I	C	S_t	S_c distribution (%)				Most engaged features (median value)								
		(%)	*P3*	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B4a	1	6	45	11	3	41	26	**82**	–	**74**	**64**	–	39	–	–
	2	7	5	71	1	23	**45**	–	**49**	–	–	33	**62**	22	7
	3	7	9	28	17	46	13	–	**44**	–	–	–	–	7	–
	4	7	35	12	20	33	10	37	–	–	–	–	12	–	–
	5	10	53	30	11	6	–	**81**	–	**77**	**69**	–	16	–	–
	6	13	1	45	40	14	**41**	–	**70**	–	–	–	–	17	13
	7	14	18	29	23	30	–	11	9	–	–	–	–	–	–
	8	15	6	29	55	10	**42**	11	**49**	–	–	–	8	19	12
	N	21													
B4b	1	19	47	20	7	26	11	**82**	–	**75**	**67**	–	25	–	–
	2	27	24	22	22	32	9	20	–	–	–	–	15	–	–
	3	39	8	38	36	18	39	–	**60**	–	–	–	–	17	12
	N	15													

Table 5.1: Final clusters by HDBScan. Samples assigned to a cluster (S_c) are shown as a percentage of S_t; their distribution over participants; and their median values. Values above 40 are marked in bold; values below 7 are not shown. Features: EyeSquint (*ES*), EyeWide (*EW*), BrowDown (*BD*), BrowInnerUp (*BIU*), BrowOuterUp (*BOU*), MouthShrug (*MS*), MouthFrown (*MF*), CheekSquint (*CS*), NoseSneer (*NS*)

and further including EyeSquint (26). This cluster aligns with the 'raised eyebrows' + 'eyes wide' definition of 'q' from the literature, with the addition of a prominent MouthFrown and a slight EyeSquint.

Second, cluster B4a.2 represents 7% of S_t. At 71%, most samples in $S_{B4a.2}$ originate from participant 5. 23% of samples originate from participant 7, while the remainder originate from participant 3 (5%) and participant 6 (1%). The most engaged features of this cluster are MouthFrown (62), BrowDown (49), and EyeSquint (45). Further features are MouthShrug (33), CheekSquint (22) and NoseSneer (7). This cluster aligns with the 'furrowed eyebrows' definition of 'q' found in the literature, with the addition of a prominent MouthFrown and EyeSquint, and a less prominent MouthShrug, CheekSquint, and NoseSneer.

Cluster B4a.3 also represents 7% of S_t. $S_{B4a.3}$ is more evenly distributed over participants than the previous cluster: 46% originates from participant 7, 28% from participant 5, 17% from participant 6, and 9% from participant 3. The most engaged feature of this cluster is BrowDown (44). Other features are EyeSquint (13) and CheekSquint (7). Like its predecessor, this cluster aligns with the 'furrowed eyebrows' definition of 'q' in the literature, differing in its additions: a slight EyeSquint and CheekSquint.

Like the previous two clusters, cluster B4a.4 represents 7% of S_t. Again, $S_{B4a.4}$ is more evenly distributed over participants than the previous cluster: 35% originate from participant 3, 33% from participant 7, 20% from participant 6, and 12% from participant 5. The most engaged feature is EyeWide (37). Other features are EyeSquint (10) and MouthFrown (12). This cluster does not align with previous definitions of 'q' from the literature. We refer back to the earlier comment on supercluster B4b.2, that these relatively neutral frames may not be characteristic for question marking as they might not be 'active', and that further detail on some frames in this cluster may be lost due to the categorisation of feature values.

The next cluster, B4a.5, represents 10% of S_t. $S_{B4a.5}$ is skewed towards participant 3 (53%) and participant 5 (30%). 11% of $S_{B4a.5}$ originates from participant 6, and 6% from participant 7. The most engaged features of this cluster are EyeWide (81), BrowInnerUp (77), BrowOuterUp (69), and to a lesser degree MouthFrown (16). Like cluster B4a.1, this cluster aligns with the 'raised eyebrows' + 'eyes wide' definition of 'q' from the literature, albeit with only the addition of a slight MouthFrown.

Cluster B4a.6 represents 13% of S_t. The majority of $S_{B4a.6}$ originates from participant 5 (45%) and 6 (40%). 14% of $S_{B4a.6}$ originate from participant 7, and 1% from participant 3. The most engaged features of this cluster are BrowDown (70) and EyeSquint (41). Other features are CheekSquint (17) and NoseSneer (13). Like clusters B4a.2 and B4a.3, this cluster aligns with the 'furrowed eyebrows' definition of 'q' found in the literature. One differentiating factor between these earlier clusters and cluster B4a.6, however, is that samples in this cluster have a higher amplitude for BrowDown than the samples assigned to other clusters in

		S_{EC}									S_t
		1	2	3	4	5	6	7	8	N	
S_{SC}	1	33	–	–	–	52	–	–	–	15	19
	2	–	7	–	24	–	–	50	4	15	27
	3	–	4	18	–	–	32	–	35	11	39
	N	–	21	–	3	3	–	–	3	70	15

(a) S_{SC} in terms of S_{EC}

		S_{SC}				S_t
		1	2	3	N	
S_{EC}	1	99	–	–	1	6
	2	–	30	21	49	7
	3	–	–	100	–	7
	4	–	93	–	7	7
	5	96	–	–	4	10
	6	–	–	100	–	13
	7	–	99	–	1	14
	8	–	8	89	3	15
	N	12	19	19	50	21

(b) S_{EC} in terms of S_{SC}

Table 5.2: Percentage composition of super-clusters (SC) and elemental clusters (EC) from B4

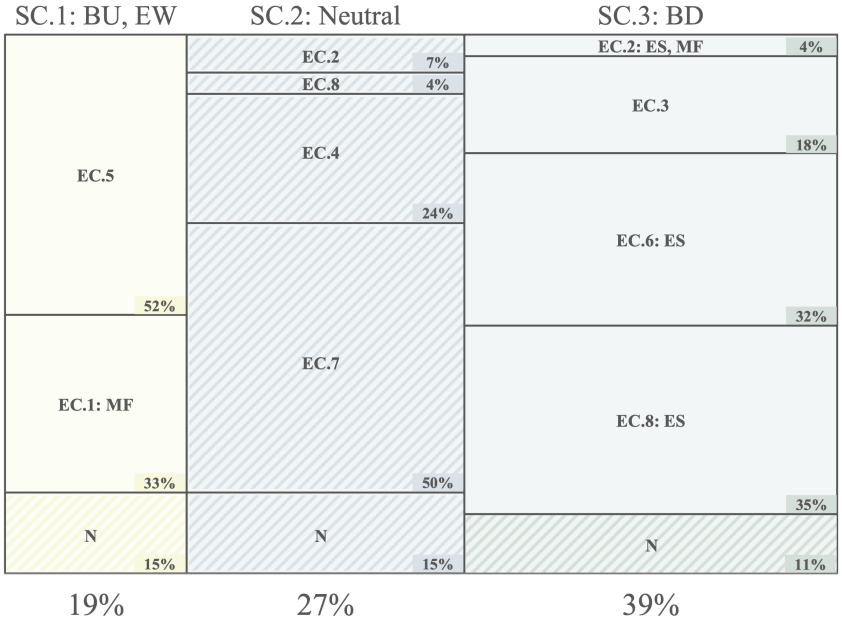


Figure 5.1: Visualisation of Table 5.2a

this category. The additional EyeSquint, CheekSquint, and NoseSneer for this cluster are similar to the additional features of cluster B4a.2, except for the MouthFrown.

Cluster B4a.7 represents 14% of S_t. $S_{B4a.7}$ is distributed fairly evenly over the participants: 30% originates from participant 7, 29% from participant 5, 23% from participant 6, and 18% from participant 3. The features in this cluster are minimally engaged: Eyewide (11) and BrowDown (9). Like cluster B4a.4, this cluster does not align with previous definitions of 'q' found in the literature. Again, the same remark applies to cluster B4a.7 as to cluster B4a.4 and super-cluster B4b.2.

Finally, cluster B4a.8 represents 15% of S_t. 55% of $S_{B4a.8}$ originates from participant 6, 29% from participant 5, 10% from participant 7, and 6% from participant 3. The most engaged features in this cluster are BrowDown (49) and EyeSquint (42). Other features include CheekSquint (19), NoseSneer (12), EyeWide (11), and MouthFrown (8). Like clusters B4a.2, B4a.3, and B4a.6, this cluster aligns with the 'furrowed eyebrows' definition of 'q' in the literature. Its features (and median values) are most similar to cluster B4a.2, barring the more defined MouthShrug and MouthFrown in that cluster.

5.1.3 Relating super-clusters and elemental clusters

Table 5.2 shows the composition of super-clusters (SC) and elemental clusters (EC) in terms of each other: how every S_{SC} relates to S_{EC}, and how every S_{EC} relates to S_{SC}. We see mostly clear and expected divisions of the clusters, that are generally in line with the definition of 'q' from the literature a cluster relates to most. This is further visualised

(a) Cluster 4 (b) Cluster 4 (c) Cluster 7 (d) Cluster 7

Figure 5.2: Visualisation of more neutral facial expressions (do not align with previous definitions of 'q')

(a) Cluster 1 (b) Cluster 1 (c) Cluster 5 (d) Cluster 5

Figure 5.3: Visualisation of facial expressions with raised eyebrows

in Figure 5.1, which shows how much of S_t each super-cluster represents; how each super-cluster is composed in terms of the elemental clusters; the main active feature of each super-cluster; and any additional features that the elemental clusters may have. The 15% of S_t that is labeled as noise is not shown in this Figure.

As expected, $S_{SC.1}$ is mostly composed of samples that are likewise in $S_{EC.1}$ and $S_{EC.5}$; the remaining $S_{SC.1}$ were labeled as noise in implementation B4a. All of these clusters mostly align with the 'raised eyebrows' + 'eyes wide' definition of 'q' found in the literature.

$S_{SC.2}$ is mostly composed of samples that are in $S_{EC.7}$ and $S_{EC.4}$. The remainder of $S_{SC.2}$ is labeled as noise in implementation B4a, or $S_{EC.2}$ and $S_{EC.8}$. The former two are expected: neither elemental cluster 7 nor 4 align with a previous definition of 'q', and mean feature values in these clusters are low. Although less expected, samples that are labeled as elemental clusters 2 and 8 likely have low feature values.

Finally, $S_{SC.3}$ is mostly composed of samples that are likewise in $S_{EC.8}$, $S_{EC.6}$, and $S_{EC.3}$. The remaining samples in $S_{SC.3}$ are labeled as noise in implementation B4a, or as $S_{EC.2}$. Again, this is expected, as these clusters align most with the 'furrowed eyebrows' definition of 'q' in the literature.

5.2 Visualising characteristic facial expressions

This Section visualises the resulting relevant facial expressions for biased polar question marking in NGT. Only the elemental clusters are considered, as they provide the highest level of detail. The prototypical facial expressions derived from the elemental clusters are visualised on a MetaHuman avatar in Unreal Engine, and as a video frame of a participant where the blendshape values lie closest to the median blendshape values for that cluster (as seen in Table 5.1). Note that the avatar visualisations are made so that the avatar's mouth is always closed, whereas the mouths of participants may be opened.

5.2.1 Neutral expressions

First, the clusters with the 'neutral' facial expressions (cluster 4 and 7) are shown in Figure 5.2. These expressions do not align with any definitions of 'q' in the literature. Although the examples of

(a) Cluster 2 (b) Cluster 2 (c) Cluster 3 (d) Cluster 3

(e) Cluster 6 (f) Cluster 6 (g) Cluster 8 (h) Cluster 8

Figure 5.4: Visualisation of facial expressions with furrowed eyebrows

these expressions on the avatars are very similar, the eyes are slightly wider for that of cluster 4. This is more visible in the participant visualisations.

5.2.2 Raised eyebrows

The clusters with facial expressions that align with the 'raised eyebrows' + 'eyes wide' definition of 'q' found in the literature (cluster 1 and 5), are shown in Figure 5.3. Again, these expressions seem quite similar at a first glance. However, in both the avatar and participant visualisations of these clusters we see that the MouthFrown is more defined for cluster 1 than for cluster 5.

5.2.3 Furrowed eyebrows

The clusters with facial expressions that align with the 'furrowed eyebrows' definition of 'q' in the literature (clusters 2, 3, 6, and 8) are shown in Figure 5.4. First, cluster 2 is the only one in its category with a high median value for MouthFrown and MouthShrug, which is clearly visible in both the avatar and participant visualisations (Figures 5.4a and 5.4b). In fact, it has the highest median values for these features of all elemental clusters. It is important to note that 71% of $S_{EC.2}$ originates from participant 5, and 23% from participant 7, whose facial expressions often contained these features.

Cluster 3 seems to be the most 'neutral' of this category. It has the lowest median values for all features. In the visualisations in Figures 5.4c and 5.4d, the most visible feature is BrowDown – more so with the participant than with the avatar. Compared to the other clusters in this category, cluster 3 is less defined in the CheekSquint feature, which can be seen in the visualisations of the avatars and the participants.

Cluster 6 is the most defined for the feature BrowDown, with a median value of 70. This is clearly visible in both the visualisation of the avatar and the participant in Figures 5.4e and 5.4f. These visualisations also show EyeSquint and CheekSquint well.

Figures 5.4g and 5.4h show that the facial expression derived from cluster 8 differentiates itself from the other expressions in this category. EyeSquint is more defined than in cluster 3, but less than clusters 2 and 6 due to the addition of EyeWide. Like clusters 2 and 6, CheekSquint is visible. Moreover, MouthFrown is more defined than for clusters 3 and 6, but significantly less than cluster 2. This is less visible in the visualisations of the participants, and more on those of the avatar.

			Window				
			1	2	3	4	5
S_{EC}	SC.1	1	11	17	23	27	22
		5	14	26	20	22	18
	SC.2	4	18	17	10	18	37
		7	46	16	10	10	18
	SC.3	2	1	2	12	36	48
		3	28	22	8	13	28
		6	12	19	23	21	25
		8	6	21	22	26	25

(a) S_{EC} over windows

		SC.1		SC.2		SC.3				
		1	5	4	7	2	3	6	8	N
S_{cond}	Baseline	8	12	13	12	4	4	2	22	23
	NegPos	3	8	6	16	5	11	17	11	23
	NegNeut	8	6	4	15	9	4	12	15	27
	NeutNeg	2	12	12	16	5	6	16	18	13
	NeutNeut	6	12	3	14	11	4	10	11	29
	NeutPos	12	25	7	9	5	9	5	14	14
	PosNeut	10	8	4	13	8	4	14	18	21
	PosNeg	–	–	7	13	5	17	23	15	20

(b) $S_{condition}$ over clusters

			Condition							
		EC	Baseline	NegPos	NegNeut	NeutNeg	NeutNeut	NeutPos	PosNeut	PosNeg
S_{EC}	SC.1	1	15	6	17	5	14	20	23	–
		5	13	10	7	14	16	27	12	–
	SC.2	4	21	12	9	21	6	10	8	14
		7	9	14	15	15	15	7	13	12
	SC.3	2	6	9	18	10	23	8	16	10
		3	7	18	7	9	7	13	7	31
		6	2	17	12	15	10	4	16	24
		8	16	9	13	14	10	9	16	13

(c) S_{EC} over conditions

Table 5.3: Relation between clusters and windows, and clusters and conditions. For each window (W), how S_W is distributed over clusters (a), for each condition, how $S_{condition}$ is distributed over clusters (b), and for each cluster, how S_{EC} is distributed over the conditions (c)

5.3 Mapping facial expressions to contexts and temporal windows

In this section, we discuss how facial expressions found in these clusters map to the context of the experimental conditions, and how they map to the temporal windows. As in the previous Section, we limit our considerations to the elemental clusters, which provide the highest level of detail.

As detailed in Section 2.3.1, the samples from each video have been distributed over five temporal windows. The first 20% of samples from each video are assigned to the first temporal window (W1), the next 20% to the second temporal window (W2), continuing up to the final temporal window (W5). These windows thus allow us to roughly model temporal development of the clusters by examining how clusters relate to these temporal windows.

5.3.1 Neutral expressions

The clusters with neutral facial expressions are clusters 4 and 7. As seen in Table 5.3a, of all the samples assigned to cluster 4, 37% occurs in W5 – at the end of a question – and 18% in W1, at the beginning of a question. Moreover, in Table 5.3c, we see that 21% of $S_{EC.4}$ occurs during the Baseline condition, and another 21% during the NeutNeg condition.

Like cluster 4, the facial expression derived from cluster 7 occurs most at the beginning and at the end of a question: of $S_{EC.7}$, 46% occurs in W1 and 18% in W5. In fact, 38% of $S_{W.1}$ are assigned to cluster 7, which is the highest percentage of S_W assigned to any cluster, including noise (Appendix C.1). $S_{EC.7}$ is more evenly spread over the conditions than $S_{Ec.4}$. It is therefore likely that the samples assigned to these clusters are simply reflections of participants' neutral facial expressions that occurred before and after the 'true' prototypical facial expressions for question marking.

5.3.2 Raised eyebrows

The clusters with facial expressions that align with the 'raised eyebrows' + 'eyes wide' definition of 'q' from the literature are clusters 1 and 5. As seen in Table 5.3a, samples in $S_{EC.1}$ most frequently occur at the end of a question: in W4 (27%), W3 (23%), and W5 (22%). In Table 5.3c, we see that participants tend to use the facial expression in cluster 1 the most during PosNeut (23% of $S_{EC.1}$) and NeutPos (20% of $S_{EC.1}$). In these conditions, participants expected a positive outcome after their interactions with Ria and Tom. This facial expression is used the least during NegPos (6%), NeutNeg (5%), and PosNeg (0%). In these conditions, participants

are less certain about the outcome, as their prior belief conflicts with the contextual evidence. During the NeutNeg condition, participants may have initially assumed a positive outcome during the interaction with Ria. Of these conditions, a minimal amount of $S_{condition}$ are assigned to cluster 1: 3% of S_{NegPos}, 2% of $S_{NeutNeg}$, and 0% of S_{PosNeg} (Table 5.3b). In fact, this is the case for almost every condition except NeutPos (12% of $S_{NeutPos}$) and PosNeut (10% of $S_{PosNeut}$), as expected.

Cluster 5 is one of the bigger elemental clusters. As seen in Table 5.3a, samples in $S_{EC.5}$ occur slightly more in the middle of a question, with 26% in W2, 22% in W4, and 20% in W3 – although the distribution over windows is slightly more equal than that of cluster 1. As seen in Table 5.3a, the facial expression derived from cluster 5 is similarly used to that of cluster 1, although there are differences. 27% of $S_{EC.5}$ occurs during the NeutPos condition, confirming the positive (Table 5.3c). However, in contrast to cluster 1 only 12% of $S_{EC.5}$ occurs during PosNeut. We see that this facial expression is least used during NegNeut (7% of $S_{EC.6}$) and PosNeg (0% of $S_{EC.5}$). The reason this facial expression is not used during PosNeg is likely the same for cluster 1: participants expected a positive outcome, and are confused by the negative contextual evidence. During NegNeut, the negative prior belief paired with neutral contextual evidence likely leads to marking the question with *furrowed* brows together with negation to confirm the negative, as the contextual evidence is inconclusive. Similarly, negation may be used in NeutNeg. In this case, the participant confirms the negation with *raised* eyebrows, as the neutral prior belief paired with negative contextual evidence may have been more surprising or unexpected than in NegNeut. As a general note, it is important to know whether a facial expression is included in combination with negation. This has not been investigated during this research project, but will be a meaningful addition in future work. In NegPos, this expression could mark surprise, more so than the facial expression of cluster 1, due to the lack of EyeSquint and the lower amplitude for MouthFrown.

Another difference between cluster 5 and cluster 1 is seen in Table 5.3b. For most conditions, only a small percentage of $S_{condition}$ are assigned to cluster 1. This is not the case for cluster 5, to which 25% of $S_{NeutPos}$ is assigned. Disregarding the samples that are labeled as noise, this is the highest percentage of $S_{condition}$ assigned to a cluster for any condition. Further, 12% of $S_{Baseline}$, $S_{NeutNeg}$, and $S_{NeutNeut}$, and 8% of S_{NegPos} and $S_{PosNeut}$ are labeled as cluster 5. We see that only 6% of $S_{NegNeut}$ and 0% of S_{PosNeg} are assigned to cluster 5. Note that 53% of $S_{EC.5}$ originates from participant 3, who does not tend to furrow their eyebrows often (see Table 5.1).

Finally, we have seen in Table 5.3b that 0% of S_{PosNeg} is assigned to either cluster 1 or 5. On the other hand, these clusters do occur during NegPos, the other experimental condition where prior belief and contextual evidence contradict. However, this amounts to only 11% of S_{NegPos}.

5.3.3 Furrowed eyebrows

The clusters with facial expressions that align with the 'furrowed eyebrows' definition of 'q' are clusters 2, 3, 6, and 8. First, in Table 5.3a, we see that the facial expression derived from cluster 2 is almost exclusively used at the end of a question; of $S_{EC.2}$, 48% occurs in W5 and 36% in W4. In contrast, only 1% occurs in W1 and 2% in W2. Moreover, Table 5.3c shows that this facial expression is mostly used in those conditions that involve neutral contextual evidence: NeutNeut, NegNeut, and PosNeut, containing 23%, 18%, and 16% of $S_{EC.2}$, respectively. The inconclusive neutral contextual evidence may have prompted participants to use this expression with the additional MouthFrown and MouthShrug markers, to confirm their prior belief.

Table 5.3a shows that most of the samples assigned to cluster 3, $S_{EC.3}$, occur at the beginning and end of the questions: 28% in both W1 and W5, 22% in W2, while only 13% occurs in W4 and 8% in W3. This cluster also displays an interesting phenomenon in Table 5.3c, regarding the distribution of $S_{EC.3}$ over the conditions. 31% of $S_{EC.3}$ occurs in PosNeg, 18% in NegPos, and 13% in NeutPos. This could suggest that this expression happens most after receiving contextual evidence conflicting with prior belief, and then again at the end of repeating the target question to confederate B; trailing the expression they used for most of that question, which was likely from cluster 6 or cluster 8. However, further work on the temporal development is necessary to check whether this is the case.

Table 5.3a shows that samples in cluster 6, $S_{EC.6}$, are spread out over the windows relatively evenly, though they are more concentrated towards the end of the questions: 25% occurs in W5, 23% in W3, and 21% in W4. In Table 5.3c, we see that 24% of $S_{EC.6}$ occurs in the PosNeg condition, 17% in NegPos, 16% in PosNeut, 15% in NeutNeg, 12% in NegNeut, and 10% in NeutNeut. In the first two conditions, contextual evidence contradicts the speaker bias. Further, we see that only a minimal amount of samples in $S_{EC.6}$ occur in the remaining conditions: only 4% in NeutPos and 2% in Baseline.

Table 5.3b shows that of $S_{NeutPos}$, 25% is assigned to cluster 5, and 12% to cluster 1: both clusters in which the eyebrows are raised. However, we also see that 14% of $S_{NeutPos}$ is assigned to cluster 8. For $S_{Baseline}$, we see that a smaller percentage is assigned to cluster 5 (12%) and cluster 1 (8%), whereas 22% is assigned to cluster 8. Although the facial expression derived from this cluster displays

furrowed eyebrows and squinted eyes like cluster 6 does, it does so in a less extreme manner and in combination with slightly widened eyes and frowning mouth. Unlike the strong expectations voiced by the heavily furrowed eyebrows in cluster 6, the facial expression of cluster 8 depicts a weaker speaker bias and more uncertainty. This is likely why more $S_{NeutPos}$ and $S_{Baseline}$ are assigned to cluster 8 than to cluster 6.

For cluster 8, Table 5.3a shows that the samples in $S_{EC.8}$ barely occur at the start of a question (only 6% in W1), but are more or less equally distributed over the remaining windows. Further, we see in Table 5.3c that samples in $S_{EC.8}$ are again spread relatively equally over the conditions, albeit less so in NegPos (9%), NeutPos (9%), and NeutNeut (10%). Perhaps the expression in cluster 8 is generally often used in polar questions, but does not correspond to any specific type of speaker prior belief, contextual evidence, or combination thereof. It is the most frequently occurring elemental cluster in the Baseline condition, and it remains quite frequent in all other conditions (never less than 10%).

Chapter 6

Discussion

In this Chapter, we will discuss the methods, results, and limitations of the present research project. This research project aspired a two-fold contribution: methodological and empirical. First, we aimed to carry out a methodological exploration of the application of Computer Vision technology and Machine Learning techniques in the field of sign language linguistics. Second, in order to explore these methods, we applied them in the specific domain of analysing non-manual markers of polar questions in Sign Language of the Netherlands.

We will first discuss the methodological exploration of RQ1 and RQ2 in Section 6.1, and compare our work specifically to that of Kuznetsova et al. (2021, 2022). Next, in Section 6.2, we discuss the prototypical facial expressions resulting from our clusters. The limitations of the current work and the subsequent avenues for future research are discussed throughout this Chapter.

6.1 Methodologies

6.1.1 Application of Computer Vision technologies

The aim of the first Research Question, RQ1, was to explore a new methodology for how we could use Computer Vision technology to collect data on non-manual markers in sign languages. Related previous work in the field of sign language linguistics have used software such as OpenFace, which extracts information about facial landmarks from 2D videos. We have investigated the application of the *Live Link Face* application, which uses Apple's ARKit and TrueDepth sensor to measure 3D information in the form of standardised ARKit blendshapes directly from the face. We hypothesized that this technology would allow us to measure many facial features to a high degree of precision and accuracy, and that the format of the resulting data would be ideal for further application of ML techniques for data analysis.

This technology had a shallow learning curve, it was relatively straightforward to acquire the necessary knowledge for application. Moreover, the output of Live Link Face was structured and displayed in a comprehensible format. The Live Link Face app has a setting that disables the view of the camera. This was a useful feature for this particular experimental setting, as the screen of the iPhone had to be on and face the participant while recording, which could be a big distraction otherwise.

The position of the depth camera was carefully chosen to be in between the participant and the confederate (although slightly to the right side of the participant). However, we had not considered the fact that most participants would be right handed, and that this position of the camera was therefore prone to occlusions of the face. Placing the camera on the participant's non-dominant side would have provided the most unobstructed view. This is thus an important consideration for the application of this technology for the collection of non-manual markers in sign languages. Nevertheless, ARKit handled occlusions of the face well. In most cases where this occurred, blendshape measurements were not taken for circa 10-20 frames. As we recorded 60 frames per second, this amounted to missing only a sixth or a third of a second. The missing values could therefore easily be interpolated.

A limitation of this method is that, while ARKit measures a total of 61 blendshapes, not all of the recorded facial features may be relevant for the purpose of collecting data on non-manual markers in sign languages. Therefore, careful considerations must be made regarding the inclusion of each recorded feature, as some may add noise to the data set. Some of these features are potentially linguistically relevant, but could not be considered due to the noise produced by the mouthings. Furthermore, a limitation of this technology is that not all of the information that may be of interest for a linguist studying non-manual markers are recorded with this specific technology. For instance, a forward movement of the entire head or body is not measured. This must thus be supplemented with either manual

annotations or body tracking technology, which can be explored in future work.

Another limitation is that the blendshapes can be subject- and location-sensitive. Live Link Face offers the option to calibrate blendshape measurements to a person's neutral facial expression. This is a useful feature, but not an all encompassing solution. The measurements of some blendshapes, specifically those regarding the rotation of the face and eyes, are influenced by the subject's position relative to the camera. For the most accurate measurements of these blendshapes, the calibration must thus be carried out if the subject moves to a different location. Restricting the location of the subject by using a chair would therefore be a useful addition to a similar experiment. Another solution could be to attach the camera to the subjects body to that the relative position of the camera to the subject's face is always the same. However, this might feel unnatural, which could influence the data. These are all avenues to explore in future work.

Despite the calibration, the measurements of the features were not necessarily in the same range between participants. ARKit measures the blendshapes on a scale of 0-1, but not all participants reached a measurement of 1 for some features. For instance, all participants (at some point in the experiment) had fully raised their eyebrows. Therefore, the output data still needed some minor modifications such as normalisation. However, this is very straightforward.

Further pre-processing of the output data was relatively simple, but still took a considerable amount of time. However, this can be mostly attributed to having to align the recordings of the main camera and the depth camera. With more careful preparation – such as linking both camera's to the same time code – this process would have taken significantly less time. Moreover, the design of the experiment required recordings to be started at the beginning of each trial, meaning that each video and CSV file needed to be trimmed to the target question.

A final caveat of this method is that the data has not yet been validated with manual annotations, which is currently a limitation. Initial testing of output data for specific features suggested that it is certainly accurate, but in future work it will be interesting to investigate the extent to which the blendshape measurements align with manual annotations.

6.1.2 Application of Machine Learning techniques

The aim of the second research question, RQ2, was to carry out a methodological exploration of Machine Learning techniques to analyse the data gathered through CV technology. We chose to apply these techniques in the domain of polar questions in Sign Language of the Netherlands, specifically to investigate the variations between prototypical facial expressions that are used for marking such questions. For that reason, we applied *clustering*, which is a technique that attempts to differentiate between groups (or 'clusters') of samples in the data. This machine learning technique is typically applied to unlabeled data in order to infer new information about it, which applied exactly to this research project. The data was unlabeled, meaning that we did not know beforehand what facial expressions resulted from specific combinations of feature values; and our aim was to infer new information about the data, specifically what 'groups' of facial expressions could be found in it.

First and foremost, we found that in order to apply ML techniques to the 3D data, it is essential to have a fundamental understanding of the data set. The process of applying various clustering algorithms and attempting to optimize their outputs lead to an increased understanding of the structure and characteristics of the data set. This was then used to make informed and targeted decisions about which samples to present to the algorithms and in what way they should be structured. For instance, we found that it was necessary to first select the samples relevant to our research question. Otherwise, the clustering algorithms would naturally take into account samples that were not of interest, such as those in which features transition from one facial expression to another.

Moreover, clusters are formed by samples that are similar to each other in terms of feature values. Initially, clusters predominantly formed on the basis of immediately neighbouring samples in the data set, as these are mostly very alike. Although downsampling the data set did have some effect, we found that samples were still too subject-specific. The continuous measurements of feature values were taken to such a specific degree of precision that clusters naturally formed around facial expressions that were personal rather than facial expressions that were linguistically relevant. Therefore, it was necessary to decrease the level of detail in the data set by transforming the values to discrete rather than continuous. While this did decrease the extent of the advantage of detailed knowledge about the amplitude of features, it did not eradicate it completely. The categorised feature values still provided more specific information about the non-manual markers, while allowing more inclusive and comprehensive clusters to be formed.

Another important aspect of this process was studying different ML techniques and the effect that certain changes to their parameters would have on the outcomes, and learning why HDBScan was a more appropriate algorithm for this data set than the more commonly applied K-means. First, HDBScan formed clusters on the basis of density, and

clusters could take the form of various shapes, sizes, and densities. Further, HDBScan labels samples that do not distinctly fall into a cluster as noise, which leads to clusters that are more clearly defined. K-means, on the other hand, essentially partitions the data. It allocates every sample in the data set to a cluster, even though it might not necessarily belong to that (or any) cluster. Despite our efforts to discard samples containing noise, they were still present in the data set. Allocating these samples to the clusters could therefore influence the facial expressions that these clusters represent. For these reasons, HDBScan was more suitable for our purpose and data set than K-means.

We found that there is not one superior configuration of the algorithm's parameters for our data set and purpose. Adjustments in the parameters lead to variations in the clusters that were found by HDBScan. Although there was not only one 'correct solution', there were certainly clusters that do not allow us to generalise over the results. These clusters typically represented a very small portion of the data set (between 1-3%), and often consisted predominantly (or entirely) of samples originating from one single participant. We therefore limited our discussion of the results to two implementations (one comprising more clusters than the other) of which we believed they accurately portrayed the clusters found by HDBScan.

One limitation of our method for ML is in the way frames were selected. In an attempt to extract only relevant and thus more constant facial expressions, we discarded frames in which feature values fluctuated. However, the measurements of the depth camera were quite sensitive. This means that although a facial expression could appear (relatively) constant to a human perceiving the expression, micro fluctuations in the expression may have caused these samples to be discarded.

Another limitation of this method is that the data set was not balanced with relation to the amount of samples for each participant, condition, and the temporal windows. This likely has a significant effect on the results, which has not been controlled for. Future work could include the creation of a more balanced data set and an analysis of how this affects the results.

6.1.3 Comparison with related work

This research project was partly inspired by the work of Kuznetsova et al. (2021, 2022). We aimed to explore new methodologies for the computational analysis of non-manual markers in sign languages. In this Section, we will compare the methods of this research project to those in Kuznetsova et al. (2021, 2022).

First, the data gathered in the previous research was simultaneously less structured and more structured than our data. The format of their data set was less structured in terms of controlling for context, considering all types of polar questions to belong to one class, and all wh-questions to belong to another. They did not consider the variation that might take place within these classes. In contrast, we focused on only one class, polar questions, and controlled specifically for variations in context within this class.

On the other hand, their data was more structured in the sense that participants all produced the same signs in the same order, for three different forms (polar questions, wh-questions, and statements). In comparison, our data was less structured in this regard, as participants were given more freedom to decide which signs to use and in what order. Whereas previous researchers were able to normalise, split, and analyse the data on specific syntactical elements, we did not have this ability. Since the sign order and number of signs varied, specific syntactical scopes of NMMs may become invisible. This makes it more difficult to determine whether a feature serves a linguistic function or not. On the other hand, it is likely that our data is more naturalistic and representative of real world situations.

Related to this point is the method of data elicitation. Our data was elicited through a more natural and open role-play type conversations, using the same instructions and experimental stimuli for each participant. The data collection in our study involved prompting participants through videos in NGT, role-play conversations, and picture prompts. Each aspect was carefully considered, as to not influence participants on the types of signs they used and the order in which they signed them. On the other hand, the data collected in Kuznetsova et al. (2021) was elicited in two different manners. The hearing signers were presented with the stimuli in written Russian, while the deaf signers were presented with the stimuli as video recordings in Kazakh-Russian Sign Language. These videos could certainly have influenced the signs produced by their participants.

Further, in Kuznetsova et al. (2022), a significant difference was found in how hearing and deaf signers used NMMs. In the present study, only deaf signers participated.

The methods between these studies also varied in how quantitative data was obtained. Kuznetsova et al. (2021, 2022) used OpenFace software to extract landmark information from 2D videos, while the present study used a 3D depth camera. In general, the depth camera provides information about more facial features than OpenFace. Whereas only the position of the eyebrows and head tilting were examined in Kuznetsova et al. (2021, 2022), the present study investigated a broader range of features, including the brow position, eye form, mouth, and the cheek and nose.

Both of these methods had their own limitations.

Kuznetsova et al. (2021) found that OpenFace produced bias results in the position of the eyebrows when the head was tilted, and had to account for this bias before examining the data. While our 3D camera did not see such bias in the measurements of features like the eyebrows, it was sensitive to the position of the subject in relation to the camera, rendering the measurements of features regarding head tilting and rotation inapplicable. We have not investigated the possibility of correcting these measurements, but the problem in this case is more complex than in the case of Kuznetsova et al. (2021). In their case, the bias presented in the more structured condition of head tilting, which is measurable and annotated. In our case, the position of the participant relative to the camera was not measured nor annotated.

Finally, the studies applied different methods for analysing the data. Kuznetsova et al. (2022) used functional data analysis, a technique which allows the modeling of the continuous movement of non-manual markers from sequential temporal data. In contrast, we used clustering to investigate the prototypical facial expressions for question marking. Our technique allowed for a rougher exploration of temporal information regarding these facial expressions, instead of information regarding their continuous movement. This enabled us to explore the specific variations between facial expressions and their potential contextual and/or temporal constraints. However, our method of modeling temporal information potentially misses certain trends in the data by examining a rougher abstraction of temporal information. While this method for data analysis fit the purpose of the present study, future work may include a more detailed exploration of the temporal development of these facial expressions.

6.2 Prototypical facial expressions

We now turn to the discussion of RQ3, of which the aim was to find out how different types of biased polar questions are marked in Sign Language of the Netherlands. Sub-questions of this research question were a) what results we would expect to see on the basis of a literature review, b) what the prototypical facial expressions are for marking polar questions (and thus what are variations of these expressions), c) how variations of these expressions map to the contexts in which they are used, and d) what the temporal progression of non-manual markers are during biased polar questions. This was first investigated through a literature review, after which we aimed to find our own results through the application of clustering on our data set.

On the basis of the literature review, we expected to find at least two facial expressions. The first – and, according to the literature, also the most used – facial expression would comprise raised eyebrows and (likely) eyes wide open; while the second would consist of at least furrowed eyebrows and (likely) squinted eyes. However, we observed additional non-manual markers were used during the experiment. Therefore, we hypothesized that we would find other variations *within* these expressions as described in the literature, such as expressions with the addition of the mouth in the form of a frown or a shrug, or with cheeks squinted. Further, we expected to find contextual and/or temporal constraints in which these facial expressions could be applicable.

Indeed, all of our clustering implementations lead to two main (super-)clusters, which mostly aligned with the expectations obtained by the literature. However, we will focus our discussion on the clusters resulting from two final implementations of HDBScan.

6.2.1 General results

We first turn to the overarching results. In our final implementation, HDBScan found not two, but three super-clusters (not including the 15% of samples labeled as 'noise'). The first super-cluster expressed raised inner and outer eyebrows, along with wide-opened eyes. Further, samples in this cluster also displayed a modest mouth frown, and to a lesser degree a slight eye squint (for an example of the simultaneous occurrence of wide eyes and squinted eyes, see Appendix B.1). However, this was certainly not the most frequently occurring facial expression for question marking during our experiment. The samples in this cluster only represent 19% of the data set. This result therefore does not completely align with the expectations based on the literature.

The second super-cluster comprised a more neutral facial expression, where barely any features were visibly expressed. We note that it is likely that the samples belonging to this cluster are not characteristic for question marking. This cluster could be a side effect from the categorisation of feature values, containing those samples that mostly comprise low feature values. Further, it could contain samples that are generally not truly 'active', but were missed by our attempts to feature these samples out.

The third super-cluster (and thus the second prototypical facial expression) displayed furrowed eyebrows and squinted eyes, and a slight cheek squint and nose sneer. This facial expression occurred the most frequently, as the samples assigned to this cluster represent 39% of the data set.

Therefore, our general results mostly align with the expectations obtained by most definitions of 'q' in the literature, as the two most characteristic facial expressions during question marking included one with raised eyebrows and one with fur-

rowed eyebrows. However, the frequency at which these facial expressions occurred does not align with most definitions of 'q' from the literature, as the facial expression with furrowed eyebrows occurred almost twice as much as the facial expression with raised eyebrows. This does, however, support the "phonetic strength" that was attributed to 'furrowed eyebrows' (i.e. BROWDOWN or AU 4) in de Vos et al. (2009); which stated that this marker was stronger than 'raised eyebrows' (i.e. BROWINNERUP + BROWOUTERUP or AU 1 + 2).

Finally, although 'wide eyes' and 'squinted eyes' did not appear in previous definitions of 'q' for NGT as found in the literature, our results certainly support the premise that they are indeed defining components of prototypical facial expressions during polar question marking in NGT.

6.2.2 Variations within expressions and their contextual and temporal constraints

In the second final implementation of HDBScan, we found that the three super-clusters comprised eight smaller elemental clusters. We turn to these clusters for the variations in prototypical facial expressions, how they relate to their contexts, and how their relate to their temporal position in a question.

In essence, these elemental clusters still belonged to the overarching facial expressions either displayed raised eyebrows, furrowed eyebrows, or a neutral expression. However, we found variations of these expressions where, for instance, one facial expression included additional features such as a mouth frown, whereas another version of that facial expression did not. There were more variations between the facial expressions with furrowed eyebrows than between the facial expressions with raised eyebrows.

When relating the facial expressions of the elemental clusters to their respective contexts and temporal windows, we expected to find clear constraints on the applicability of these facial expressions. However, from the results so far we can only conclude two definite constraints. The first is that facial expressions with raised eyebrows (found in super-cluster 1 and elemental clusters 1 and 5) do *not* occur in contexts in which a person holds a positive prior belief and is subsequently presented with negative contextual evidence. These facial expressions do occur in conditions in which a person holds a negative prior belief and is presented with positive contextual evidence, but only to a small extent. The second constraint is that facial expressions involving significantly frowning mouth (found in elemental cluster 2) do *not* occur in the beginning of a question, but only in the temporal windows W3-W5.

However, we may find further constraints if we cross contexts with temporal windows, or if we take the presence of negation into account. For instance, it may be that a certain type of expression never occurs in W1-W3 in the NeutNeg condition, or it may be that a certain type of expression only occurs in the PosNeg condition if it co-occurs with negation. In principle, we could also cross contexts with grammatical role (subject, object, verb) if we integrate the blendshape data with the manual annotations. These considerations were not explored during the present research project, but provide interesting avenues for future research.

Finally, we included some features in this research that were not known to be linguistically relevant. For two of them, MOUTHSHRUG and MOUTHFROWN, we had reason to believe they might be characteristic for facial expressions in question marking. It is possible that CHEEKSQUINT and NOSESNEER are more side-effects from EYESQUINT, and are thus not necessarily linguistically relevant. It could well be that these features are just general human expression, but it is difficult to determine the exact category in which these features fall. However, as it is possible to have squinted eyes without these additional features, we decided to include them in our research. Nevertheless, future research should also explore this data set without these last two features and investigate how the resulting clusters compare. Additionally, this could then be repeated while limiting the data set entirely to features measuring the eyebrows and the eyes, as these are known to be linguistically relevant. Clusters would then likely distinguish between more detailed variations within these features. This might also lead to more distinct constraints in which these facial expressions can be used.

Bibliography

Apple (2022). ARKit blendShapes - Apple Developer Documentation. `https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation`. Accessed: 13/12/2022.

Baker, A., van den Bogaerde, B., Pfau, R., and Schermer, T. (2016). *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66.

Benitez-Quiroz, C. F., Gökgöz, K., Wilbur, R. B., and Martinez, A. M. (2014). Discriminant features and temporal structure of nonmanuals in american sign language. *PloS one*, 9(2):e86268.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Cecchetto, C. (2012). 14. sentence types. In *Sign Language*, pages 292–315. De Gruyter Mouton.

Chen, L., Wu, Z., Ling, J., Li, R., Tan, X., and Zhao, S. (2022). Transformer-s2a: Robust and efficient speech-to-animation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7247–7251. IEEE.

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

Crasborn, O. (2006). Nonmanual structures in sign language. In *Encyclopedia of Language & Linguistics (Second Edition)*, pages 668–672. Elsevier, Oxford, second edition edition.

Dardas, N. H. and Georganas, N. D. (2011). Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Transactions on Instrumentation and measurement*, 60(11):3592–3607.

de Coster, M. and Dambre, J. (2022). Leveraging frozen pretrained written language models for neural sign language translation. *Information*, 13(5):220.

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.

Domaneschi, F., Romero, M., and Braun, B. (2017). Bias in polar questions: Evidence from english and german production experiments. *Glossa: a journal of general linguistics*, 2(1).

Ehret, J., Bönsch, A., Aspöck, L., Röhr, C. T., Baumann, S., Grice, M., Fels, J., and Kuhlen, T. W. (2021). Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception (TAP)*, 18(4):1–15.

Ekman, P., Friesen, W., and Hager, J. (2002). Facial action coding system (facs). *A Human Face, Salt Lake City*.

Epic Games (2022a). METAHUMAN - High-fidelity digital humans made easy. `https://www.unrealengine.com/en-US/metahuman`. Accessed: 11/12/2022.

Epic Games (2022b). Recording facial animation from an ios device. https://docs.unrealengine.com/5.0/en-US/recording-face-animation-on-ios-device-in-unreal-engine/. Accessed: 11/12/2022.

Esselink, L., Oomen, M., and Roelofsen, F. (2022a). Biased polar questions in sign language of the netherlands - 3d data set. https://uvaauas.figshare.com/articles/dataset/Biased_polar_questions_in_Sign_Language_of_the_Netherlands_-_3D_data_set/21746216/1.

Esselink, L., Roelofsen, F., Dotlačil, J., Mende-Gillings, S., de Meulder, M., Sijm, N., and Smeijers, A. (2022b). Exploring automatic text-to-sign translation in a healthcare setting. Manuscript under review at Universal Access in the Information Society.

European Union of the Deaf (2022). Netherlands. https://www.eud.eu/member-countries/netherlands/. Accessed: 25/11/2022.

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):e0233731.

Klomp, U. (2021). *A descriptive grammar of Sign Language of the Netherlands*. LOT.

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 LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*. European Language Resources Association (ELRA).

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.

Loukas, S. (2020). K-means clustering: How it works & finding the optimum number of clusters in the data. https://towardsdatascience.com/k-means-clustering-how-it-works-finding-the-optimum-number-of-clusters-in-the-data-13d18739255c.

Luo, L., Weng, D., Songrui, G., Hao, J., and Tu, Z. (2022). Avatar interpreter: Improving classroom experiences for deaf and hard-of-hearing people based on augmented reality. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

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.

Miyawaki, R., Perusquia-Hernandez, M., Isoyama, N., Uchiyama, H., and Kiyokawa, K. (2022). A data collection protocol, tool and analysis for the mapping of speech volume to avatar facial animation.

Oomen, M. and Roelofsen, F. (2022a). Biased polar questions in Sign Language of the Netherlands - Methods description. https://uvaauas.figshare.com/articles/preprint/Biased_polar_questions_in_Sign_Language_of_the_Netherlands_-_Methods_description/21701954.

Oomen, M. and Roelofsen, F. (2022b). Biased polar questions in Sign Language of the Netherlands - Stimuli context videos. https://uvaauas.figshare.com/articles/media/Biased_polar_questions_in_Sign_Language_of_the_Netherlands_-_Stimuli_context_videos/21695150.

Oomen, M. and Roelofsen, F. (2022c). Biased polar questions in Sign Language of the Netherlands: Video data. https://uvaauas.figshare.com/articles/media/Biased_polar_questions_in_Sign_Language_of_the_Netherlands_Video_data/21666203.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pfau, R. and Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. *Cambridge language surveys*.

Roelofsen, F., Esselink, L., Mende-Gillings, S., Meulder, M. d., Sijm, N., and Smeijers, A. (2021a). Online evaluation of text-to-sign translation by deaf end users: Some methodological recommendations. In *First International Workshop on Automatic Translation for Sign and Spoken Languages (AT4SSL)*, pages 82–87.

Roelofsen, F., Esselink, L., Mende-Gillings, S., and Smeijers, A. (2021b). Sign language translation in a healthcare setting. In *Translation and Interpreting Technology (TRITON)*, pages 110–124.

Schmitt, D. and McCoy, N. (2011). Object classification and localization using surf descriptors. *CS*, 229:1–5.

Volksgezondheid en Zorg (2022). Gehoorstoornissen. `https://www.vzinfo.nl/gehoorstoornissen`. Accessed: 25/11/2022.

World Health Organization (2021). World report on hearing. `https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/highlighting-priorities-for-ear-and-hearing-care`.

Zeshan, U. (2004). Interrogative constructions in signed languages: Crosslinguistic perspectives. *Language*, pages 7–39.

Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., and Yang, J. (2011). A framework for hand gesture recognition based on accelerometer and emg sensors. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 41(6):1064–1076.

Zhao, S. (2020). K-means clustering clearly explained. `https://medium.com/@luo9137/k-means-clustering-clearly-explained-44746ccc3621`.

Appendices

Appendix A

Data set design, collection, and preparation

A.1 Situations

For ease of the reader, the information in this Section of the Appendix is taken directly from Appendix A of Oomen and Roelofsen (2022a). The entirety of this Section is therefore not personally written, but courtesy of Oomen and Roelofsen (2022a).

We created five situations and one practice situation. Video recordings of the contexts (two per situation) in NGT, as they were also shown to the participants, are available at `https://doi.org/10.21942/uva.21695150`. Confederate responses provide positive ('+'), neutral ('0'), or negative ('−') evidence for the target question (final participant utterance).

A.1.1 Practice situation: Is there a metro station nearby?

1. ***Original speaker bias***

Context 1: You recently moved to the center of Amsterdam. You would like to take the metro to Artis [zoo in Amsterdam]. You don't know if there's a metro station nearby. You meet Ria, who lives close to Artis. Ask her.

Participant: "Is there a metro station nearby Artis?"

Confederate A:
+ "Yes, there is a metro station close to Artis."
0 "I don't know, I never take the metro."
– "No, there's no metro station near Artis."

2. ***Contextual evidence***

Context 2: You're meeting your new neighbor Tom for the first time. Ask him whether he knows the way to Artis.

Participant: "Do you know the way to Artis?"

Confederate B:
+ "There's a metro station here around the corner. You should take line 51 to Weesperplein, which is close to Artis."
0 "It's best to go by public transport."
– "You can't take the metro, because there's no metro station near Artis. You should take tram 17."

3. ***Target question***

Picture prompt:

Participant: Variation on "Is there a metro station nearby?"

A.1.2 Situation 1: Is Kim a vegetarian?

1. *Original speaker bias*

Context 1: You're organizing a dinner. You've also invited Kim, but you don't know if Kim is a vegetarian. Ria knows Kim well. Ask her.

Participant: "Is Kim a vegetarian?"

Confederate A:
+ "Yes, Kim is a vegetarian."
0 "I don't know if Kim is a vegetarian."
– "No, Kim is not a vegetarian."

2. *Contextual evidence*

Context 2: You and Tom are cooking dinner together. You're making meatballs. Ask Tom how many meatballs you should make.

Participant: "How many meatballs should we make?"

Confederate B:
+ "You don't have to make any for Kim, she is a vegetarian"
0 "Let's make two for everyone, except for the vegetarians."
– "We should definitely make enough for Kim, she loves them!"

3. *Target question*

Picture prompt:

(Version 1)

(Version 2)

Participant: Variation on "Is Kim a vegetarian?"

A.1.3 Situation 2: Is the park open?

1. ***Original speaker bias***

Context 1: You want to go to the Efteling [Dutch theme park] this weekend, but you're not sure it's open. You meet Ria, who has a subscription to the park. Ask her.

Participant: "Is the Efteling open this weekend?"

Confederate A:
+ "Yes, the Efteling is open this weekend."
0 "It's open on Saturday but I don't know about Sunday. I never go on Sunday."
– "It's open on Saturday but I think I read in the newspaper that it's not open on Sunday."

2. ***Contextual evidence***

Context 2: Later that day, you meet Tom. He works at the Efteling. You know he has the weekend off. Ask him if he'd like to come to the Efteling with you this weekend.

Participant: "Do you want to go to the Efteling with me?"

Confederate B:
+ "Fun! Shall we go on Sunday?"
0 "I can't this weekend."
– "The Efteling is only open on Saturday. I'm available then."

3. ***Target question***

Picture prompt:

(Version 1)

(Version 2)

Participant: Variation on "Is the Efteling open this weekend?"

A.1.4 Situation 3: Is entrance free of charge?

1. ***Original speaker bias***

Context 1: You would like to visit the Veluwe [Dutch national park] tomorrow. You don't know if entrance is free of charge. Ria is a volunteer at the park. Ask her.

Participant: "Is entrance to the Veluwe free of charge?"

Confederate A:
+ "Yes, you don't have to pay a fee."
0 "I don't know."
– "No, a ticket costs 10 euros."

2. ***Contextual evidence***

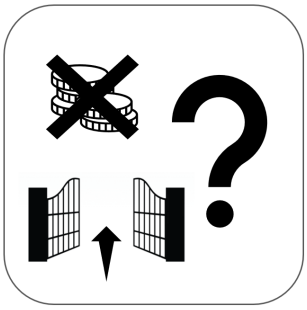
Context 2: A day later, you're at the Veluwe parking lot. You can't find the entrance to the park. At the parking lot, you meet Tom, another visitor to the park. Ask him.

Participant: "Do you know where the entrance is?"

Confederate B:
+ "The entrance is there by the white flag. You don't need a ticket."
0 "The entrance is there by the white flag."
– "The entrance is there by the white flag, but you need to get a ticket at the ticket counter over there first."

3. ***Target question***

Picture prompt:



(Version 1)

(Version 2)

Participant: Variation on "Is entrance free of charge?"

A.1.5 Situation 4: Is Kim home?

1. *Original speaker bias*

Context 1: You're a student and you're living together with Ria, Tom, and Kim. You're planning to visit your parents this weekend. You know that Ria and Tom will also be away. You don't know if Kim will stay at home. Ask Ria.

Participant: "Will Kim stay at home?"

Confederate A:
+ "Yes, she needs to study all weekend."
0 "I don't know if she'll stay at home."
– "I thought Kim said she going to spend a weekend at sea."

2. *Contextual evidence*

Context 2: On Saturday morning, you unexpectedly have to return home early, but you forgot your keys. On the way home, you call Tom; you can't get a hold of Kim. Ask Tom if Kim could open the door for you.

Participant: "Can Kim open the door for me?"

Confederate B:
+ "Yes, I just talked to her and she's there."
0 "I don't know. You should send her a text."
– "Kim is away for the weekend."

3. *Target question*

Picture prompt:

(Version 1)

(Version 2)

Participant: Variation on "Is Kim home?"

A.1.6 Situation 5: Is there a train at 9am?

1. *Original speaker bias*

Context 1: Tomorrow morning, you'd like to take the train from Amsterdam to Paris. You'd prefer to leave at 9am. But you don't know if there's a train at 9. Ria has a public transportation travel planner app on her phone. Ask her.

Participant: "Is there a train from Amsterdam to Paris at 9am tomorrow?"

Confederate A:
+ "Let me check. Yes, there's a train at 9am"
0 "Oh, the app doesn't work, so I don't know."
– "Let me check the app. No, I don't see a train at 9am."

2. *Contextual evidence*

Context 2: You live close to the train station, so you decide to walk to the ticket counter to buy a ticket. Ask the ticket seller how much a ticket costs for the train to Paris tomorrow.

Participant: "How much does a ticket for the train to Paris tomorrow cost?"

Confederate B:
+ "For the 9 o'clock train, a ticket costs 100 euros."
0 "It depends on what time you'd like to leave. There are multiple trains going tomorrow."
– "There's only one train tomorrow, which leaves at 10am. A ticket costs 100 euros."

3. *Target question*

Picture prompt:

(Version 1)

(Version 2)

Participant: Variation on "Is there a train at 9am?"

Appendix B

Prototypical facial expressions

B.1 Simultaneous features

Figure B.1: Simultaneous occurrence of of EyeWide and EyeSquint

B.2 HDBScan clusters

B.2.1 Original and downsampled data sets

I	C	S (%)	S distribution (%)				Most engaged features (median value)								
			P3	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
	1	1	–	–	–	100	**42**	**80**	–	**75**	**42**	**44**	**67**	18	–
	2	1	–	100	–	–	**49**	–	**58**	5	–	30	**53**	21	9
	3	2	–	100	–	–	**54**	–	**64**	3	–	**51**	**40**	23	9
1a	4	2	25	29	–	46	28	23	–	3	–	30	**74**	9	–
	5	13	59	27	8	5	6	**82**	–	**77**	**70**	–	21	–	–
	6	53	11	33	33	23	23	3	**44**	2	–	3	4	12	9
	N	29													
	1	12	61	28	5	6	6	**82**	–	**77**	**70**	–	20	–	–
1b	2	56	10	36	32	22	25	2	**46**	2	–	3	5	12	9
	N	32													

(a) Original dataset

I	C	S (%)	S distribution (%)				Most engaged features (median value)								
			P3	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
	1	2	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	3	–	100	–	–	**51**	5	**60**	2	–	**45**	**46**	21	9
	3	2	38	4	39	19	1	3	–	**81**	**76**	–	18	5	3
2a	4	2	23	21	–	56	24	16	–	5	–	23	**73**	8	1
	5	14	52	23	16	9	6	**80**	–	**76**	**69**	–	21	–	–
	6	55	12	37	31	20	26	–	**46**	–	–	–	–	13	9
	N	22													
	1	16	44	20	10	26	7	**82**	–	**75**	**66**	–	25	2	–
2b	2	67	13	35	30	22	26	7	**42**	3	–	4	8	13	9
	N	17													

(b) Downsampled dataset

Table B.1: Clusters formed by HDBScan for the original dataset (a) and the downsampled dataset (b). Features are shown as the cluster's median value for that feature; values above 40 are marked in bold.

B.2.2 Categorised data sets

I	C	S (%)	S distribution (%)				Most engaged features (median value)								
			P3	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B10a	1	3	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	2	–	99	1	–	**49**	–	**58**	5	–	31	**54**	22	11
	3	2	20	25	–	55	24	25	–	4	–	32	**74**	9	1
	4	15	54	25	11	9	6	**81**	–	**77**	**69**	–	21	–	–
	5	62	13	33	32	22	24	5	**41**	2	–	3	5	12	9
	N	17													
B10b	1	16	51	24	11	14	6	**81**	–	**76**	**68**	–	21	–	–
	2	68	14	33	31	22	24	8	37	3	–	3	7	12	8
	N	17													

(a) B10

I	C	S (%)	S distribution (%)				Most engaged features (median value)								
			P3	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B6a	1	3	–	–	–	100	34	**82**	–	**72**	**43**	36	**66**	16	–
	2	1	–	99	1	–	**48**	–	**61**	3	–	34	**47**	21	12
	3	3	13	43	6	38	23	25	–	3	–	26	**74**	8	2
	4	15	53	24	10	13	6	**81**	–	**76**	**69**	–	21	–	–
	5	57	12	33	34	21	24	6	**42**	2	–	3	5	12	9
	N	21													
B6b	1	20	44	21	8	27	9	**81**	–	**73**	**66**	–	24	1	–
	2	69	14	36	30	21	27	7	42	2	–	4	7	13	9
	N	11													

(b) B6

I	C	S (%)	S distribution (%)				Most engaged features (median value)								
			P3	*P5*	*P6*	*P7*	*ES*	*EW*	*BD*	*BIU*	*BOU*	*MS*	*MF*	*CS*	*NS*
B4a	1	6	45	11	3	41	26	**82**	–	**74**	**64**	4	39	–	–
	2	7	5	71	1	23	**45**	4	**49**	4	–	33	**62**	22	7
	3	7	9	28	17	46	13	2	**44**	2	–	2	2	7	6
	4	7	35	12	20	33	10	37	–	4	–	–	12	4	–
	5	10	53	30	11	6	3	**81**	–	**77**	**69**	–	16	–	–
	6	13	1	46	40	14	**41**	–	**70**	3	–	5	1	17	13
	7	14	18	29	23	30	5	11	9	2	–	2	4	3	4
	8	15	6	29	56	10	**42**	11	**49**	3	–	3	8	19	12
	N	22													
B4b	1	18	47	20	7	26	11	**82**	–	**75**	**67**	–	25	1	–
	2	27	24	21	22	32	9	20	–	3	–	2	15	4	3
	3	39	8	38	36	18	39	–	**60**	2	–	4	2	17	12
	N	15													

(c) B4

Table B.2: Clusters formed by HDBScan for B10 (a), B6 (b), and B4 (c). The most engaged features are shown as the cluster's median value; values above 40 are marked in bold.

Appendix C

Results

C.1 Cluster samples per window

		Cluster								
		1	2	3	4	5	6	7	8	N
	1	4	–	12	8	8	9	38	6	15
	2	6	1	9	7	15	14	12	18	18
S_W	3	9	5	4	4	12	18	8	21	19
	4	7	11	5	6	10	12	6	18	25
	5	5	12	7	9	6	11	9	14	26

Table C.1: Relation between clusters and windows. For each window (W), how S_W is distributed over clusters (a).

C.2 Visualisations of facial expressions

(a) Cluster 4

(b) Cluster 4

(c) Cluster 7

(d) Cluster 7

Figure C.1: Visualisation of more neutral facial expressions (do not align with previous definitions of 'q').

(a) Cluster 1

(b) Cluster 1

(c) Cluster 5

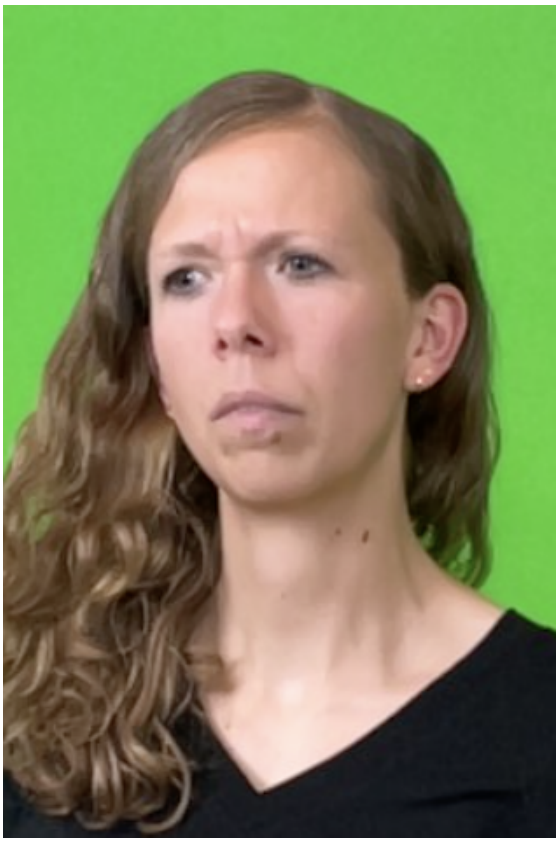
(d) Cluster 5

Figure C.2: Visualisation of facial expressions with raised eyebrows.

(a) Cluster 2

(b) Cluster 2

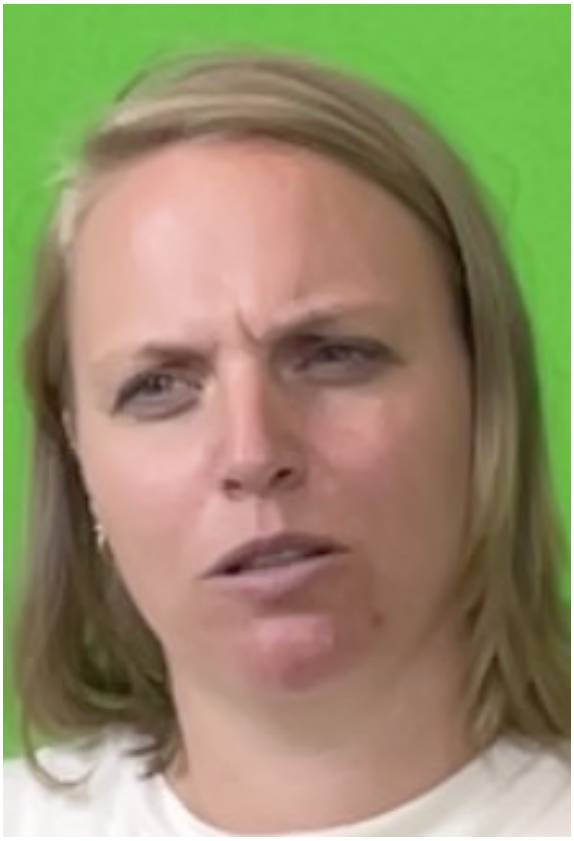
(c) Cluster 3



(d) Cluster 3

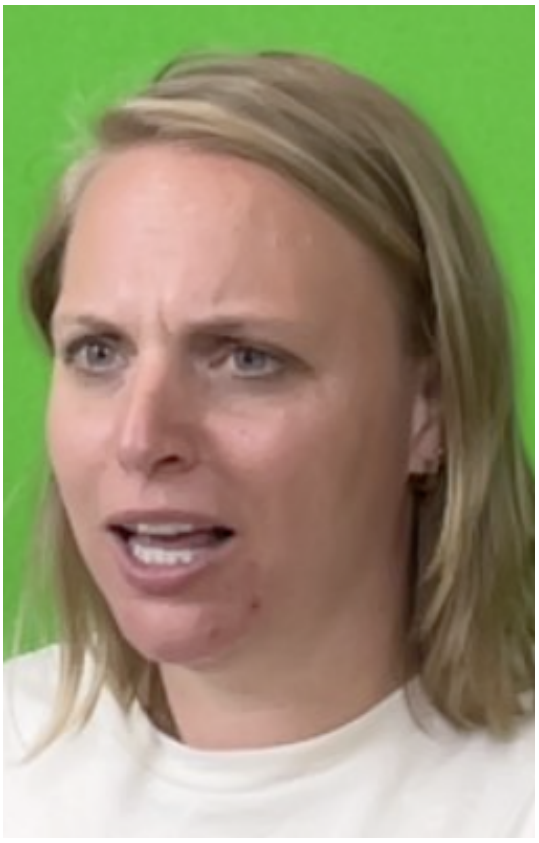
Figure C.3: Visualisation of facial expressions with furrowed eyebrows.

(a) Cluster 6



(b) Cluster 6

(c) Cluster 8



(d) Cluster 8

Figure C.4: Visualisation of facial expressions with furrowed eyebrows