An Analysis of Visual and Morphosyntactic Cues in Biased Polar Questions in Dutch

MSc Thesis *(Afstudeerscriptie)*

written by

Elynn Louise Weijland
(born December 28th, 2001 in Amersfoort, the Netherlands)

under the supervision of **Prof Dr Floris Roelofsen** and **Dr Marloes Oomen**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
26th of February, 2024	Dr Balder ten Cate Dr Sonia Ramotowska Dr James Trujillo

Institute for Logic, Language and Computation

Abstract

This thesis project analysed the visual cues that mark different types of polar questions in Dutch. From previous multimodal studies, we know that visual cues, such as eyebrow movements and head tilts, may accompany spoken language questions [Nota et al., 2021, 2023, Zygis et al., 2017, da Silva Miranda et al., 2020, Miranda et al., 2021]. Similarly, preceding research has shown that visual cues play a significant role in sign languages [Baker et al., 2016]. Such cues, also referred to as non-manual markers (NMM), can play a role at the morphological, phonological, syntactic and pragmatic level [Pfau and Quer, 2010]. A recent development in sign language research concerns the quantitative studies of NMM in polar questions. Particularly, Esselink [2022] examined the NMM that mark *biased* polar questions in Sign Language of the Netherlands (NGT). It was reported that different types of polar question in NGT are marked with different combinations of NMM. For instance, these involve frowned eyebrows, in combination with squinted eyes, or raised eyebrows and widened eyes, which do not occur in contexts in which positive prior belief is contradicted with negative contextual evidence.

Spoken languages are multimodal: auditory cues, involving speech, are combined with visual cues to relay communicative intention. It is well-known that in Dutch, prosodic patterns and differences in word order mark polar questions [Englert, 2010, Borràs-Comes et al., 2014, Gaasbeek, 2023]. However, the manner in which spoken Dutch employs visual cues as a means to mark polar questions is relatively understudied.

The current project had three objectives: identifying the question structures used most frequently in different types of Dutch biased polar questions, obtaining the most prototypical facial expressions marking these different types of questions, and comparing these results to those found in NGT [Esselink, 2022].

The data that was obtained for this project was elicited by means of an experiment in which native Dutch speakers interacted with *confederates* in small role-plays. This experimental design closely followed the design by [Oomen and Roelofsen, 2023a, Esselink, 2022], in order to be able to compare results between NGT and Dutch.

During the experiment, participants were recorded by three cameras, as well as one 3D depth camera using the Live Link Face software [Epic Games, 2023]. This camera measured the participants' activity of 61 facial landmarks, also referred to as *blend shapes*, and assigns them a value between 0 and 1 (indicating a low and high level of activity, respectively).

The data was analysed in three ways. Unfortunately, due to the scope of this project, only a *preliminary* analysis was performed. First, the video data was manually annotated, using the ELAN software [ELAN, 2023]. Both the most prominent visual cues and the used question structure were annotated. Furthermore, two methods of quantitative data analysis were employed on the 3D data set: the

temporal progression of the blend shape data was visualised and the HDBScan clustering algorithm was implemented. The latter provided us with the most prototypical combinations of facial features.

The five most prototypical facial expressions marking Dutch questions involve:

[1] Raised eyebrows and wide eyes (occurring mostly in situations where positive prior belief is later contradicted with negative evidence)

[2] Frowned eyebrows and squinted eyes (found frequently in situations where neutral contextual evidence is provided)

[3] Squinted cheeks, squinted eyes and a sneered nose

[4] Simultaneous squinted and wide eyes

[5] The neutral facial expression

The first two of these were also found to be question markers in NGT [Esselink, 2022], however, this was not the case for the latter three. Furthermore, the pattern of engagement of the corresponding features does not always resemble the patterns found in NGT.

Lastly, future avenues of research are discussed in detail, regarding changes in the experimental setup, pre-processing and data analysis.[1]

[1] The materials, 2D and 3D data, manual annotations of the video data, written code and some results are publicly available, see: `https://doi.org/10.6084/m9.figshare.c.7054445.v1`

Acknowledgements

I want to thank Floris Roelofsen, Marloes Oomen, Lyke Esselink, Pieter Manders, Yente Dodemont, Kyo Gerrits and all participants for their contribution to this project.

Personally, I want to thank my mom, dad, sister, grandparents, Angelina, Leonoor, Olivia, Tamana, the people I studied with at the Master of Logic, Loki and Kiwi, and Anouk ♡.

Contents

Chapter 1

Introduction

Human language is multimodal: auditory cues, involving speech, are combined with visual cues, including body movements and facial expressions, to relay communicative intention. In this project, the visual cues exhibited during the utterance of different types of biased polar questions by native speakers of the Dutch language are investigated. There are plenty of ways to ask polar questions, i.e. questions to which the answer is either *'yes'* or *'no'*. For instance, it is possible to add particles to the end of the question (think of *'right'*), add words indicating polarity (*'no'* or *'not'*) or change the word order of the question; these indicate different question structures. Furthermore, polar question utterances may be accompanied with different facial expressions, containing facial features such as frowned eyebrows, wide eyes or squinted cheeks; these are visual cues. We expect that speaker bias has a significant influence on the manner in which one decides to ask a polar question, both in terms of question structure and use of visual cues. Although much research has been done on question marking in Dutch (see [Borràs-Comes et al., 2014, Englert, 2010, Gaasbeek, 2023, Nota et al., 2021, 2023]), as well as the contexts in which certain sentence types and (spoken) polarity markings are present (see [Englert, 2010, Gaasbeek, 2023]), the combination between question structure and visual cues has been understudied for biased polar questions in the Dutch language. One aim of this project is therefore to investigate the different visual cues that are used when native Dutch speakers utter different types of polar questions and which cues are most prominently present in which context.

The current project was carried out at the SignLab in Amsterdam. A previous project at this institution has investigated the influence of bias on the non-manual markers (NMM)[1] used for polar question marking in Sign Language of the Netherlands (NGT) (this study will be referred to as the *BpQ/NGT-experiment* henceforth). The further aim of the current study is to compare the results that were found, regarding the Dutch language, to those found in the BpQ/NGT-experiment, for NGT. Conclusions can then be drawn, regarding which cues are prototypical for question marking in certain contexts in general, as well as about which cues are

[1] In sign language research, it is often spoken about NMM, a term which parallels the term visual cues for spoken languages.

specific to either Dutch or NGT. Therefore, the experiment set up and conducted for the current project is analogous to the BpQ/NGT-experiment, however, it was adapted to study Dutch polar question utterances in comparison to NGT. Participants engaged in various role-plays with two *confederates* (all of whom are native speakers of the Dutch language). These confederates presented both speaker bias and contextual evidence to the participants (which was either positive, negative or neutral), after which the participants asked the target question.

Not only is the experimental design similar to that of the BpQ/NGT-experiment, one of the current methods used for data analysis mirrors the data analysis for the BpQ/NGT-experiment as well (see [Esselink, 2022]). First, computer vision (CV) technology was used to collect participants' data (by means of a 3D depth camera, see Section 2.2.2), in addition to video data. Furthermore, the machine learning (ML) technique *clustering* was used to analyse this aforementioned data, in addition to a second method of data analysis.

The current chapter provides an introduction on the topic. First, Section 1.1 reviews the most prominent results found in the literature, regarding the NMM used for question marking in sign language (and specifically, NGT) (see Section 1.1.1), as well as the visual and auditory cues found during question utterances in spoken languages (Section 1.1.2) and specifically in Dutch (Section 1.1.3). The research questions this project aims to answer are then reported in Section 1.2. Section 1.3 presents the hypotheses, based on the literature overview provided.

1.1 Literature Overview

1.1.1 NMM in Sign Language of the Netherlands

Previous research has shown that NMM, such as eyebrow movements, head tilts and body movements, play a significant role in sign languages [Baker et al., 2016]. These markers can provide additional linguistic information at the phonological and morphological level (for instance, being an essential part of a lexical sign, or modifying the meaning of a sign without the need for an additional one, respectively), as well as the syntactic level and pragmatic level (for instance, to convey a biased polar question) [Pfau and Quer, 2010]. In particular, during question utterances, many sign languages make use of NMM. This phenomenon is widely studied. Since the current project aims to analyse the used visual cues in different types of biased polar questions in Dutch, as well as to compare these results to the BpQ/NGT-experiment, it is important to review the most prominent NMM used for question marking in sign languages, in addition to the research methods that these studies employed, particularly for NGT. This section provides such an overview.

Across sign languages, the manual signs of declarative statements and their corresponding polar question are often identical [Pfau and Quer, 2010]. NMM are generally the only feature distinguishing these two utterances from each other, hence they function as a marker for the syntactic structure. For polar questions specifically, these NMM include raised eyebrows, a head that is tilted forward and a forward body tilt [Coerts, 1992, Pfau and Quer, 2010, Zeshan, 2004]. This combination of features is often referred to as *'q'* [Coerts, 1992]. Furthermore, this definition of 'q' can include widened eyes and eye contact with the addressee of the question, even though this is not always the case [Coerts, 1992, Cecchetto, 2012]. Since NMM are exhibited simultaneously with manual signs, the markers indicate scope. The features in 'q' are often exhibited during the entirety of question utterances: their activity sternly increases at the start of the question, is continuous during the utterance, and quickly decreases at the end of the question utterance [Coerts, 1992]. For an overview of NMM marking questions in sign languages, see [Cecchetto, 2012, Coerts, 1992, Pfau and Quer, 2010, Zeshan, 2004].

Similar results have been obtained for NGT, specifically. Coerts [1992] found that the most prominent NMM marking polar questions in NGT include raised eyebrows and a head tilted forward. Furthermore, in some cases, a body tilt or wide eyes are exhibited. However, the presence of these features is not significant enough for it to be included in the definition of 'q' for NGT [Coerts, 1992]. Again, analogous to the scope of 'q' across sign languages, the scope of the NMM marking polar questions in NGT includes the entire question utterance.

Additionally, polar questions in NGT may be marked with a headshake, which is referred to as an *inquisitive* headshake in the literature [Oomen and Roelofsen, 2023b]. The shaking of the head primarily indicates a negative polarity, in contrast to a headnod, indicating positive polarity. However, this is not its only function: in question utterances, a headshake may indicate confusion or the need for a response by the interlocutor [Oomen and Roelofsen, 2023b]. Oomen and Roelofsen [2023b] further make a distinction between questions containing only a sentence radical (for instance *'Is Kim een vegetariër?'*, which translates to *'Is Kim a vegetarian?'*), or questions containing one or more *tags*. Tags are defined as 'particular combinations of manual and non-manual markers following the sentence radical, fulfilling a specific pragmatic function.' Think of questions as *'Kim is een vegetariër, toch?'* (*'Kim is a vegetarian, right?'*) or *'Kim is een vegetariër, toch, of niet?'* (*'Kim is a vegetarian, right, or not?'*), containing one and two tags, respectively. The inquisitive headshake, marking polar questions, is displayed most often during the utterance of these sentence-final phrases [Oomen and Roelofsen, 2023b].

Besides, next to raised eyebrows, head tilts, body tilts, wide eyes and inquisitive headshakes, signers of NGT occasionally manually mark polar questions by pointing their palms up. This marker never replaces 'q', but rather acts as an additional marker indicating that a polar question is being asked [Coerts, 1992].

Furthermore, De Vos et al. [2009] studied the affective function of the eyebrows in NGT. They remarked that whilst polar questions are often marked with raised-eyebrows, this eyebrow position further indicates surprise. The double function of this feature can lead to linguistic conflict for NGT signers, especially when both of these functions need to be employed at the same time. De Vos et al. [2009] found that in surprised polar questions, raised eyebrows are therefore strongly present, more so compared to its neutral counterparts. Additionally, it was found that at the end of target question utterances, frowned eyebrows are often present, regardless of its linguistic or affective function [De Vos et al., 2009].

More specifically, Esselink [2022] researched NMM in biased polar questions in NGT: the BpQ/NGT-experiment. As has been previously stated, one aim of the current project is to compare our findings to those found by Esselink [2022]. Accordingly, we will discuss the BpQ/NGT-experiment in more detail here, regarding both the experimental design and the data analysis. Note that the current experimental setup largely adopted the setup of the BpQ/NGT-experiment.

The aim of the BpQ/NGT-experiment [Esselink, 2022] was to investigate whether bias could account for the variation of NMM marking polar questions in NGT. During the experiment, participants asked elicited polar questions in a role-play setting with two interlocutors: *confederates*. These confederates provided participants with original speaker belief, as well as contextual evidence, which was either positive, negative or neutral. During the target question utterances, a 3D camera (with the aid of the *Live Link Face* application [Epic Games, 2023]) measured the extent of engagement of 61 *blend shapes*: facial landmarks corresponding to facial features. The HDBScan clustering algorithm was then implemented to extract the most prototypical facial expressions found in the data. Hence, this study combined computer vision (CV) technology (the 3D depth camera) with machine learning (ML) techniques to analyse the data (the clustering algorithm).

Two overarching facial expressions marking polar questions were found. The first contains raised eyebrows and wide eyes, and does not occur in a context in which a positive prior belief is contradicted with negative evidence. The second involves furrowed eyebrows and squinted eyes [Esselink, 2022].

Esselink [2022] further demonstrated a methodological refinement regarding NGT research. Most studies on NMM in sign languages analyse the captured data on the basis of manual annotations. Thus, for each moment in time, the position and behaviors of facial landmarks such as eyebrows, the mouth and cheeks, among others, are reported upon. Whilst this method of data analysis leads to important insights, this process is laborious, categorical and prone to subjective judgements. For instance, it is difficult to determine whether the position of the eyebrows is lowered, raised, or neutral, when participants are in the process of raising their eyebrows. Furthermore, inter-annotator disagreement might occur [Oomen et al., 2023], in addition to an annotator not agreeing with their own annotations at a later time.

Therefore, more recent studies employed CV technology and ML techniques to analyse sign language data, since these methods are more objective and reproducible in comparison to manual annotations. However, Metaxas et al. [2012] found that in a large number of cases, this CV technology was sensitive to occlusions of the face. Furthermore, only the low-level features, such as head and eyebrow moments, were recognised, compared to high-level features, combinations of these low-level features. Consequently, Metaxas et al. [2012] and Liu et al. [2014] took a different approach: they extracted 3D data from the 2D video data. Features of the low and high levels were now extracted based on the landmarks present in individual and multiple frames, respectively.

Kuznetsova et al. [2022] and Kimmelman et al. [2020] further built on this framework: they used OpenFace [Baltrusaitis et al., 2018] software to extract 3D data from the 2D video data. However, it was found that the OpenFace software returned biased data when participants' heads were tilted.

The use of the Live Link Face application [Epic Games, 2023] by Esselink [2022] refines this methodology. As stated, with the use of a 3D depth camera, 61 facial features were measured, referred to as blend shapes. For instance, these are *EyeWideLeft* and *JawOpen*. The corresponding measurements range between 0 and 1, indicating no activity or the highest level of engagement, respectively. Since this application directly measures the activity of facial features, therefore creating 3D data, instead of this 3D data needing to be extracted and translated from 2D data, the blend shape data is more reliable and precise. Therefore, the current project adopts the method used by Esselink [2022]: participants' blend shape measurements were captured by a 3D depth camera, in combination with the Live Link Face software, during the experiment.

1.1.2 Visual, Auditory and Morphosyntactic Cues in Spoken Languages

Not only are non-manual markers an essential part of sign languages, spoken languages also heavily depend on nonverbal cues. Polar question utterances often contain morphosyntactic cues, think of the French question particle *est-ce que*. Furthermore, some languages use a different word order to distinguish between declarative and interrogative clauses. For instance, in Dutch and English, SVO word order indicates an interrogative statement, whilst VSO word order marks polar questions [Borràs-Comes et al., 2014].[2] Furthermore, besides morphosyntactic markers, spoken languages make use of prosody to signal a polar question. According to Bolinger [1989], the high pitch that is present during the elicitation of polar questions is found across languages, and can be considered a linguistic universal.

The importance of the use of prosody as a marker for polar questions is additionally emphasised by Borràs-Comes et al. [2014]. They describe an experimental setup in which participants, native speakers of Catalan, played two versions of the game *Guess Who?*[3]: the *question-elicitation* variant, in which participants asked polar question to their opponent, and the *statement-elicitation* variant, in which participants described their mystery person using declarative statements. Data analysis showed that in Catalan, specific intonational patterns are used in order to distinguish a declarative statement from its interrogative counterpart. This effect was found to be considerably stronger compared to languages that do not 'lack such a *lexico-morphosyntactic* distinction' [Borràs-Comes et al., 2014]. Additional results from this study indicate that the use of visual cues, on top of auditory and morphosyntactic cues, provides a higher accuracy of understanding for the addressee, in comparison to situations in which is only relied on auditory cues [Borràs-Comes et al., 2014]

[2] SVO and VSO word order refer to the order in which the subject, object and verb are placed within a sentence. Hence, using SVO word order, the subject is uttered before the verb (*'Kim is a vegetarian.'*), whilst for VSO word order, this is the opposite (*'Is Kim a vegetarian?'*).

[3] In this game, players are presented with 24 drawings of faces. To win, players need to guess the other players' chosen *mystery person*. By asking polar questions to their opponent, they are able to eliminate drawings.

Zygis et al. [2017] further investigated the influence of sentence type on both the auditory cues, as well as oro-facial expressions, in German. To study this effect, Zygis et al. [2017] set up a production task, containing clauses representing eight different pragmatic uses: assertion, exclamation and six types of declarative questions[4] (such as those seeking justification or confirmation). During the experiment, participants were asked to read a scenario, which provided appropriate context to the target question; this context introduced the question type. The declarative clause was then read aloud by the participant, similarly to how they would utter this clause in the aforementioned scenario. During these utterances, facial movements were measured using OpenFace software [Baltrusaitis et al., 2018]. Zygis et al. [2017] found three main results. First, the position of the eyebrows was the highest for the experimental trials containing exclamations and assertions, whilst this was the lowest for echo, guessing and incredulity questions. The latter corresponds to a question in which a positive prior belief is contradicted with negative evidence. Lastly, the right eyebrow always presented a higher activity in comparison to its left counterpart [Zygis et al., 2017].

The influence of visual and auditory cues on question perception in Brazilian Portuguese was further studied by [Miranda et al., 2021]. Video-recordings were made of native speakers of Brazilian Portuguese, in which a sentence, which can either be interpreted as a declarative or an interrogative clause, was uttered. Participants were asked to judge whether these videos contained a declarative statement, or an echo question. It was found that, based on the use of prosody and nonverbal cues, participants were able to distinguish these two clauses with ease. Prosodic patterns for judged declarative clauses were falling, whilst this was rising for interrogative clauses. Additionally, visual cues marking questions include a right head tilt, eye blinks, lowered eyebrows, nose sneers and eye squints. Furthermore, in noisy conditions, visual cues improve the interpretation of auditory information [Miranda et al., 2021]. da Silva Miranda et al. [2020] build on this research, by comparing these findings to Mexican Spanish. Similar results were obtained regarding the auditory and visual cues. However, stretching ones lip also indicates the use of a polar question in Mexican Spanish [da Silva Miranda et al., 2020].

1.1.3 Visual, Auditory and Morphosyntactic Cues in Dutch

As the previous section describes, spoken languages make use of nonverbal cues, in addition to morphosyntactic and auditory cues, to mark polar questions. Accordingly, this phenomenon is also found in the Dutch language. This section briefly describes four studies in which the these cues marking polar questions in Dutch were investigated.

First, the experiment described above, investigating Catalan, was also carried out in Dutch [Borràs-Comes et al., 2014]. Hence, Dutch participants played the question-elicitation and statement-elicitation version of the *Guess Who?* game. The most prominent morphosyntactic cue found in this data was the use of VSO word order, compared to SVO word order.[5] This difference in word order is sufficient to

[4] Questions containing SVO word order.

[5] Continuing this report, we may refer to the word order of utterances in terms of inversion

distinguish declarative statements from polar questions. Additionally, a rise in intonation was found when participants neared the end of their question utterances, which further marks these as polar questions. Furthermore, Borràs-Comes et al. [2014] report that Dutch relies more heavily on the syntactic structure marking questions than on prosody patterns, and that these phenomena are related: when a SVO word order is present (hence, when the question is not syntactically marked), Dutch speakers make use of rising intonation more than they would in a question utterance containing VSO word order [Borràs-Comes et al., 2014]. Lastly, similarly to Catalan, Borràs-Comes et al. [2014] found that the addition of nonverbal cues to auditory information enhances the accuracy of understanding by the addressee.

The syntactic structure employed during Dutch question utterances was examined by Englert [2010]. She provided an overview of the ways in which native Dutch speakers formulate their utterances, in order to indicate them as questions. For polar questions, it was found that most utterances include either VSO word order (therefore, these are interrogative polar questions), or include SVO word order (declarative polar questions) in addition to a tag (a term that was introduced previously for NGT: a sentence-final particle). It was further found that polarity can be incorporated into the formulation of polar questions, by means of additional particles. Think of *'nee'* (*'no'*), or the Dutch *'wel'*.[6] As was mentioned before, the corresponding visual cues are a headshake or a headnod. Dutch native speakers further prefer to use a positive assertion in combination with a negative tag [Englert, 2010].

Lastly, [Nota et al., 2021] and [Nota et al., 2023] investigated the multi-modality of the Dutch language, the latter focusing primarily on eyebrow movements.

In [Nota et al., 2021], clusters of facial expression corresponding to question and answer utterances were obtained from corpus data, using decision tree models and multiple correspondence analysis. The features that were most frequently present in the data were frowned and raised eyebrows, squinted eyes, blinks, gaze shifts and smiles. Zooming in on question data specifically, the obtained clusters primarily contained frowned eyebrows, whilst those found in answer data primarily contained eyebrow raises and gaze shifts. Furthermore, in the data that consisted of response utterances, features were often more highly engaged in comparison to the question utterances, with the exception of eyebrow frowns and eye squints. When question utterances did exhibit a high activation of a feature however, this feature was significantly more engaged than during the response utterances. Lastly, the onset of widened and squinted eyes often occurred after the start of the question or answer utterances, whilst this was not the case for other facial features.

[Nota et al., 2023] studied the ways in which eyebrow movements mark polar questions in Dutch in a more detailed manner. Participants were presented with videos containing avatars uttering clauses. In these videos, eyebrow movements

instead of SVO or VSO position. In a clause containing VSO word order, the subject and verb are switched in comparison to a declarative clause: this clause therefore contains *inversion* of the subject and verb. In contrast, a clause with SVO word order does not include inversion.

[6] A proper translation of *'wel'* in English does not exist. This is a positive polarity marker, emphasising the fact which the question tries to confirm.

accompanying the utterances were manipulated. Participants were instructed to judge, as quickly as they could, whether this avatar uttered a declarative or interrogative clause. Nota et al. [2023] concluded that questions accompanied with eyebrow frowns are more accurately judged in contrast to those without any eyebrow movement. For eyebrow raises, a similar pattern was not found. Furthermore, the response time of participants was smaller for questions containing eyebrow frowns, in comparison to those without eyebrow movement. Again, a similar pattern was not observed for eyebrow raises. Additionally, the earlier the onset of the eyebrow movements, and the longer the duration of these movements, the faster the response time. Thus, this research suggests that polar questions are often accompanied with frowned eyebrows, since this leads to higher accuracies in understanding by the addressee and less response time.

1.2 Contribution

This project aims to investigate the different visual cues that are exhibited during different types of biased polar question utterances by native Dutch speakers. As previously described, much research has studied question marking in Dutch, as well as the contexts in which sentence types and (spoken) polarity markings are present. For instance, the most prominent question structure marking polar questions in Dutch includes the inversion of the verb and subject. Additionally, a rise in intonation is an important auditory question marker in Dutch [Borràs-Comes et al., 2014, Englert, 2010]. On the other hand, frowning ones eyebrows seems to be the most prominent nonverbal question marker [Nota et al., 2021, 2023]. As one can clearly see, these studies investigate the auditory, morphosyntactic or visual cues used for polar question marking, in addition to the enhancement of one of such cues when the others are not present (see [Borràs-Comes et al., 2014]). However, the combination between the visual cues and the used question structure has not been examined for *biased* polar questions in the Dutch language, using a elicitation task. Hence the current project provides an empirical contribution to this field. Furthermore, previous studies have not used the Live Link Face [Epic Games, 2023] software to investigate visual cues in Dutch. Consequently, blend shape data has not been analysed in combination with auditory data that was simultaneously captured. This thesis project therefore provides this additional methodological contribution.

Given that we are interested in the visual question markers of different types of polar questions, we first need to ascertain what these different types of question structures, regarding polarity markings, word order and sentence type, actually are. Moreover, we aim to determine how bias influences the use of these different types of questions; hence, how these question types correspond to context. This leads us to our first research question:

(RQ1) **What are the possible question structures in Dutch and how do these correspond to context?**

After this question has been answered, we then aim to discover how these different types of biased polar questions are marked in spoken Dutch. To answer this question, not only are we interested in studying which combinations of visual cues are most prototypical during question marking, we aim to find how these combinations of cues corresponded to question structure (again, these structures regard the use of word order, sentence type and polarity markings), and how these combinations of cues progress over time. The second research question, and its four sub-questions, are formulated as follows:

(RQ2) **How are different types of biased polar questions visually marked in Dutch?**

(a) What are the prototypical combinations of visual cues that mark polar questions?

(b) To what extent does context determine which combinations of visual cues occur in question utterances?

(c) What is the temporal progression of these various combinations of visual cues?

(d) How do these various combinations of visual cues interact with question structure?

This research question and its four sub-questions concern the visual cues exhibited during Dutch polar question utterances. The interaction with question structure is encompassed in the latter sub-question; the visual cues expressed by participants and their simultaneously used question structure are investigated.

The third and last research question is presented below. A comparison is made between the results of the BpQ/NGT-experiment and the current experiment. In this way, we inquire whether the prototypical cues for question marking, and their characteristics, are comparable in Dutch and in NGT, or whether they are are language-specific. Note that for the BpQ/NGT-experiment, Esselink [2022] does not report on the relation between question structure and visual cues, whilst this is the case for the current project. Therefore, only the results regarding visual cues and NMM are compared.

(RQ3) **How do the results of this study compare to those found in Sign Language of the Netherlands?**

1.3 Hypotheses

Based on the literature overview of Section 1.1, some hypotheses are formulated regarding the visual cues and question structures we expect to find in our data set.[7]

[7] Note that apart from the BpQ/NGT-experiment [Esselink, 2022], the previous described studies do not take into account the influence of bias on the presented visual cues. Thus, even though the hypotheses are based on the reviewed studies, we keep in mind that these are not substantially comparable.

First, regarding the question structures, we hypothesise that inversion of the subject and verb is a prominent marker of polar questions in Dutch, which would be in line with findings by [Borràs-Comes et al., 2014, Englert, 2010].

Additionally, rising intonation is commonly known to signal polar questions in Dutch, especially if these questions involve SVO word order [Borràs-Comes et al., 2014]. This research project, however, does not investigate prosodic patterns in the data. Only Section 4.3.1 briefly discusses some observations.

Furthermore, we hypothesise that target question utterances might contain tags [Oomen et al., 2023, Englert, 2010] (such as *'toch'* (*'right'*)), or particles signaling polarity [Englert, 2010] (such as *'niet'* (*'no'*) or *'wel'*, marking negative and positive polarity, respectively). Moreover, as [Englert, 2010] found, Dutch native speakers prefer to use a positive assertion with a negative tag. Therefore, we hypothesise that when polarity is marked, using particles, this polarity is negative.

One visual cue we expect to see is the frowning of the eyebrows. Both Nota et al. [2021] and Nota et al. [2023] concluded that Dutch native speakers often frown their eyebrows in question utterances. Zygis et al. [2017] further found that in German, a language which is relatively similar to Dutch, frowned eyebrows were often present during question utterances for which a positive prior belief was contrasted with negative evidence. A comparable result was found in NGT: a facial expression consisting of raised eyebrows and wide eyes was fully absent in this condition. Consequently, we hypothesise that this pattern is similar for Dutch, and therefore that we frequently find facial expressions containing frowned eyebrows, particularly in the condition in which positive speaker belief is contradicted with negative contextual evidence.

On the other hand, we do not expect to discover that the facial expression containing raised eyebrows is such a prominent question marker in Dutch. Even though Esselink [2022] and Coerts [1992] reported that wide eyes and raised eyebrows mark polar questions in NGT, Nota et al. [2021], Nota et al. [2023] and Zygis et al. [2017] found that these visual cues do not show a significant activity during question utterances. Therefore, we hypothesise that this facial expression is used during question utterances, albeit to a lesser extent than the expression containing lowered eyebrows.

The structure of this thesis report is as follows. The following chapter, Chapter 2, describes the design of the experiment that was conducted in more detail, as well as the participants, the experimental procedure and the setup of the recording studio.

Chapter 3 proceeds, discussing the ways in which the video data, as well the data captured by the 3D depth camera, was pre-processed in order to fit data analysis. This data was first synchronised, after which noise was removed and only relevant data was selected from the 3D data. Next, the data captured by the 3D depth camera was transformed: dimensions were reduced, after which this data was normalised and ranged, in order to perform the two data analysis methods.

Chapter 4 describes the procedure of manually annotating the video data in ELAN [ELAN, 2023]. The ELAN software is briefly introduced, after which the annotation template that was designed specifically this project is described. The chapter concludes with a preliminary analysis of the manually annotated video data, thereby answering our first research question. Based on this analysis, subsets of the complete normalised and ranged data sets are additionally obtained.

What follows is Chapter 5, which describes the data analysis of the captured 3D data that was implemented for this project. As a first step, an additional dimension reduction was performed, to remove even more noise. The two methods of 3D data analysis are then reviewed in more detail. First, the mean measurements captured by the 3D depth camera were visualised over time, based on the ranged data set. Next, the HDBScan clustering algorithm was implemented, based on the normalised data set, which provided us with the most prototypical combinations of facial expressions exhibited during target question utterances.

Chapter 6 reports the most striking findings resulting from these two forms of data analysis. Furthermore, these results are compared and summarised.

Lastly, Chapter 7 briefly reviews the results, and compares the current findings with the results found in the literature. Thus, a comparison is made here between the results of the BpQ/NGT-experiment and the current experiment, answering our final research question. Chapter 8 concludes, and suggests further avenues of research.

Chapter 2

Experiment Design and Data Collection

This chapter provides a detailed description of the experiment design and data collection. As stated before, the experiment conducted for this thesis project aimed to investigate the visual and morphosyntactic cues that signal biased polar questions in Dutch. The objective was to conduct an experiment similar to the experiment described in Oomen and Roelofsen [2023a]. Whilst that experiment was designed to investigate NGT, the experiment described here investigated spoken Dutch. In this section on experimental design and data collection, the experimental stimuli, materials, recording studio setup and experimental procedure that are used in this elicitation experiment, are reported upon.

Section 2.1.1 first provides a general description of the experiment design, after which Section 2.1.2 discusses the experimental stimuli that have been designed for this experiment. The participants whose data was captured are discussed in Section 2.2.1, whilst Section 2.2.2 describes the setup of the recording studio in which the experiment took place. The experimental procedure is discussed in Section 2.2.3. The subsequent sections loosely follow the structure of the article by Oomen and Roelofsen [2023a], which describes the BpQ/NGT-experiment design this study was based on.[1]

2.1 Experimental Design

2.1.1 General Description

During the experiment, participants, native speakers of the Dutch language, were prompted to ask questions to two different confederates in a role-play setting. These confederates, A and B, were referred to as *Robin* and *Sam* respectively, during the experiment. In response to the utterances by the participants, the confederates read out prescripted answers, hereby introducing original speaker bias and contextual

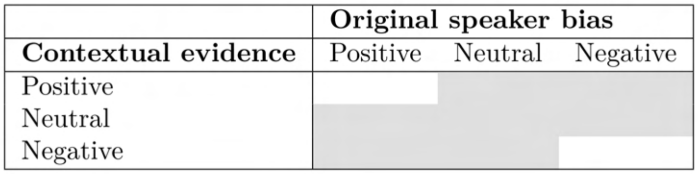
[1] The materials used during the experiment, as well as the captured 2D video data, are publicly available, see: `https://doi.org/10.6084/m9.figshare.c.7054445.v1`

evidence. This original speaker bias and contextual evidence was either positive, negative or neutral. Lastly, the participants were prompted to ask a target question to the second confederate, Sam. This process was repeated for multiple variations of several situations.

The two confederates are both native Dutch speakers. They were hired by the SignLab at the University of Amsterdam, specifically for their contribution to this project. Both of the confederates have acting experience. Because of this, they were able to introduce this original speaker bias and contextual evidence in a convincing manner.

2.1.2 Experimental Stimuli

For this experiment, six situations were designed to elicit polar questions from participants. The first of these situations is used for practice (see Section 2.2.3). The situations are heavily based on the scenarios used in the experiment by Oomen and Roelofsen [2023a], which in their turn were loosely based on the experimental stimuli in Domaneschi et al. [2017]. For every situation, there were seven trials that were conducted during the experiment. These trials contain different combinations of introduced original speaker bias and presented contextual evidence, which can both be either positive, negative or neutral. The only combinations that have not been investigated are *PosPos* and *NegNeg*, since asking the target question in these cases would be highly unnatural. This is because the contextual evidence confirms the participants' original speaker bias. See Table 2.1 for an overview of the experimental conditions. Appendix A provides the script of all situations and their possible variations. Note that in this appendix, and in the example that will be given below, the script has been translated to English. During the experiment itself, only the Dutch language was used.



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

Table 2.1: Experimental Conditions [Oomen and Roelofsen, 2023a].

Every trial had the same structure, and always contained three stages: the introduction of original speaker bias, the introduction of contextual evidence, and the elicitation of the target question. In the first stage, the participants interacted with confederate A, whilst they interacted with confederate B in the second and third stage. Figure 2.1 visualises the structure of all trials.

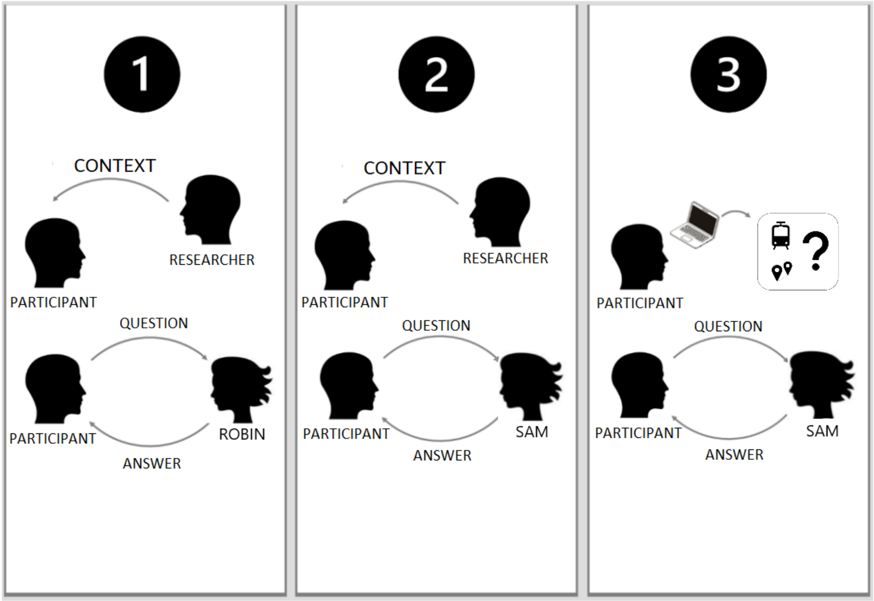


Figure 2.1: Visualisation of the structure of the trials (adapted from Esselink [2022]).

A description of the three stages now follows, by means of an example trial from the practice situation (adapted from Oomen and Roelofsen [2023a]).

Stage 1: The introduction of original speaker bias

The first stage of each trial started with one of the researchers reading out some context to the participants. This context described the current situation and afterwards prompted the participants to ask confederate A a polar question. The participants, at this stage, did not yet have any bias concerning the answer to this question, since the trial had just started. It is important to note that the content of the polar question asked by the participants was always identical, or very similar, to the target question that was elicited from the participants in stage three. After the participants asked the polar question, confederate A (*Robin*) responded to this polar question, introducing original speaker bias, which was again either positive, negative or neutral. In the illustrated example here, the participants were prompted to ask whether there is a metro station nearby, to which confederate A responded as stated, therefore introducing negative speaker bias to the question asked by the participants in this stage.

Context 1: You recently moved to Amsterdam. You are currently at your house, but would like to go the city center. You don't know if there's a metro station nearby. You're meeting Robin, your new neighbor. Ask Robin.

Participant: 'Is there a metro station nearby?'

Confederate A: 'No, there's no metro station nearby.'

Stage 2: The introduction of contextual evidence

The second stage of each trial started similarly to the first stage: the participants were read another context to the situation by one of the researchers, which prompted the participants to ask confederate B (*Sam*) a polar question. This question, content-wise, was not the same as the question the participants had asked in the previous stage. In fact, these questions were often not directly related. Confederate B then responded to this polar question, hereby introducing contextual evidence, which again was either positive, negative or neutral. In the example here, the polar question uttered by the participants in this stage concerns the way to the city center, instead of a metro station nearby. This question is not directly related to the question from the previous stage and therefore also not directly related to the target question. However, this question does provide an opportunity for confederate B to inform the participants on the content of the target question, thereby introducing contextual evidence. Hence, the participants were prompted to ask the way to the city center, to which confederate B responded as stated, hereby introducing positive contextual evidence to the target question.

Context 2: On your way, you meet Sam. Ask Sam whether she knows the best way to the city center.

Participant: 'Do you know the best way to the city center?'

Confederate B: 'There's a metro station here around the corner. You should take line 51 to Weesperplein, which is close to the city center.'

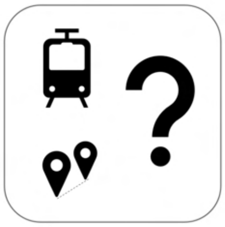
Stage 3: The elicitation of the target question

Each trial concluded with the participants asking the target question to confederate B. In contrast to the past stages, the participants asking the question now directly followed the interaction from the previous stage. The target question, which was almost (always) identical to the question elicited in stage one, was elicited by means of a picture prompt. These images contained multiple icons representing concepts that were present in the target question. Participants were instructed that these concepts had to be present in their formulation of the target question (see Section 2.2.3). We had explicitly chosen to give participants linguistic freedom, to ensure the delivery of the target question and the accompanying visual, auditory and morphosyntactic cues came as naturally as was possible to the participants. The expectation was that the formulation of the target question would differ for the different combinations of original speaker bias and contextual evidence.

Going back to the example given here, the picture prompt contained an icon of a metro and a symbol representing the concept *nearby.* The question mark next to these two symbols was a reminder for the participants to formulate a question, and not a declarative statement. Therefore, the target question was some variation on the question *'Is there a metro station nearby?'* The confederates in the current example provided negative original speaker bias, but positive contextual evidence to the target question. Hence, this example trial corresponds to the *NegPos* condition. One could expect that this conflicting information triggers question forms that signal this discrepancy somehow. For instance, think of the use of particles indicating negation, or the addition of sentence-final particles such as *'right'* (*tags*). After the participants uttered the target question, confederate B answered with a brief and unscripted response.

It is important to note that it was a deliberate decision to elicit the target question using picture prompts instead of video recordings. This is because the lexical order of the target questions uttered by the participants should be influenced as minimally as possible. For this same reason, the icons in the prompts were placed on top of each other instead of next to each other. Lastly, there were two variations of each picture prompt, in which the two icons next to the question mark were switched [Oomen and Roelofsen, 2023a] (see Appendix A for the different prompts). One half of the participants saw the first version of the picture prompts, while the second half was presented with the second version. Note that the icons in the picture prompt could still influence the lexical order of the target questions, since some participants might read these from top to bottom. For half of the trials, the participants were therefore presented with a picture prompt in which the upper icon corresponded to the concept expected to be used first in the target question (such as in the picture prompt below). In the other half of cases, the participants viewed the symmetric counterparts. As a result, the influence of the picture prompt on the participants' chosen lexical order was minimized even more.

Picture Prompt:



Participant: A variation on the question: 'Is there a metro station nearby?'

2.2 Data Collection

The collection of data for this project took place in the months of October and November 2023. The eleven experimental sessions were held on six days, the first of which contained two pilot sessions. Hence, the recorded data of nine participants was used for data analysis. On average, the experimental sessions lasted for an hour and a half, including breaks.

2.2.1 Participants

For this project, we collected data of eleven participants, between the ages of 18 and 25, all from the Randstad area (a region in the Netherlands including Amsterdam and its surroundings). Four participants are male, and seven are female. The participants were recruited in a variety of ways: via flyers that were hung in public spaces at the University of Amsterdam, messages spread amongst students of the General Linguistics Master and the Master of Logic at the University of Amsterdam, in addition to personal recruitment. All participants are native speakers of Dutch and use the Dutch language on a daily basis. Furthermore, they all speak some variation of (standard) Dutch and therefore do not only speak a Dutch dialect.

Prior to the experimental sessions, participants were asked to give their informed consent for the collection, processing and analysis of the data, which was required to participate in this experiment. Additionally, we asked the participants to give their informed consent for the online publication of the recorded data, as well as the discussion of this anonymised data in academic publications. This consent was not required to participate in the experiment. Moreover, participants were asked a few personal questions about their language background. These questions concerned their first and second languages, the first language of their parents and the area they grew up in.

As stated in Section 1.1.1, 3D data was captured during the experiment. According to Esselink et al. [2023], the effect of glasses on measurements by the 3D depth camera is not yet investigated. However, it is expected that glasses have an influence on certain blend shape measurements, regarding for example the squinting of the eyes and raising of the eyebrows. To avoid any imprecision this may cause, we asked participants to not wear glasses during the experimental trials. Furthermore, to avoid that participants blended in with the background, we requested they wore non-green, plain or monochrome clothing.

2.2.2 Recording Studio Setup

This experiment was conducted in the sign language studio at the University of Amsterdam. This studio contains a green wall, used in a similar fashion as a green screen, as well as studio lamps to ensure a good video quality. The setup of the recording studio for this project is similar to the setup used in Oomen and Roelofsen [2023a]. Figure 2.2 provides an overhead view, Figure 2.3 shows the studio setup during the pilot sessions (during which two iPhones were used, see below).

The participants stood in front of the green wall (marked with the black X in Figure 2.2), with two studio lamps on either side, illuminating the wall. Another studio light illuminated the participants. The confederates' position is marked with the gray X in Figure 2.2.

The participants were recorded by three cameras (Sony FX30 Cinema Line), all positioned on a tripod. The cameras filmed the participants from the side and from the front (the *participant* camera), capturing their body movements (such as head and body tilts) and facial expressions, whilst the third (*confederate*) camera recorded an overview of the studio during the sessions. Therefore, this camera did not only record the participants, but also recorded the confederates and the researchers. The decision was made to record the entire studio, to ensure that double checks could be performed if needed; for example in cases where participants uttered unexpected responses. Naturally, this footage was not used for data analysis. The cameras, with the confederate camera as an exception, were connected to a *Tentacle Sync* device (see Section 3.1) [Tentacle Sync, 2023a]. Furthermore, one iPhone 13 containing a depth camera for 3D recording, in combination with the *Live Link Face* application [Epic Games, 2023] (which was discussed in detail in Section 1.1.1), captured the participants' facial features during the experiment. The iPhone was connected by bluetooth to one of the Tentacle devices that was attached to a camera. Initially, the decision was made to use two iPhones, recording the participants from either side. However, during the pilot sessions, it became clear that handling two iPhones at the same time, as well as properly storing the captured data between experimental sessions, was too complex. Therefore, we determined it was best to use only one iPhone, similarly to the setup by Oomen and Roelofsen [2023a]. Since the back cameras of iPhones do not contain Apple's True Depth Sensor, the iPhone needed to face the participants during the sessions. In order to avoid the participants being distracted by being visible on the iPhone screen, the screen of the iPhone was blacked out.

It was necessary for the angles and distances of cameras recording the participants to be identical between participants, as well as to the participants in the BpQ/NGT-experiment [Esselink, 2022]. The reason for this is that Esselink et al. [2023] found that camera angle has a small, yet significant effect on the measurements by the 3D depth camera. Furthermore, the distance between the depth camera and the participants has a slight effect on the measurements as well, albeit less than that of angle. Because we want to compare results between participants, as well as to compare the results of this study to the BpQ/NGT-experiment [Esselink, 2022], we have closely followed the setup of the cameras used in that experiment.

To ensure a proper audio quality, a unidirectional microphone (Mke 600 Shotgun Microphone (Sennheiser)) was placed in front of the participants. This microphone recorded all sound signals in the studio, but was facing the participants since only their utterances were needed for data analysis.

Besides the participants and the two confederates, two experimenters were present during the experimental sessions. The lead experimenter was positioned behind the laptop (see Figure 2.2), and was responsible for guiding the sessions by means of giving instructions, projecting stimuli and handling the participant camera and depth camera, which could be controlled via the laptop. The monitor and the laptop (see Figure 2.2) were connected to ensure that this experimenter could present the right stimuli at the right time to the participants. The second experimenter was positioned between the confederates and the studio lamp next to the confederates. Their job was to handle the confederate camera and the camera facing the participants from the side. Furthermore, this experimenter kept track of the trial numbers during the experimental sessions.

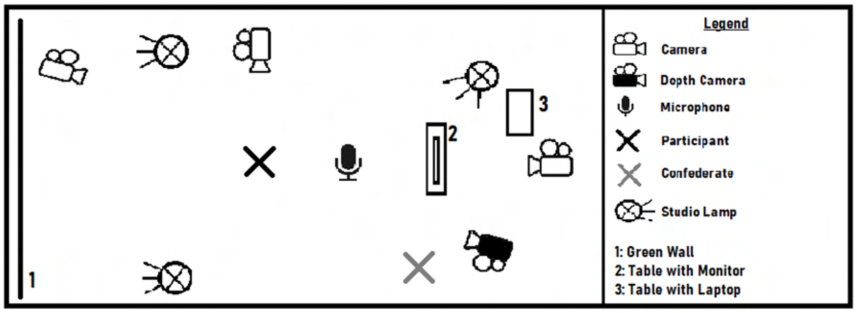
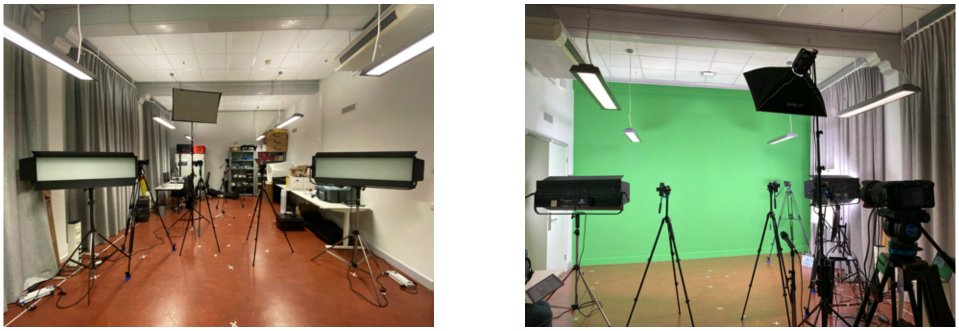


Figure 2.2: Overhead view of the recording studio setup.



((a)) Front view.

((b)) Back view.

Figure 2.3: Recording studio setup during the pilot sessions.

2.2.3 Experimental Procedure

Participants were given a brief introduction by the lead experimenter after arriving to the studio. The participants were reminded that the experimental session would be both audio- and visual-recorded, and that they had given permission for the analysis and publication of the data that would be collected. This permission had been given prior to the experimental session, in writing. Furthermore, they were assured that the experimental session could be stopped at any time, and that they could request their data be removed after the session.

After the introduction, the participants watched multiple pre-recorded and Dutch instruction videos on a laptop. In the first of these videos, participants were informed that during the experimental session, they would be interacting with the two confederates *Robin* and *Sam*, and that they would have to ask these confederates questions during small role-plays. After watching this instruction video, the participants watched a live demonstration of two example trials (from the practice situation, see Appendix A) in which one of the researchers took on the role of the participant. This was shown to the participants to help them understand the structure of the trials and the kind of utterances we were after.

Next, the participants watched two more pre-recorded videos, in which the structure of the trials was explained. These videos used Figure 2.1 for visualisation. Participants were instructed to always ask the confederates polar questions and were informed they were free to use the word order and facial expressions that they liked. Additionally, participants were instructed to keep their productions brief, at maximum one sentence, and to speak as naturally as possible.

The last instruction video explained that there were six different situations: one practice situation and five experimental situations. Every situation had seven different variations, hence there were thirty-five experimental trials. Moreover, participants were told that the context videos and picture prompts in every situation were always identical, but that the responses from the confederates might differ in each trial. Lastly, participants were explained they might like to change their way of asking questions, based on the confederates' responses.

After the participants had finished watching the pre-recorded instruction videos, they were given the opportunity to ask any questions they might have. Additionally, they were told they can ask clarifying questions at any time during the experimental session. Furthermore, it was again stated that there would be regular breaks and they could request to take a break when needed.

Having finished the instructions, the 3D depth camera was calibrated on the participants' neutral faces. The blend shape values measured in this calibration image served as a baseline, to which the values captured during the experiment were adjusted [Epic Games, 2023].

Next, the four practice trials were conducted, to make sure the participants understood the trial structure and experienced the different ways in which they might decide to ask the target question. Before each of the trials began, a clapper was used to ensure that the auditory and visual data could be easily synchronised later on [see Nota et al., 2021], using Tentacle Sync synchronisation equipment Tentacle Sync [2023a] (see Section 3.1 for a detailed overview).

After all these preparatory steps, the experimental trials began; thirty-five trials in total. We found that often, participants did not need to be presented with context in every variation of a situation (similarly, this was found to be the case during the BpQ/NGT-experiment [Oomen and Roelofsen, 2023a, Esselink, 2022]). Of course, this context was presented when needed.

Even though some participants did not use either VSO word order or a tag during the target questions utterances for certain trials, the decision was still made to count these occurrences as questions, on the grounds that these utterances still elicited a response by the second confederate. However, if the target question utterance was clearly a statement, the participants' question was off-target, or participants stated their utterance did not feel natural, the situation was re-recorded.

Following the seven trials of a situation, the participants were asked to transform the target question into a statement (for instance: *'Is Kim een vegetariër?'* (*'Is Kim a vegetarian?'*) became *'Kim is een vegetariër.'* (*'Kim is a vegetarian.'*)), which we recorded as a baseline.

As a final step, *calibrations* were recorded, for every participant, in which they engaged the following facial features one by one, to the fullest of their abilities: raised eyebrows, frowned eyebrows, wide eyes, squinted eyes, squinted cheeks, sneered nose, shrugged mouth and frowned mouth.[2] Around three calibrations were recorded per feature. Pictures of these facial features were provided in case of confusion. This step, suggested by Esselink et al. [2023], ensured that the normalisation of the data during data pre-processing was relatively straightforward. Initially, during the pilot sessions, this step was performed before the experimental trials began. However, we expected participants therefore focused on their facial expressions during the trials, and consequently, that their natural facial expressions would not be reflected in the data.

After all the data was collected, participants were thanked and paid for their contribution to the project and were given the opportunity to share their experience and ask more questions.

[2] Esselink [2022] found that these facial features did not produce any noise in the data set. For this project, we have adopted this selection of facial features to investigate. See Section 3.3.1 and Section 5.1.1 for a more in depth explanation.

Chapter 3

Pre-processing the Data

After the data had been collected, it required pre-processing in order to perform data analysis in the later stages of this project. In this chapter, the procedure of pre-processing the data is discussed at length. First, the recorded camera and 3D depth camera footage needed to be synchronised, see Section 3.1. The second step was to select the relevant data and remove noise, which is reviewed in Section 3.2. Lastly, the data was transformed, which Section 3.3 describes.[1]

3.1 Synchronising the Footage

For this thesis project, the aim was to perform a frame-by-frame analysis of the data. This analysis was executed using both the captured blend shape values by the 3D depth camera, as well as the manual annotations in ELAN [ELAN, 2023] (see Chapter 4). Therefore, it was necessary to synchronise the footage of the two participant cameras and the 3D depth camera, since these cameras were not turned on at the exact same time. As stated previously, the recorded data by the confederate camera, facing the entire studio, will not be used for data analysis. Consequently, these recordings were excluded during pre-processing.

In order to be able to align the recorded footage from the three cameras, we have used Tentacle Sync E equipment [Tentacle Sync, 2023a]. These are small devices that can be connected to recording devices, generating a timecode. As stated before (see Section 2.2.2), the two participant cameras were connected to a Tentacle Sync E device by cable, whilst the iPhone containing the 3D depth camera was connected to one of these two devices using Bluetooth.[2] The Tentacle devices now generated a timecode for the device to which they were connected. Using the phone application *Tentacle* [Tentacle Sync, 2023b], the timecodes of the Tentacle Sync devices were now manually synchronised, therefore the timecodes of the cameras connected to these devices were also synchronised. The frame rates of the cameras and the 3D

[1] The RStudio and Python code written for pre-processing, in addition to the resulting 3D data set, are publicly available, see: `https://doi.org/10.6084/m9.figshare.c.7054445.v1`

[2] The Live Link Face application [Epic Games, 2023] contains a feature which lets the Sync E device determine the timecode.

depth camera was set to 59.94 and 60 frames per second (fps) respectively, whilst the Tentacle Sync device was set to 29.97 fps in the Tentacle application.

Since both the cameras and 3D depth camera now contained the same timecode, it initially seemed straightforward to align the recorded footage by loading these recordings into an application such as Adobe Premiere Pro [Adobe, 2023] and synchronising these based on timecode. Unfortunately, this was not possible. During testing, before the experimental sessions took place, we found that the device connected to a Tentacle Sync device by Bluetooth did not pick up the timecode generated by that Tentacle device. Therefore, we could not synchronise the recorded footage by the 3D depth camera to the other camera footage, based on the timecode. Furthermore, it was found that the timecode generated by the two Tentacle Sync E devices connected to the cameras, always differed with the same number of frames. For some recording sessions, the timecodes were exactly one frame apart, while for others this was twelve or 32 frames. This number was consistent for every recording session. Thus, when cameras were turned off, this number would change for the next recording session. Unfortunately, this problem was not solved before the experimental sessions took place. This meant that the footage recorded by the two cameras and the 3D depth camera needed to be synchronised afterwards, either on sound wave signals, or manually. Nevertheless, the decision was made to use the Tentacle Sync equipment anyway, since the timecode of the cameras were the same number of frames apart for every participant. This could prove to be beneficial in case the recordings needed to be manually synchronised.

As stated, the recorded data by the three cameras needed to be aligned after the experimental sessions took place. The synchronisation was performed using Adobe Premiere Pro software [Adobe, 2023]. First, for each participant, the relevant recordings were selected for every condition in each situation. Hence, the recordings in which the participants' target question was off-target, or they asked this question in a way that felt unnatural to them, were discarded. Afterwards, the recordings of the trial by the three different cameras were loaded into Adobe Premiere Pro.

As was described in Section 2.2.3, a clapperboard was used at the start of each recording. Initially, the purpose of this clapperboard was to review whether the footage was properly synchronised by the Tentacle equipment [Tentacle Sync, 2023a]. However, as stated, we found during testing that this was not the case. We therefore tried to align the recordings, from the three different angles, based on the sound waves of this clapperboard. But, synchronising the recorded data based on sound signals does not synchronise this recorded data visually: the recordings were still a few frames apart for every recorded trial. A probable reason for this is that not every camera had the same distance to the clapperboard. Therefore, each device might pick up the audio signals of the clapperboard at a different time than the remaining devices.

Since the recorded data could not be synchronised based on timecode or audio signals, it had to be manually aligned. For every recording, the uniquely identifying frame in which the clapper closed was used to align the three video recordings.

Unfortunately, some footage was lost for the third participant (p3), situation 5 condition *PosNeut*. The video data captured by the 3D depth camera did not match the other camera recordings. The iPhone storage became full during the trial, hence this trial needed to be re-recorded. We suspect that we lost the 3D depth camera footage because of this, although the exact reason is not completely clear. The recorded data could therefore not be aligned and was excluded. Furthermore, in two trials, the camera filming the participant from the side was not recording. This was the case for p4 situation 3 condition *NeutPos* and p8 situation 1 condition *PosNeg*. Even though the participants were not recorded from the side during these trials, we did not exclude these trials from the data set. The reason for this is that the 3D depth camera did capture the blend shape values for this trial. Furthermore, the footage of this trial could still be annotated, albeit we could not examine the participants' body movements in as much detail as for other trials.

3.2 Selecting the Relevant Data

After the recorded data of both the two cameras and the 3D depth camera were aligned, the relevant data needed to be selected. The devices recorded the entirety of the role-play for each trial. Therefore, the footage in which the participant uttered the elicited target question needed to be extracted for manual annotations (see Chapter 4), as well as the corresponding blend shape data captured by the Live Link Face application [Epic Games, 2023] for data analysis (see Chapter 5). The relevant video footage was selected using Adobe Premiere Pro [Adobe, 2023], whilst the extraction of the blend shape data captured during these utterances was performed in RStudio [Posit Software, 2023]. A detailed description of the latter process now follows.

For every recording by the 3D depth camera, a corresponding CSV file was created. This is a $63 \times (n + 1)$ file, in which the timecode, number of blend shapes and the 61 blend shape values were documented for all n frames within the duration of the recording. For each of these CSV files, the blend shape data corresponding to the target question utterance was manually extracted. Afterwards, all of these CSV files were compiled into one dataset.

To accomplish this, we adapted the code used in Esselink [2022] to fit the current data set. The first step was to load all of the trimmed CSV files, containing the captured blend shape data, into RStudio [Posit Software, 2023]. In this phase of pre-processing, we discarded the trial for which the 3D depth camera data was lost (p3 situation 5 condition *PosNeut*). Furthermore, the situations were discarded in which the participants' uttered target question was off-target. This was the case for four situations. Lastly, we found that the Live Link Face application [Epic Games, 2023] did not always capture the blend shape data during the target questions utterances. Consequently, the blend shape data in these situations could not be analysed. This was the case for ten situations. The number of CSV-files we therefore ended up with equals 345.[3]

[3] This number includes the *Declarative* conditions.

For these files, we then identified each frame with a video ID (unique for each of the 345 files), the participant ID $(3-11)$, the scenario ID $(1-5)$, the condition ID (e.g. *PosNeut*) and the ID for a frame number (unique for each frame). All of these IDs identifying each frame are added as columns to the dataframe.

As a last step, these 345 CSV files were now combined to form the *untrimmed* data set. This data set contained 68 columns, one for each of the IDs, the timecode, the number of blend shapes and each of the 61 measured blend shapes, and 68.031 samples.

In the BpQ/NGT-experiment [Esselink, 2022], it was found that participants' signs sometimes caused occlusions of the face during the target question utterances. This missing data was then interpolated. Just in case the current data set contained occlusion (for instance, in cases where participants showed hand movements with an affective function), the current data set was also checked for missing values and interpolated.

Even though the CSV files were manually trimmed to contain only the data corresponding to the target question utterances before they were loaded into RStudio, this data still contained noise. In some cases, the start and end of the target question were marked a second too early or late. In other cases, the participant expressed their surprise or confusion (*'Oh'* / *'Huh'*) before they asked the target question. Since we are only interested in those frames that are contained within the target question utterance, these frames were again marked with much precision for each of the 345 recordings, to remove as much noise as was possible from the data set.

The frame numbers for which the participant started and stopped asking the elicited target question were marked in an XLSX file. Moreover, this file included the video ID, participant ID, scenario ID, condition ID, the duration of the utterance of the target question in number of frames, and the number of frames per window (the previous number divided by five, see below). The information contained in this XLSX file was then incorporated into the untrimmed data set. Namely, four additional columns were added: the number that indicated the frame ID for the start and end of the target question utterance, the number of frames of the duration of the utterance of the target question, and the number of frames per window. Besides this, a column was added to encode which frames were used during data analysis. The values this column contained were binary: 1 indicated the frame was used, whilst 0 indicated the frame was not used for data analysis. The last supplemental column gave each frame, labelled as used for data analysis, a uniquely identifying frame ID. For the frames labelled as not used, this frame number equalled 0. The data set now comprised 74 columns and 68.031 samples.

In this thesis project, the aim was to not only investigate which facial expressions are used while asking polar questions, but also to model temporal development of these facial expressions. A common approach used in previous (sign language) research is to normalise the recordings of the questions asked by participants, so that their utterances started and ended at exactly the same frames (for instance, see Kuznetsova et al. [2022]). However, as Esselink [2022] rightly remarked, this approach is not feasible for analysing a data set using a clustering algorithm, for which the participants were given linguistic freedom during the experiment. Since the participants could use their preferred word order and formulation of the question in the experiment conducted for this project, the used word order differed significantly between situations and participants. For example, participants added tags (*'Er gaat een trein om 9 uur, toch?'* (*'There is a train at 9am, right?'*)), embedded their question (*'Ik dacht dat er een trein ging om 9 uur?'* (*'I thought that there would be a train at 9am?'*)) or added some extra information (*'Gaat er een trein om 9 uur naar Parijs?'* (*'Is there a train at 9am to Paris?'*)).

Since we could not normalise the data on time for the implementation of the clustering algorithm in the further stages of this project, we adopted the method used to model temporal development by Esselink [2022]. All video frames were assigned to one of five temporal windows, placing the frames within a time frame contained in the target question recording. Hence, if a frame was assigned to the first window, we know this frame was captured somewhere at the start of the recording. Because the duration of the video recordings of the target question differed with regards to frame numbers, the sizes of the windows also differed between recordings. However, within a recording, this number was consistent, and equalled the number of frames captured during the trial divided by five. It is important to note that only those frames that did not contain noise were assigned to the temporal windows.

As discussed, Live Link Face [Epic Games, 2023] captures the timecode and the number of blend shapes that were measured during the recording. These columns were discarded from the combined CSV file. The data set now contained 72 columns and 68.031 samples.

Lastly, the *trimmed* data set was created, which only consisted of those frames that did not contain noise. To build this CSV file, the frames labelled as *not used for data analysis* were discarded. Furthermore, the irrelevant columns containing the starting and end frames of the target question utterance, the duration of the video in frame numbers, the previous unique frame ID and the column encoding which frames are used for data analysis, were discarded. As a last step, the unique frame IDs were reset, so as to start their count from 1. The trimmed data set contained 67 columns and 33.525 samples.

3.3 Transforming the Data

After the trimmed data set was obtained, combining data from all conditions, situations and participants, this data set now required to be transformed. The code that was used for this stage of the project was written by Esselink [2022], and adapted as to fit the current data set. First, we performed dimension reduction (even more of these dimensions were later reduced, see Section 5.1.1), in which highly correlated and symmetrical features were combined. Next, new data sets were created, containing only normalised and ranged blend shape values. These steps will now be reviewed in more detail.

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 3.1: Symmetric features, their correlation and their combined feature name. Courtesy of Esselink [2022].

3.3.1 Dimension Reduction

As one might expect, many facial features are highly correlated. This, for instance, is often the case for blend shapes describing the same feature, one for the right and one for the left side of the face. Whilst this information, describing the differences between these symmetric features, proves to be useful when creating realistic animations [Apple, 2023], this information is not relevant for the purpose of this project. Therefore, we would like to simplify the data set by combining these correlated features into one. Since the acquired data set for this project is quite similar to the BpQ/NGT-experiment, and given that the correlated behavior of the combined features in [Esselink, 2022] was observed within the current data set as well, the selection of features to combine was adopted from Esselink [2022].[4]

[4] Note that this dimension reduction does not take into account that certain features, such as *BrowInnerUp* and *BrowOuterUp*, or *BrowDown* and *EyeSquint*, frequently occur together, for anamotical reasons. See Section 7.2 and Chapter 8 for a more in depth discussion.

Table 3.1 presents all recorded blend shapes containing a version for both the left and the right side of the face, and their computed correlation. As can be seen, most symmetrical features are highly correlated, hence a new feature is created containing the mean value of its left and right counterparts. The features for which this is not the case are *JawLeft, JawRight, MouthLeft, MouthRight, BrowInnerUp, BrowOuterUpLeft* and *BrowOuterUpRight*. Hence, these blend shapes are still present in their original state in the new and reduced data set. Lastly, this was also the case for a few features that do not contain a left and right variant: *CheekPuff, HeadPitch, HeadRoll, HeadYaw, JawForward, JawOpen, MouthClose, MouthFunnel, MouthPucker* and *TongueOut*. The resulting data set therefore comprises 39 features.

3.3.2 The Normalised Data Set

After the first dimension reduction took place (see Section 5.3 for a description of the second reduction), two new data sets were created, containing both the normalised and ranged values, respectively.

For each participant, the range in which they are able to engage certain facial features differs. For the purpose of this project, we were not interested in the participants' ability to engage facial features to a maximum extent. Rather, we aimed to investigate the ways in which the participants engaged their facial features during the experiment, in comparison to their own maximum ability. To eliminate these differences between participants, which could have had a significant effect on the outcomes of data analysis, the data set was normalised. As a first step, the blend shape values that were measured during the calibration recordings (see Section 2.2.3) were added to the data set. The minimum and maximum values for each blend shape value were obtained from this combined data frame, containing both the recordings of the calibrations and the target question utterances. The latter recordings were inspected, in case these contained even higher measurements of the 39 blend shapes, compared to those captured during the calibration recordings. After this, the normalised values were computed and stored as a new data set (in which the calibration recordings were discarded). The formula that was used to compute the normalised values is presented below [Esselink, 2022]. Here, x is the measured blend shape value, whilst x' is its normalised counterpart, expressed as a percentage of the participants' range.

$$x' = \frac{x - min}{max - min} \times 100 \tag{3.1}$$

Consider the example given in Table 3.2. In this table, the maximum and minimum values for two measured blend shapes (f_1 and f_2) are given for two participants (p_1 and p_2). Furthermore, three example values are given: v_1, v_2 and v_3. As can be seen in Table 3.2, the blend shape values for f_1 of the first participant range between $0 - 97$, whilst this is between $0 - 80$ for the second participant. Similarly, for the second feature, measurements of p_1 range between $3 - 49$, in contrast to p_2,

	f_1		f_2	
	p_1	p_2	p_1	p_2
min	0	0	3	1
max	97	80	49	76
v_1	31	27	6	6
v_2	56	56	28	32
v_3	92	77	42	65

((a)) Original Values

	f_1		f_2	
	p_1	p_2	p_1	p_2
min'	0	0	0	0
max'	100	100	100	100
v_1'	32	34	7	7
v_2'	58	70	54	41
v_3'	95	96	85	85

((b)) Normalised Values

	f_1		f_2	
	p_1	p_2	p_1	p_2
min''	0	0	2	2
max''	89	89	63	63
v_1''	28	30	6	6
v_2''	52	62	35	27
v_3''	85	85	54	58

((c)) Ranged Values

Table 3.2: Comparison between original, normalised and ranged blend shape values. Note that the values are multiplied by 100 for readability. The resulting values are rounded to the nearest integer.

for which this range is $1 - 76$. Hence, it can be concluded that p_1 is able to engage f_1 to a higher extent, in contrast to f_2, for which p_2 can express this feature to a higher extent.

These values are now normalised using Equation 3.1. For f_1, value v_2', the calculations are as follows:[5]

$$v_2' = \frac{56 - 0}{97 - 0} \times 100 = 58 \tag{3.2}$$

$$v_2' = \frac{56 - 0}{80 - 0} \times 100 = 70 \tag{3.3}$$

Even though both participants engaged f_1 to the same extent, their normalised blend shape values now differ significantly. The fact that p_2 is not able to engage this facial feature in a similar manner as p_1 is now captured within the data.

Besides, the normalised values of two participants for a certain feature might be the same, whilst their measured blend shape values differ. For example, see value v_3' of f_2:

$$v_3' = \frac{42 - 3}{49 - 3} \times 100 = 85 \tag{3.4}$$

[5] Note that for all equations in this chapter, the outcomes are rounded to the nearest integer.

$$v'_3 = \frac{65 - 1}{76 - 1} \times 100 = 85 \tag{3.5}$$

The normalised data set now contains a fair representation of the participants' engaged facial features, since the blend shape measurements for each participant now fell within the same range (in comparison to only their range, as was the case before). Therefore, this normalised data set was used for machine learning purposes (see Section 5.3).

3.3.3 The Ranged Data Set

As has been stated before, next to the reduced and normalised data set, a data set was created containing the ranged values per feature. The aim for this project was to not only perform machine learning techniques onto our data set, but also to visualise the mean blend shape values over time, for different conditions, situations and participants. The normalised values were representative of the data set for machine learning purposes, since the differences between participants could now be computed, as stated. However, these normalised blend shape values could not be used for our visualisation of the data set, since these values did not capture the actual magnitude of the expressed facial feature. For instance, consider a facial feature for which all participants are only able to engage this feature within the range $0 - 70$. This means that, according to the normalised values, when Live Link Face [Epic Games, 2023] captures a value equal to 70 for this blend shape, this is represented as having value 100 in the normalised data set. Consequently, one might conclude that the feature is therefore maximally engaged in this frame. However, this is not the case, since this facial feature can be engaged to a higher extent; this was just not found in our data set.

To account for this misrepresentation of the normalised data set and to be able to visualise the data later on (see Section 5.2), we furthermore computed the measurements as a range of mean values [Esselink, 2022]. First a new minimum and maximum value were computed, by taking the mean of all participants' minimum and maximum value, respectively. For f_2, we get the following:

$$min'' = \frac{3 + 1}{2} = 2 \tag{3.6}$$

$$max'' = \frac{49 + 76}{2} = 63 \tag{3.7}$$

The next step was to compute the ranged values, which were found using the following formula [Esselink, 2022]. Here, max'' and min'' are the newly computed maximum and minimum values, x' is the normalised value, and x'' is its corresponding ranged value.

$$x'' = (\frac{x'}{100} \times (max'' - min'')) + min'' \tag{3.8}$$

Hence, for the first value (value v''_3), and the second feature (f_2) we get the following (for p_1 and p_2, respectively):

$$v_3'' = (\frac{85}{100} \times (63 - 2)) + 2 = 54 \tag{3.9}$$

$$v_3'' = (\frac{91}{100} \times (63 - 2)) + 2 = 58 \tag{3.10}$$

Table 3.2 contains further examples of these original, normalised and ranged values. The calculations of these values will not be provided here, but one can check these examples for their own comprehension.

Chapter 4

Analysis of Video Data

As has been previously stated, this thesis project aims to perform a frame-by-frame analysis of the captured video data, using both the recorded blend shape values, as well as manual annotations. This chapter reports on the method of manually annotating the data using ELAN [ELAN, 2023], as well as some preliminary results obtained through a brief analysis of these annotations. Section 4.1 first provides a general introduction on ELAN and the data that has been annotated. The tier structure and accompanying controlled vocabularies are discussed in Section 4.2. Section 4.3.1 and Section 4.3.2 conclude this chapter, identifying and analysing the prominent visual cues and question structures that were present in the data, respectively.[1]

4.1 Introduction to ELAN

Non-manual markers (NMM) play a significant role in sign languages, as Section 1.1.1 remarked. These NMM such as facial expressions and body movements do not only have an affective function, they often provide linguistic information. Therefore, investigating NMM plays a prominent role in current-day sign language research. More often than not, sign language studies analyse the NMM present in the data set using manual annotations, which are captured in software such as ELAN [ELAN, 2023]. ELAN is an annotation tool, developed by the Max Planck Institute for Psycholinguistics [MPI, 2023], in which both visual and auditory data can be annotated. Not only does the use of ELAN annotations lead to important insights in sign language research, these manual annotations have proven to be useful for research into multimodal communication as well [Wittenburg et al., 2006] (see for instance [Sugahara et al., 2022, Turchyn et al., 2018]). For this reason, we have used the ELAN software in this project.

ELAN [ELAN, 2023] contains tiers, which are layers in the program that form a hierarchical structure. This structure can be manually created, by assigning tiers a

[1] The manual annotations that were performed, as well as the designed annotation template, are publicly available, see `https://doi.org/10.6084/m9.figshare.c.7054445.v1`

parent tier when they are first constructed. As a consequence, the annotations on the *child tier* can only fall within the annotations that were created on its parent tier.

Annotations can be assigned to tiers and span a certain time frame. One creates such an annotation by selecting a time frame in which a non-manual marker is distinctly present, and annotating the corresponding feature using free text or a controlled vocabulary. A controlled vocabulary consists solely of concepts or phrases that can be used for the specific tiers this controlled vocabulary is linked to. For example, a tier that corresponds to the position of the eyebrows can adopt the controlled vocabulary consisting of the possible annotations *lowered, neutral* or *raised.* Other tiers, such as those containing glosses or comments, often allow free text annotations.

As previously stated, ELAN supports both video and audio files. Therefore, for instance, word order and prosody can be annotated, in addition to the accompanying visual cues. The PRAAT software [Boersma and Weenink, 2023] visualises the progression of intonation over time. The corresponding files can be imported into ELAN, simplifying the annotation process of the prosodic patterns.

After the uploaded files are annotated, the resulting file is saved as an EAF file. The annotation template, containing the tier structure and corresponding controlled vocabularies, can be saved as an ETF file and an ECV file, respectively.

For this thesis project, the aim was to investigate the visual cues that are present when one asks different types of biased polar questions. Therefore, the blend shape data captured by the 3D depth camera was not sufficient for our data analysis, due to the fact that this data does not document the head and body movements, prosodic patterns, sentence type, polarity markers and word order used by the participants. The latter three types of information, regarding morphosyntactic cues, needed to be incorporated into our data set before starting data analysis. ELAN [ELAN, 2023] is a suitable tool for capturing this additional information. Furthermore, although the Live Link Face application [Epic Games, 2023] has proven to be relatively accurate [Esselink, 2022], the angle and distance between participants and the camera has a significant effect on the captured blend shape values [Esselink et al., 2023] (see Section 2.2.1). Therefore, manual annotations provide an additional reliability to the current data analysis. Accordingly, considering these two objectives, our data was annotated in ELAN in this stage of the project.

However, there are some limitations to working with ELAN [ELAN, 2023]. Apart from the fact that manually annotating video and audio data in much detail is laborious and time-consuming, the process is also categorical and not objective. NMM are gradable, yet need to be converted into categorical annotations. This both results in information loss as well as creates a major source for inter-annotator disagreement. Illustrating: when considering eyebrow movements, eyebrows can be annotated as either *lowered, neutral* or *raised.* It cannot be the case that their annotated position falls somewhere in between. Consequently, different annotators might not agree on the eyebrow position in this transition period. Besides, one annotator might annotate one frame as having neutral eyebrows, while annotating this frame as having raised eyebrows at different moments in time [Esselink, 2022, Oomen et al., 2023]. Furthermore, Oomen et al. [2023] rightly remarked that, even though sign language research heavily relies on manual annotations in software such as ELAN, the inter-annotator agreement is often not properly assessed. It is important to consider whether two different annotators use the same annotation labels at the same time, before conclusions can be drawn based on their annotations. In Oomen et al. [2023], the inter-annotator agreement was assessed. It was found that, specifically in transitional periods, annotators did not always agree on the annotation label. Because of the above described limitations, not only the manual annotations in ELAN used were used for the purpose of this project, but rather a combination of the blend shape data captured by the 3D depth camera and the manual annotations (see Chapter 5). Furthermore, only tentative conclusions, based on the annotations, are drawn currently.

4.2 Tier Structure

After the decision was made to use the ELAN software [ELAN, 2023] for manual annotations, a tier structure now had to be created that fit our data set and aims for this project. The hierarchical structure that was developed, was inspired by the annotation scheme constructed by Oomen et al. [2023], which was developed to annotate NMM in sign language research. In this structure, NMM such as eyebrow movements, eye shapes, body positions, head positions, lip movements and nose movements can be annotated, next to the Dutch and English glosses corresponding to the used signs. Since the data in our experiment contains spoken Dutch utterances, we could not adopt this hierarchical structure as is. Furthermore, not all visual cues needed to be annotated in as much detail: validating the blend shape values captured by the 3D depth camera using our manual annotations was not feasible for the scope of this project. Moreover, this level of detail was not required for the performed data analysis (see Chapter 5), whilst this was the case for the data captured in the BpQ/NGT-experiment [Esselink, 2022, Oomen et al., 2023]. Consequently, the tier structure that was developed as to fit the current data set was a simplified version of the structure developed by Oomen et al. [2023], and was furthermore adapted to fit spoken Dutch instead of NGT. What follows is a detailed description of the hierarchical tier structure. See Figure 4.1 for an example of a fully annotated question using the tier structure described below.

The first tier, and only tier that does not have a parent tier, is a tier called *Sentence-type*. The purpose of this tier was to annotate whether the polar questions uttered by the participants only contained a sentence radical (*SR*), such as *'Is Kim een vegetariër?'* (*'Is Kim a vegetarian?'*), or might also contain *tags*, such as *'toch'* (*'right'*) or *'of niet'* (*'or not'*) (see Section 1.1.1). If the participant used one or two of these tags, for instance, *'Kim is een vegetariër, toch, (of niet)?'* (*'Kim is a vegetarian, right, (or not)?'*), the annotation created on the Sentence-type tier would be *SR-T1* or *SR-T1-T2*, respectively (for instance, see Figure 4.1). Besides these three annotations, two possible annotations for this tier include *EMB* and *EXCL*. The first applies to cases in which participants embedded their polar question (*'Ik dacht dat Kim een vegetariër was?'* (*'I though that Kim was a vegetarian?'*)),

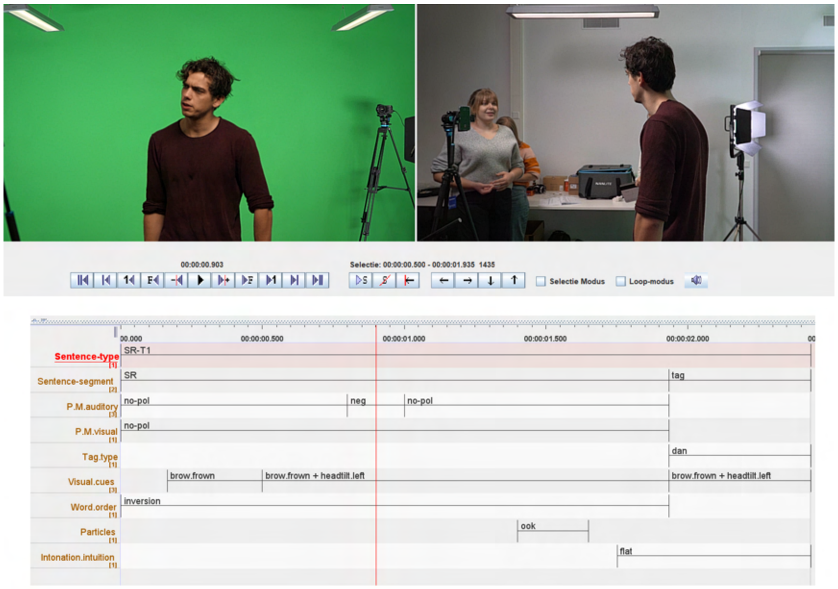


Figure 4.1: Example of a fully annotated target question in ELAN [ELAN, 2023], using the tier structure as described above. The video still aligns with the red line.

whilst the second was used in cases where the participants' target question was off-target[2] (*'Dat is Kim toch?'* (*'That's Kim right?'*), in which the word *'vegetariër'* (*'vegetarian'*) was not included). Because the annotations on this tier covered the sentence type, the span of all of these annotations equalled sentence length. Furthermore, since there are only five possible sentence types, this tier is linked to a controlled vocabulary containing the values *SR, SR-T1, SR-T1-T2, EMB* and *EXCL*.

Next, the *Sentence-segments* tier was created. This tier has the *Sentence-type* tier as its parent tier. The annotations that were created on this tier indicated when the participant uttered both the sentence radical and the tag. The time frame in which the participant uttered the sentence radical is now annotated as *SR*, whilst the time frame in which the participant utters the tag is annotated as *tag* (again, see Figure 4.1 for an example). Therefore, the controlled vocabulary accompanying this tier contains only these two possible annotations. If the target question only contained a sentence radical and no tag, the entire sentence was annotated as *SR*.

It is important to note that in case the sentence type was annotated as *EMB* or *EXCL*, no further annotations were added on the subsequent tiers, given that the sentence did not contain a sentence radical. As a consequence, the tiers having the *Sentence-segments* tier as their parent tier could also not take on any annotations.

The excluded trials were naturally not used for data analysis. Moreover, the preliminary analysis regarding manual annotations was not performed on the data captured during trials in which the target question involved embedding. There

[2] As Section 3.2 describes, this only occurred four times.

headtilt.left	headmove.left	bodytilt.left	eye.lookleft	mouth.frown
headtilt.right	headmove.right	bodytilt.right	eye.lookright	mouth.leftcornerup
headtilt.forward	headmove.forward	bodytilt.forward	eye.wide	inq.hs
headtilt.backward	headmove.backward	bodytilt.backward	brow.frown	
hand.nod	pointing	cheek.squint	brow.raise	

Figure 4.2: Annotated visual cues

are multiple different ways in which one can embed a polar question. To ensure patterns would be visible after this preliminary data analysis, the target questions needed to be categorised based on the different kinds of embedding. This was required before further manual annotations and data analysis can be performed. However, this was not feasible for the scope of this thesis project. Since the embedded and excluded trials were not used for the (preliminary) data analysis and were therefore discarded during this stage of the project, the decision was made to not provide the video and audio recordings of these trials with further annotations.

The following six tiers are assigned the *Sentence-segments* tier as their parent tier. Note again that these tiers can only take on annotations when the participant has used a sentence radical in their formulation of the target question.

The first two of these child tiers are the *P.M.auditory* and *P.M.visual* tier. On these tiers, the spoken and visual polarity markers were annotated (see Section 1.1.1). These annotations indicate either positive polarity (*pos*), negative polarity (*neg*) or no polarity (*no-pol*) found during the target question utterance.[3] Hence, both of these tiers are accompanied by a controlled vocabulary containing these three possible annotations. Particles indicating negative polarity include *'niet'* or *'geen'* (with the respective English translations *'not'* and *'no'*), whilst the positive spoken polarity marker is *'wel'*.[4] Visual polarity markers include headnods, to express positive polarity, and headshakes, to express negative polarity. The latter should not be confused with an inquisitive headshake [Oomen and Roelofsen, 2023b], which signals uncertainty or requests a response by the interlocutor, as stated in Section 1.1.1.

Another tier, having *Sentence-segments* as their parent tier, is the *Visual cues* tier. This tier has a free vocabulary, hence any possible annotation could be created. The main purpose of this tier is to annotate the most prominent visual cues that were observed. For instance, if a participant notably raised their eyebrows, or tilted their body forward, this would be annotated. Moreover, if inquisitive headshakes were present, annotations on this tier would indicate these. A list of all annotations used on this tier, for all target questions by participants, is presented

[3] For the spoken polarity markers, this applies only when the participant utters the sentence radical.

[4] As mentioned in Section 1.1.3, *'Wel'* does not have a corresponding English translation. When used it stresses the statement that is made. For instance: *'Kim is wel een vegetariër?'* translates to *'Kim is a vegetarian?'*, in which *'wel'* sheds some extra light on the fact that Kim is a vegetarian.

in Figure 4.2. The *Visual Cues* tier illustrates that the annotation template used for this project contains notably less detail than the template used in Oomen et al. [2023], considering that the latter contains separate tiers for different NMM, whilst the hierarchy developed for this project only contains one of these tiers. Note that the *Visual Cues* tier does not allow for two different cues to be annotated at the same time. Therefore, when two visual cues were simultaneously present, for instance when the participant would show frowned eyebrows and a head tilted to the left, this was annotated as *brow.frown + headtilt.left* (see for instance Figure 4.1).

Additionally, the *Tag type* tier is also a child tier of *Sentence-segments.* This tier specifies the tag(s), sentence-final particles succeeding the sentence radical, that is (are) used by the participant. The controlled vocabulary corresponding to this tier contains the following tags (with the respective English translations in brackets): *'toch'* (*'right'*), *'dus'* (*'so'*)[5], *'dan'* (*'then'*), *'he'* (*'right'*)[6], *'of niet'* (*'or not'*), *other.*

On the *Word order* tier, which again has *Sentence-segments* as its parent tier, it was annotated whether or not the participants' utterance of the sentence radical involved inversion. As stated in Section 1.1.3, inversion is indicated by a VSO word order, hence the verb is placed before the subject: *'Is Kim een vegetariër?'* (*'Is Kim a vegetarian?'*), whilst an SVO word order indicates no inversion: *'Kim is een vegetariër?'* (*'Kim is a vegetarian?'*).

The last tier with *Sentence-segments* as its parent tier is *Particles.* Similarly to the *Tag type* tier, annotations on this tier indicate whether a particle is used. Note here that a particle is different than a tag, since its span is contained within the sentence radical. Again, the *Particles* tier contains a controlled vocabulary, consisting of the following particles (with their respective English translations in brackets): *'toch'* (*'right'*)[7], *'dus'* (*'so'*), *'eigenlijk'* (*'actually'*), *'misschien'* (*'perhaps'*), *'ook'* (*'also'*), *'maar'* (*'but'*), *'gewoon'* (*'just'*), *other.*

The final tier in our hierarchical structure is the *Intonation intuition* tier, which has *Sentence-structure* as its parent tier. As stated before in Section 4.1, ELAN [ELAN, 2023] allows the intonation patterns by participants, visualised by the PRAAT software [Boersma and Weenink, 2023], to be loaded into ELAN to allow for further annotations. However, considering the scope of this thesis project, an analysis of prosodic cues using PRAAT before annotating these patterns in ELAN was not feasible. As an alternative measure, prosodic patterns were annotated based on intuition. The controlled vocabulary for this tier contained three annotation options: *rising, falling* and *flat.* For all target questions, the span of these annotations only concerned the end of the sentence radical and the tag.

[5] In Dutch, the word *'dus'* is used as a tag (although it can also be used as a particle): *'Kim is een vegetariër, dus?'.* For the English translation, the word *'so'* is not used as a tag, but rather as a particle (see the *Particles* tier): *'So, Kim is a vegetarian?'*

[6] At first glance, the words *'toch'* and *'he'* seem to have the same meaning, since their English translation is the same. However, there is some nuance to their definition. Both of these tags indicate some positive prior belief, for instance about the fact that Kim is a vegetarian. But, *'he'* expresses this belief more strongly than *'dus'.* See [Gaasbeek, 2023] for an extensive analysis.

[7] Note here that *'toch'* can be used both as a particle and a tag, whilst for the English translation *'right'*, this can only be used as a tag.

4.3 The Observed Visual Cues and Question Structures

This chapter concludes by giving an overview of the most prominent visual cues, particles and sentence types that were found in the data set. This overview is based on a brief analysis of the manual annotations, for which the frequency counts of certain phenomena were obtained. In addition, the ways in which these cues correspond to the different experimental contexts, is further discussed. The findings presented here give us a glimpse into the results we could expect to obtain from the analysis of the 3D data.

4.3.1 Observed Spoken and Morphosyntactic Cues

Whilst every participants' way of asking the target questions differed, there were quite a lot of similarities between participants, especially with regards to the types of question structures they employed during the experiment.

One of the most noticeable occurring morphosyntactic features was the frequent use of embeddings: over 1 in 6 target questions involved embedding. All but one participant (p6) used a target question structure involving embedding at least once during the 35 trials. For one participant (p5), this even occurred in over 40% of cases (15/35 times). The most common ways in which target questions were embedded were *'Ik dacht dat ...'* (*'I thought that ...'*) and *'Weet je of ...'* (*'Do you know whether/if ...'*).

Another apparent phenomenon was the infrequent use of tags (30/252 times), which contrasts with the BpQ/NGT-experiment, for which around a third of the target questions contained a tag [Oomen et al., 2023]. Only one participant regularly used tags: p7 (14/35), whilst this was not the case for the other eight. In fact, two participants did not use any tags during the experiment (p8 and p10), in addition to one participant only who only used tag (p9). Only three tag types were frequently used, which were *'toch'* (15/30), *'dus'* (3/30), and *'dan'* (11/30). The latter was used by only one participant (p7), however this participant used this tag frequently. Aside from these three tags, one other tag was found: *'weet je dat'*, which translates to *'do you know this'*. This tag was only used once, by p3. One more thing to note is that double tags were never found, which again contrasts with the BpQ/NGT-experiment, for which this was the case in a small number of instances [Oomen et al., 2023].

With regards to word order, question structures involving both inversion and no inversion were present in the data set, at around the same rate (129 vs. 123 cases). Some participants were more prone to using inversion, such as p7 (21/35) and p8 (30/35). Other participants did not show a clear preference. It is important to note that for questions containing a tag, inversion was rarely used. This is because for most Dutch tags, similarly to their English translations, this is grammatically incorrect.[8]

Particles indicating polarity were certainly present within the data (113/252 trials contained a spoken polarity marker). Similarly to word order, some partici-

[8] This is not the case for the tags *'of niet'* (*'or not'*) or *'dan'* (*'then'*). In these cases, inversion is possible, and in the second case it is even preferred.

pants were prone to mark their sentences with either positive or negative polarity, whilst this was less of a preference for other participants. For instance, p3 relatively infrequently marked polarity in their speech (8/35), whilst the other participants, in particular p8 and p10, did this quite often (15/35 and 17/35, respectively). Furthermore, generally speaking, negative polarity (in 95 out of 113 cases in which polarity was marked, this polarity was negative) was marked more often than positive polarity (18/113).

As for the remaining particles, many instances were found in the data (111/252). The most commonly used particles in this experiment were *'dus'* (19/111) and *'ook'* (78/111). The first of these was used predominantly at the start of each sentence radical, whilst this was not the case for the latter.[9] Furthermore, the frequency at which these particles were used, in addition to the position they had in the sentence, differed for each participant. For instance, p10 often used *'dus'* and *'ook'* in the middle of their utterances, whilst p9 had a clear preference for using *'dus'* at the start of their sentence.

Lastly, all participants used prosody as a prominent question marker. These patterns were found for all but one participant. In cases where only a sentence radical was used, the intonation at the end of the target question was often rising, which is a typical Dutch question marker. When a tag was used, however, the intonation contour on the sentence radical was falling, followed by a rising intonation during the utterance of the tag. One participant (p7) showed the same results to a degree. However their intonation at the end of the target questions was often flat (17/35 trials).

4.3.2 Observed Visual Cues

Similarly to the spoken cues, participants frequently engaged certain facial features, amongst other visual cues, while asking the target question. However, the variety of visual cues that were used by participants was significantly less than for the spoken cues that were found in the data set. Especially p5 and p9 rarely used any visual cues (13/25 and 2/25, respectively).

The most prominent facial feature that was found in the annotations was the frowning of the eyebrows (107/252 trials). More than half of the participants (p3, p4, p6, p7 and p11) frowned their eyebrows frequently during the experiment (12/35, 15/35, 11/35, 25/35 and 16/25 instances, respectively). Figure 4.1 serves as an example. Not only were eyebrow frowns found in the data set, participants also raised their eyebrows at times (39/252 times - often simultaneous with widening their eyes). However, this was not nearly as frequent as the lowering of the eyebrows. These observations are in line with those from previous studies (see Section 1.1 [Esselink, 2022, Nota et al., 2021, 2023]).

Furthermore, participants often moved their head while asking the target question. This was either in the form of a headshake, indicating negative polarity (or in the case of p4, sometimes an inquisitive headshake, found in 5 of 35 trials), a head tilt to the left or right (see Figure 4.1 again for an example of this), or a

[9] Hence, *'ook'* was only used mid-sentence radical. For instance: *'Kim is ook vegetarier, toch?'* (*'Kim is also a vegetarian, right?'*). Contrary to English, starting the sentence radical with the Dutch *'ook'* yields an ungrammatical sentence.

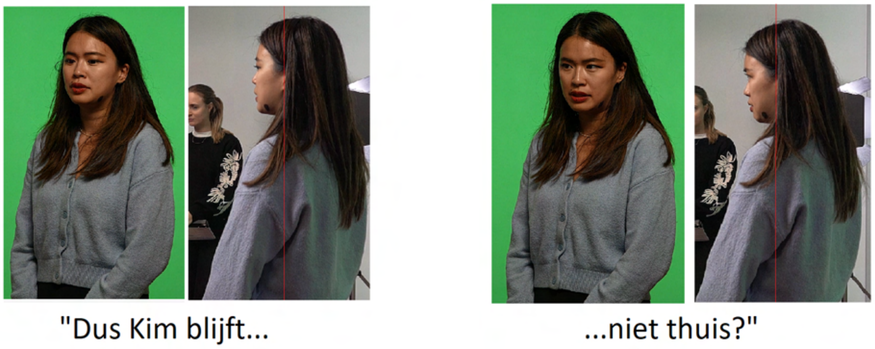


Figure 4.3: Example of a head tilted forward, found in situation 4 condition *PosNeg*. The red line visualises the body axis. *'Dus, Kim blijft niet thuis?'* translates to *'So, Kim does not stay home?'*

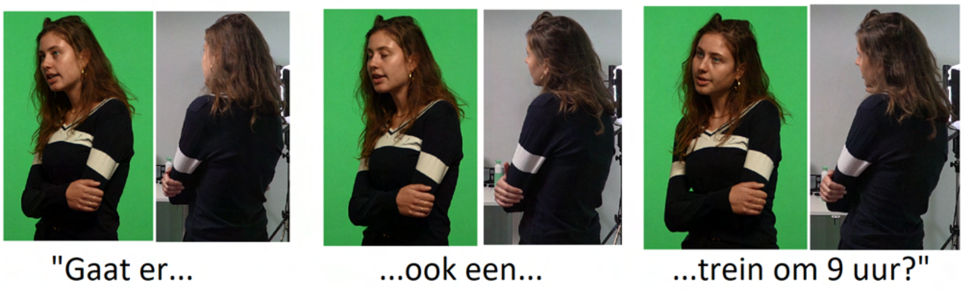


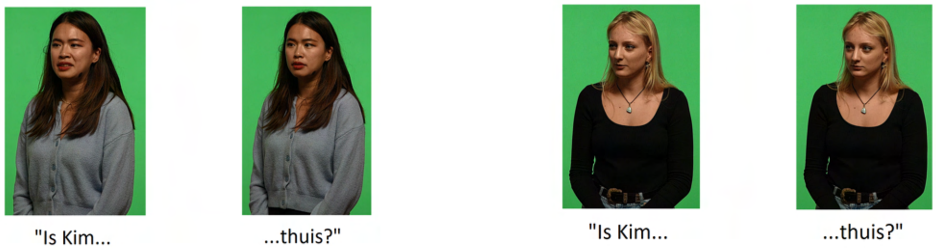
Figure 4.4: Example of a body tilt in combination with a head tilt, found in situation 3 condition *NeutNeut*. *'Gaat er ook een trein om 9 uur?'* translates to *'Is there also a train at 9am?'*

head tilt forward (see Figure 4.3). Some participants were more prone to use head movements than others; these movements were found in all but p5 and p9's video data. As stated, these participants rarely exhibited visual cues.

Besides head movements, body movements were sometimes found in the data set, primarily for p8 (9/35 instances). The participants often moved their body in combination with a head tilt. See Figure 4.4 for an example.

Lastly, two facial features were only used once in the entire data set: the cheek squint and the left mouth corner that was raised. Figure 4.5 provides an illustrations of these features.[10]

[10] As one will notice, during the analysis of the 3D data (see Chapter 5), it was found that many recorded frames comprised a high engagement of the *CheekSquint* feature. This was primarily the case for p5. Since this participant squinted their cheeks to such an extent for the entirety of target question utterances, this was not observed during manual annotation: there were barely any 'neutral' facial expressions exhibited by this participant that acted as a comparison between the two.



((a)) Instance of the left corner of the mouth being raised, found in situation 4 condition *NegNeut*.

((b)) Instance of a cheek squint, found in situation 4 condition *PosNeut*.

Figure 4.5: Features that are used only once. *'Is Kim thuis?'* translates to *'Is Kim home?'*

4.3.3 Combinations Between Various Cues

As stated before (see Section 4.1), for this thesis project, the aim was to investigate the visual cues that are present when one asks different types of biased polar questions. Therefore, the blend shape data captured by the 3D depth camera was not sufficient for our data analysis, due to the fact that this data does not document the word order, particles indicating polarity and sentence type used by the participants. Furthermore, although manual annotations in ELAN capture this kind of data very well, the process is categorical and prone to subjective judgements by the annotator. Therefore, the decision was made to include both ELAN annotations and blend shape data during data analysis for this project. However, we have not yet provided an explanation on how these two forms of data were combined.

We are not only interested in the various combinations of facial features that are present over the entire data set. Rather, we are interested to investigate the interaction between the used facial expressions to, for instance, certain combinations of speaker bias and contextual evidence, temporal progression, use of word order, polarity markings and use of tags. Machine learning techniques, such as clustering algorithms, capture the most prominent facial expressions in the first of these two situations very well, but this is not the case for the latter three. However, the manual annotations complement the clustering algorithm, as to process the data in a desired manner.

Python code [Van Rossum and Drake Jr, 1995] was implemented to obtain the frequency counts of the combinations of polarity markings (either visually or spoken), word order and sentence type participants used during their target question utterances. Table 4.1 provides an overview of the combinations that were considered, in which the most frequent combinations are made bold. These combinations are *no-pol/inv/SR*, *neg/inv/SR*, *no-pol/no-inv/SR*, *pos/no-inv/SR* and *neg/no-inv/SR*.[11]

[11] The first of these three variables describes whether positive, negative or no polarity was marked, either visually (with the use of a headshake or headnod) or spoken (using particles

		No Polarity	*Positive*	*Negative*
Inversion	*SR*	**84**	8	**25**
	SR-T1	3	0	9
No Inversion	*SR*	**20**	**30**	**55**
	SR-T1	4	7	7

Table 4.1: Combinations of polarity markings, used word order and sentence found in the data set. The five most frequent combinations are marked bold.

As can be seen, the most recurrent combinations do not contain any tags. This was expected: tags were only used by participants in 30 out of 252 cases. Furthermore, positive polarity was only frequent in target questions containing SVO word order and no tag.

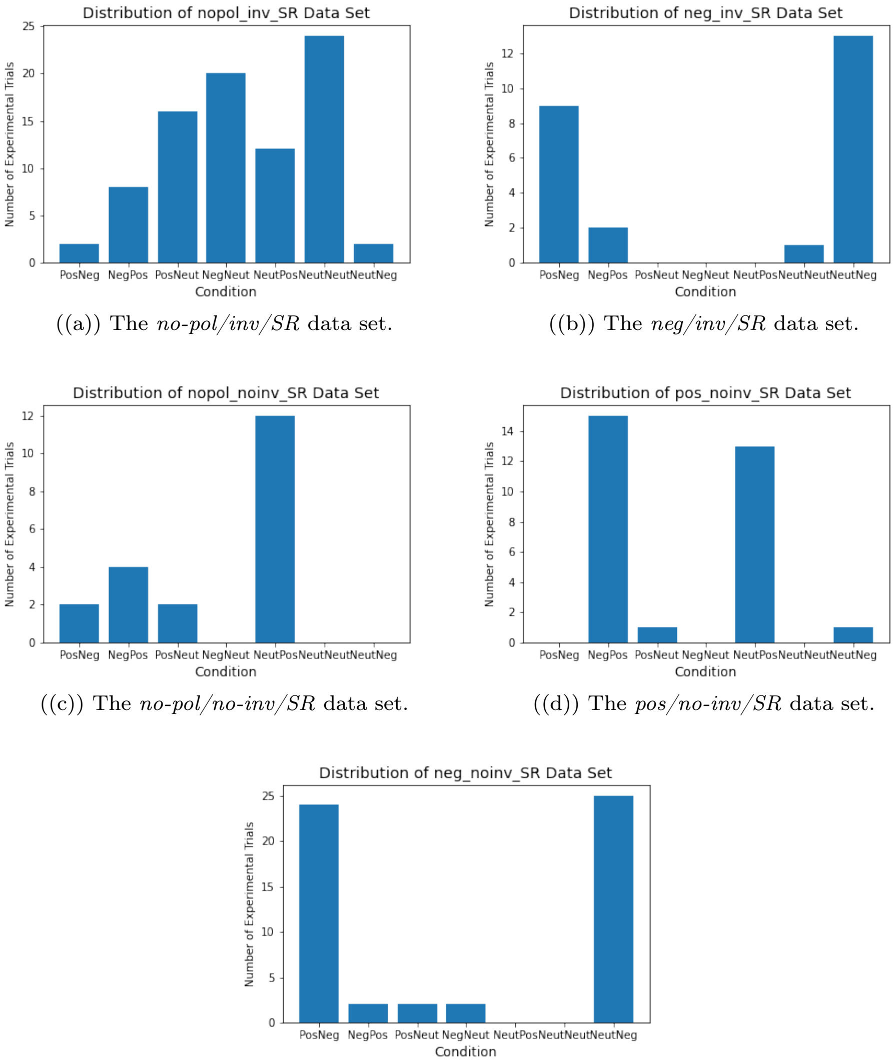
To determine the correspondence to context of these most frequently used question structures, the distribution of values over these structures was obtained. See Figure 4.6 and Figure B.1 for a visualisation. As one can observe, the only data set for which the distribution is relatively evenly spread over conditions is the *no-pol/inv/SR* data set. This is not surprising, given that this data set contains the highest number of data points, compared to the remaining subsets of data. For the data sets comprised of data captured during target question utterances including negative polarity, and no tag, a clear pattern can be established: these data sets comprise mostly of data captured during the *NeutNeg* and *PosNeg* conditions, in which participants were presented with negative contextual evidence in the latter part of the trial. A similar pattern is found for the subset of data containing positive polarity: in the conditions in which confederates provided positive contextual evidence, the participants marked these question utterances with positive polarity, either visually or using particles. For the question utterances containing no polarity markings, analogous findings were not obtained. These results suggest that in cases where participants mark polarity, either visually or spoken, they frequently adopt the polarity of the last heard utterance: the provided contextual evidence.

Additionally, participants heavily favoured using no polarity markers, inversion or tags in the *NeutPos* condition (in comparison to the remaining conditions). A reason for this could be that during these trials, participants aimed to confirm their attained information (for instance, think of *'Dus, Kim is een vegetariër?'* (*'So, Kim is a vegetarian?'*)).

Further, as stated, such prevalent patterns are not found for the distributions of data sets presented in Appendix B. The most probable explanation for this is that the number of data points in these data sets is quite small. Therefore, these data sets are often comprised of data captured during a small number of conditions.

The analysis of the 3D data was performed on both the complete data set, as well as the data sets corresponding to the five most frequent combinations (see Table 4.1). The following chapter provides an extensive overview.

indicating polarity). The second describes whether inversion was used, or not (VSO vs. SVO word order, respectively). The latter describes that only a sentence radical was used.



((a)) The *no-pol/inv/SR* data set.

((b)) The *neg/inv/SR* data set.

((c)) The *no-pol/no-inv/SR* data set.

((d)) The *pos/no-inv/SR* data set.

((e)) The *neg/no-inv/SR* data set.

Figure 4.6: Distribution of the most frequent question structures over conditions.

Chapter 5

Analysis of 3D Data

As Chapter 3 describes, the video and 3D data were first synchronised. Next, the corresponding blend shape data was trimmed as to exclude any noise, after which dimension reduction of symmetric and highly correlated features took place. As a next step, the data set was normalised and ranged. Hence, the steps that were taken during the pre-processing phase now left us with these two data sets. Section 4.3.3 further reports on the most frequent combinations of used word order, polarity markings and sentence type that were present in the data set. The corresponding subsets of the normalised and ranged data sets, as well as the complete data sets, were examined during data analysis.

This chapter describes the subsequent process of the two types of quantitative data analysis that were conducted for this project. First, Section 5.1.1 discusses another reduction of dimensions that was performed (as opposed to the dimension reduction described in Section 3.3.1), to ensure only those features contained in the data set were analysed that are not sensitive to noise. Section 5.2 and Section 5.3 then describe the two methods of data analysis. Section 5.2 recounts the process of plotting the mean blend shape values over time, for specific conditions and situations, whilst Section 5.3 describes the HDBScan clustering algorithm that was implemented.[1]

5.1 Preparing 3D Data for Analysis

5.1.1 Selecting Features

As stated, after pre-processing, the resulting data sets contained the normalised and ranged blend shape values that participants showcased during all of their target question utterances. The dimensions this data set encompassed were already reduced (see Section 3.3.1). Hence, this data set now comprised 39 columns: one for each of the symmetric and combined features (22), the seven features that did not show a significant correlation with their symmetric feature, as well as the ten

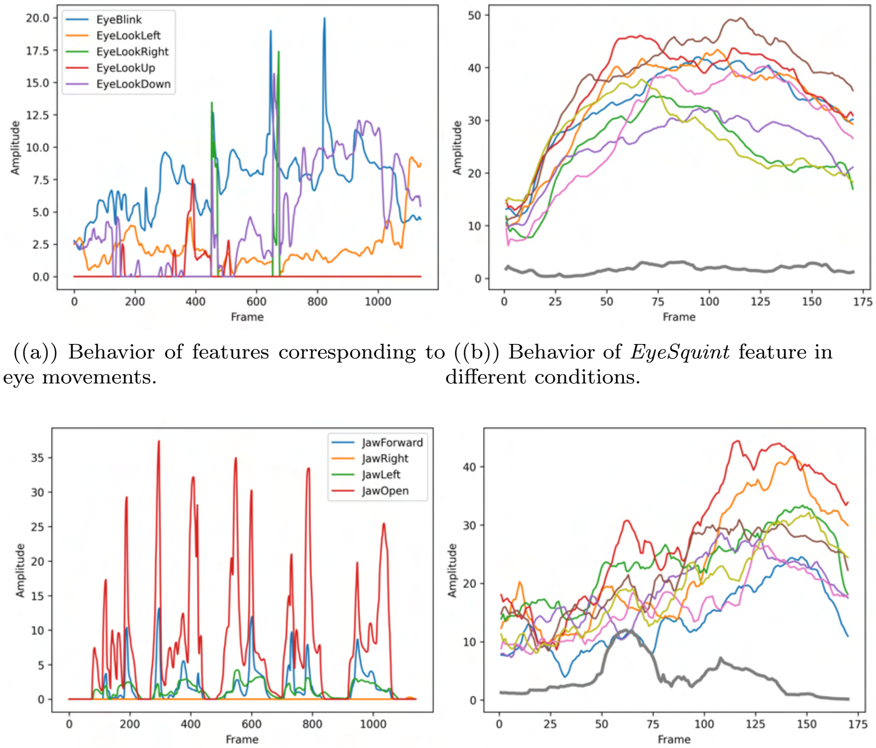
[1] The Python code written for the analysis of the 3D data is publicly available, see: `https://doi.org/10.6084/m9.figshare.c.7054445.v1`

blend shapes that did not have a symmetric counterpart. Even though this already substantially reduced the noise from the data set, the dimensions required additional reduction. This is because not all of the 39 remaining blend shapes are relevant for question marking, and some of these might still produce noise into the data set.

For the BpQ/NGT-experiment, similar to the experiment conducted for this project, Esselink [2022] selected nine of the 39 remaining features that did not introduce noise to the data set. These features were selected in three different ways. Either their relevance in the existing literature was prominent, or their behavior in the declarative statements was not comparable to that of the experimental conditions. Or, this selection was made based on so-called *baseline* recordings, in which the participants mouthed the target question with a neutral facial expression. The visualisation of the measured blend shapes during these recordings provided a means to conclude which facial features were significantly active during the mouthing of the target question. Therefore, these blend shapes were not necessarily part of the facial expressions used during target questions utterances: they produced noise. Since the target questions for the BpQ/NGT-experiment and the current experiment were the exact same, we expected that the features active in the baseline NGT recordings by Esselink [2022] were very similar, or the same for Dutch spoken language. Consequently, the decision was made to take on the selected, not noise-producing, features by Esselink [2022]. What follows is a brief description on why each of the 39 features was or was not selected By Esselink [2022] at this stage. In some cases, the visualisations by Esselink [2022] are provided.

At first, the features *BrowInnerUp* and *BrowOuterUp* were selected, given that raising the eyebrows is the most typical non-manual marker for question marking in sign languages, hence also in NGT (see [Pfau and Quer, 2010, Coerts, 1992] - Section 1.1.1). To be able to compare the results of the BpQ/NGT-experiment [Esselink, 2022] to the current findings, we have selected these two features. Furthermore, since frowned eyebrows are sometimes used as a way to mark questions in NGT [De Vos et al., 2009, Esselink, 2022], and seem to be a prominent question marker in Dutch as well (see [Nota et al., 2021, 2023] - Section 1.1.3), the corresponding blend shape feature *BrowFrown* was also selected for data analysis.

On the contrary, the features *EyePitch*, *EyeRoll*, *EyeYaw*, *HeadPitch*, *HeadRoll* and *HeadYaw* were not selected. These six features correspond to head and body movements. As stated before, camera angle and distance to the 3D depth camera has a small, yet significant effect on the measured blend shape values. The 3D depth camera was calibrated on the participants' faces at the start of the experimental sessions, in which they faced the confederate as they would during the experiment. However, participants did not face the confederate with the same angle, or stand in the same spot during the entirety of the experimental session, especially after a break had been taken. Due to these facts, and the fact that the calibration was only captured at the beginning of each experimental session, the influence of angle and distance on the measurements of head and body movements was too significant. Therefore, since these features were sensitive to noise (see [Esselink, 2022] for a visualisation), they were not selected.



((a)) Behavior of features corresponding to eye movements.

((b)) Behavior of *EyeSquint* feature in different conditions.

((c)) Behavior of features corresponding to jaw movements.

((d)) Behavior of *MouthFrown* feature in different conditions.

Figure 5.1: Visualisation of mean movements over time for certain features. Taken from Esselink [2022]. In figures ((b)) and ((d)), colours indicate: gray - *Declarative*, blue - *PosNeg*, orange - *PosNeut*, dark green - *NeutPos*, red - *NeutNeut*, purple - *NeutNeg*, brown - *NegNeut*, pink - *NegPos*, light green - *Baseline*.

Sometimes, the NMM active during question marking in sign languages include eyecontact with the adressee (see [Pfau and Quer, 2010, Zeshan, 2004] - Section 1.1.1). The corresponding blend shapes are *EyeLookLeft*, *EyeLookRight*, *EyeLookDown* and *EyeLookUp*. 5.1(a) visualises the behavior of these features during the baseline recordings. One can clearly see that they are sensitive to noise (participants were not looking in one specific direction), and therefore these features were not selected.

Previous literature has shown that widening the eyes can be a non-manual marker for question marking in sign languages, even though it has not proven to be a question marker for NGT (see [Pfau and Quer, 2010, Coerts, 1992, De Vos et al., 2009]). However, the *EyeWide* feature was selected: participants frequently had their eyes open wide during their target question utterances, and therefore created the possibility that this feature was a question marker in NGT (this turned out to be the case, see [Esselink, 2022]). Perhaps, this is the case for Dutch as well.

The following features were sensitive to noise, as was apparent from the baseline recordings, and were therefore not selected for data analysis: *EyeBlink*, since this feature was continuously engaged, *CheekPuff*, *TongueOut*, the following features concerning the mouth; *MouthLeft* and *MouthRight* (these features seem to be opposites: when one is engaged, the other is not), *MouthPress*, *MouthFunnel*, *MouthPucker*, *MouthLowerDown*, *MouthUpperUp*, *MouthSmile*, *MouthClose*, *MouthDimple*, *MouthStretch*, *MouthRollLower*, *MouthRollUpper*, and features relating to the jaw; *JawOpen*, *JawForward*, *JawLeft*, *JawRight*.[2] For an extensive review on why these features were not selected, see [Esselink, 2022].

Lastly, five other features were selected for data analysis. The first of which is *EyeSquint*, which often coincided with the *BrowDown* feature. The differences in engagement of this feature between the experimental conditions and the declarative conditions is clearly visible, see 5.1(b).

Next, *CheekSquint* was selected, which again was simultanuously active when *EyeSquint* and *BrowDown* were engaged. Even though this feature was only found once while manually annotating the current data (see Figure 4.5), there was a distinct difference between its behavior between the declarative and experimental conditions for the comparable NGT data. For similar reason, *NoseSneer* was selected for data analysis.

Finally, *MouthShrug* and *MouthFrown* were selected. Even though both of these features are slightly engaged during the baseline target question utterances, this effect was very minimal. For the *MouthShrug* blend shape, the differences at the start of the declarative condition did not differ significantly from the experimental conditions. However, this difference is clearly visible at the end of the utterances, and therefore this feature was selected. The *MouthFrown* feature was selected for the same reason. See 5.1(d) for a visualisation of the latter.

Concluding: the dimensions were reduced by selecting only those nine out of the 39 features that were not sensitive to noise. The selected features include *BrowInnerUp*, *BrowOuterUp*, *BrowDown*, *EyeWide*, *EyeSquint*, *CheekSquint*, *NoseSneer*, *MouthShrug* and *MouthFrown*.

5.1.2 Creating the Final Data Sets

Now that the final dimension reduction had taken place, the two created data sets, containing normalised and ranged measured blend shape values, were ready to be altered as to fit the data analysis methods used in this project. The two subsequent sections describe this process.

As a first step, the thirty features that were not selected in the previous procedure were removed from the normalised and ranged data sets. The resulting data sets comprised 14 columns, namely the nine selected features and the IDs describing the

[2] The features relating to the mouth and the jaw were often active when participants mouthed the target questions during the baseline recordings (5.1(c) visualises the latter). Think of letters such as *b* and *m*, causing *MouthPress* to be engaged, or words such as *'uur'* and *'open'*, causing the participants to pucker their mouth.

participant, scenario, condition, frame and temporal window, and 22.350 samples.

Esselink [2022] found that the HDBScan clustering algorithm, described in Section 5.3, did not provide sufficient results on the data set containing the normalised blend shape values. Therefore, no conclusions could be drawn from the resulting clusters (see Section 5.3 for a more detailed explanation). To avoid this problem, the normalised data was categorised and put into so-called bins: samples that have roughly the same blend shape value now belong to the same bin. The decision was made to categorise the data into eight of these bins: values $0 - 5$ were set to 0, values $5 - 10$ were set to 7.5, values $10 - 15$ were set to 12.5, values $15 - 20$ were set to 17.5, values $20 - 30$ were set to 25, values $30 - 45$ were set to 37.5, values $45 - 65$ were set to 55 and values $65 - 100$ were set to 82.5 (see Table C.1 - $B2$). Multiple variations of categorisation were implemented, in which the bins contained various upper- and lower bounds, as well as sizes. Appendix C describes these implementations and further accounts for the decision made to choose the bins for which we have obtained the final results.

Next, subsets of the categorised data set and the ranged data set were extracted, based on the most frequent combinations of used word order, polarity markings and tag use (see Section 4.3.3). As a reminder, the combinations were *no-pol/inv/SR*, *neg/inv/SR*, *no-pol/no-inv/SR*, *pos/no-inv/SR* and *neg/no-inv/SR*. First, the data captured during the declarative trials, in which the participants transformed the target questions into declarative statements, were removed from the data sets. The resulting data sets now comprised 304 trials, 14 columns and 19.921 rows. Subsequently, the subsets of these data sets, corresponding to the most frequent combinations, were extracted and saved as a new data set (see Section 4.3.3). The resulting data sets comprise 4.803, 1.556, 1.231, 1.815 and 3.465 samples, respectively.

Concluding, we have ended up with twelve data sets total. These are the complete ranged data set, the complete normalised and categorised data set, and their five subsets, respectively. The ranged data sets were used for visualising the mean blend shape values over time (see Section 5.2), whilst the HDBScan clustering algorithm was performed on the normalised and categorised data sets.

5.2 Visualisation of Temporal Progression

The 3D depth camera captures an abundant amount of data: the level of activity of all features for every frame for the duration of the recording is measured. To be able to draw conclusions from this collection of data, it would be helpful to visualise the activity of these features over time. Accordingly, this was the first step that was taken during data analysis.

Due to the fact that not every target question utterance has the exact same length, in order to visualise the behavior of features over time, the ranged data set needed to be normalised on time. To accomplish this, the column in the data set containing the unique frame IDs for each trial is adapted; these values were normalised, using the following equation:

	ID
v_1	1
v_2	2
v_3	3
$\vdots$	$\vdots$
v_4	54
v_5	55
v_6	56

((a)) Original Frame IDs

	ID'
v'_1	1.11
v'_2	2.22
v'_3	3.33
$\vdots$	$\vdots$
v'_4	59.94
v'_5	61.05
v'_6	62.16

((b)) Normalised Frame IDs

Table 5.1: Comparison between original and normalised unique frame IDs. The mean duration d_μ, calculated over all videos, equals 62.16 in this example.

$$f'_i = \frac{f_i}{d_i} \times d_\mu \tag{5.1}$$

In this formula, f_i stands for the unique frame ID of the video with ID i, f'_i is its updated normalised value, d_i is the duration in frames of the video with ID i, and d_μ is the mean duration in frames of all videos. Hence, the unique frame IDs were divided by the duration of the corresponding target question utterance, and afterwards multiplied by the mean duration of target question utterances. The resulting graph now has this mean duration as its limit on the x-axis. Table 5.1 provides an example, illustrating the workings of the equation above.

After the data had been normalised on time, the gradient of the mean blend shape values was visualised. This was implemented for the conditions per feature, the features per condition, the features per scenario and condition (hence for the features per trial), the participants per condition and feature, the participants per feature, condition and scenario (hence for the participants per feature and trial), and the scenarios per feature. Section 6.1 provides an overview of the most significant results, in addition to more results being presented in Appendix D.

5.3 HDBScan Clustering Algorithm

Whilst the first method of data analysis equips us with the means to draw conclusions about the behavior of features over time, it does not provide us with information on the most prototypical facial expressions. This is why we have implemented the HDBScan clustering algorithm for this project, which allows for a frame-by-frame analysis of the captured data. Clustering algorithms are unsupervised machine learning techniques; they take on unlabelled data and discover new insights without any other input. For clustering algorithms specifically, the unlabelled data is divided into clusters, which are groups of data points that are bundled together because of their similarities.

Note that there are dozens of clustering algorithms, each with their own advantages and disadvantages (think of K-means, HDBScan, BIRCH, GMM, OPTICS).

For the purpose of this project, the decision was made to implement the HDBScan clustering algorithm. Esselink [2022] implemented both the K-means and the HDBScan algorithm. The K-means algorithm takes on a specified k, indicating the number of clusters that will be created. It then partitions the data into k clusters, calculates the center point of these clusters, and then re-assigns all data points to the cluster for which the center is closest. This process repeats until the data points are stable within their own cluster. It was found that the K-means algorithm did not provide informative results for the NGT data. First, given that all data points must belong to a cluster, the data points producing noise were also assigned to their most probable cluster. Second, it was difficult to determine the value for k that would lead to optimal results. Consequently, Esselink [2022] implemented the HDBScan algorithm, which did lead to informative results. Given that Esselink [2022] found that the HDBScan algorithm provided insightful information on the most prototypical facial expressions found in the NGT data set, and the fact that the data collected for this project has the same structure as that data, the decision was made to only implement the HDBScan clustering algorithm here.

The HDBScan algorithm is both a density based and a hierarchical clustering algorithm. The most prominent reasons for choosing this algorithm, in comparison to others, is that the HDBScan algorithm allows for different densities within clusters, different cluster sizes, and most importantly, does not require every data point to be assigned to a cluster. Hence, data points that do not belong to a particular cluster are classified as noise. Even though, during pre-processing, we significantly reduced the dimensions present in our data set, in addition to removing as much noise as possible by discarding irrelevant frames (see Section 3.2, Section 3.3 and Section 5.1.1), it was quite likely that our data set still contained noise. This is because not all facial features that were engaged during the target question utterances were contained in the prototypical facial expressions for question marking. Additionally, since the collected data contained a large number of points, it was to be expected that the most prototypical facial expressions did not occur with the same frequency. Therefore, given that HDBScan allows clusters to have different sizes and densities, the use of this algorithm proves to be beneficial.

The algorithm works as follows. First, for every data point, a *core distance* is calculated, which equals the distance between the data point and its k-nearest neighbor.[3] Next, a division is made between dense and sparse areas within the data space. The dense areas contain data points with relatively small core distances, while the sparse areas' data points have relatively large core distances (therefore, HDBScan is a density based algorithm). The data points contained in the dense areas are now kept at the same distance, while the data points within the sparse areas are pushed further away from each other. The reason for this is to avoid that two clusters are accidentally seen as one. This might happen when one data point containing noise is positioned between two dense clusters, acting as a bridge between the two [McInnes et al., 2016]. As a next step, a hierarchical tree is created, containing all data points. To accomplish this, first, a distribution is obtained of all data points included in the data set. A threshold can then be set, indicating that the dense areas within this distribution that peak above this threshold are

[3] Here, k equals the value set for the parameter *Minimal Samples*, see below.

clusters. However, specifying this threshold is a difficult task. Consequently, the large peaks within this distribution are obtained for a large number of thresholds, after which a hierarchical tree is created [McInnes et al., 2016]. For each of the internal nodes, it is checked whether this contains more or less data points than the sum of its children nodes. If this is less, the children nodes are seen as separate clusters. If this is more, the children nodes and the parent node are combined into one cluster. This process repeats until no new clusters are found. The algorithm now extracts the clusters and returns them.

Using the HDBScan algorithm, one can set two parameters as desired: *Minimal Cluster Size* and *Minimal Samples*. The first describes the minimal number of data points a cluster needs to contain, while the second indicates the minimal number of data points required to be in the close proximity of another data point, in order for the latter to not be classified as noise. A low value for these parameters results in a more lenient clustering, while a high value produces only strict clusters [McInnes et al., 2016]. Based on trial and error, these parameters were set to 35 and 390 for the complete data set, respectively. However, using these parameters, it was found that the euclidean distance between some clusters was somewhat low (see Section 6.2). Therefore, following Esselink [2022], a few *super-clusters* were further extracted from this data set, by means of a more strict clustering: the parameters were set to 60 and 600, respectively. This allowed us to both find the most prototypical facial expressions in the data, as well as more fine-grained clusters, corresponding to facial expressions that occurred less often.

In order to compare the clusters from the complete data set with those formed within its subsets, we were less interested in the fine-grained differences between facial expressions within the latter data sets. Therefore, for each subset of the complete data set, the aim was also to extract a few super-clusters, corresponding with the prototypical facial expressions found specifically within this data. The *Minimal Cluster Size* and *Minimal Samples* parameters were set as follows, respectively: 35 and 200 for the *no-pol/inv/SR* data set, 15 and 100 for *neg/no-inv/SR*, 28 and 150 for *pos/no-inv/SR*, 12 and 90 for *neg/inv/SR*, and 15 and 100 for the *neg/no-inv/SR* data set. One can see that, in order to obtain optimal results, the clustering algorithm required more leniency for the smaller data sets, since these comprised significantly less data points than others, even though the clustering on these data sets was relatively more strict in comparison to the fine-grained clusters obtained from the complete data set.

As was previously stated, Esselink [2022] found that performing the HDBScan algorithm on the normalised data set did not provide sufficient results, whilst this was the case for the corresponding categorised data. The reason for this was that the formed clusters within the normalised data set were now based on dense areas, containing only their immediate neighboring samples. Consequently, the resulting clusters were particular to participants and generalisations could not be made from these clusters. Accordingly, as was discussed in Section 5.1.1, the data was categorised. But, as Appendix C reports, these results do not provide a space that is optimal for drawing conclusions. Therefore, the results presented in the subsequent chapter are based on a *preliminary* analysis.

Chapter 6

Results of 3D Data Analysis

In this chapter, the results arising from the two methods of data analysis are reviewed. Section 6.1 describes the behavior of certain features over time and over conditions, as well as the features some participants preferred to use over others. Additionally, Section 6.2 discusses the clusters found within the normalised and categorised data sets and the accompanying facial expressions. Note that both of these sections only review the most prominent results. Additional visualisations of results can be found in appendices Appendix D and Appendix E.

6.1 Visualisation of Temporal Progression

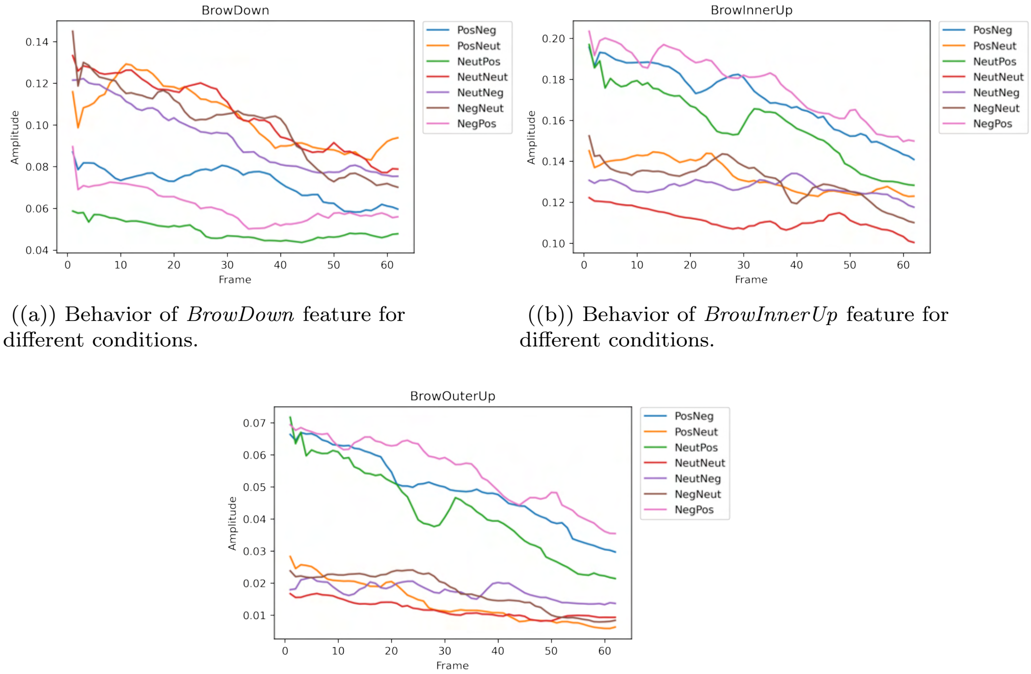
As Section 5.2 has previously described, the gradient of the mean blend shape values was visualised, both for the full ranged data set and its five subsets. For this thesis project, we have visualised the relation between the condition and the accompanying features, the features per condition, the features with the accompanying scenario and condition (thus - the features per trial), the participants and the behavior of the used features per condition, the participants and the accompanying features used per condition and scenario (trial), as well as the scenarios per feature.[1] This section reviews the resulting visualisations for each of the implementations in more detail, for both the complete data set and its five subsets. Note that in many cases, because the mean blend shape value was taken over all target question utterances, the mean engagement of features is often considerably low. The figures in this section depict patterns that are clearly visible, but sometimes seem more grandiose than they are, due to the scale of the visualisations.

6.1.1 Complete Ranged Data Set

Condition per Feature

Firstly, we will take a look at the graphs in which for each condition, the gradient of the behavior of specific features was plotted. Figure 6.1 and Figure 6.2 present five

[1] Of course, not all obtained figures are presented in this report. Therefore, the remaining figures are publicly available, see `https://doi.org/10.6084/m9.figshare.c.7054445.v1`



((a)) Behavior of *BrowDown* feature for different conditions.

((b)) Behavior of *BrowInnerUp* feature for different conditions.

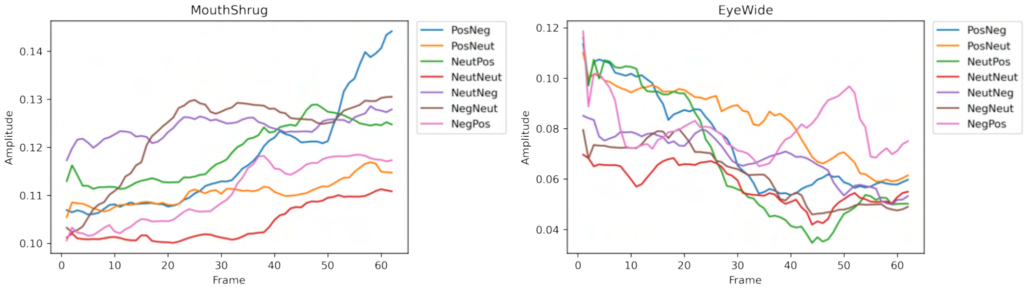
((c)) Behavior of *BrowOuterUp* feature for different conditions.

Figure 6.1: Visualisation of mean movements over time for features describing eyebrow movements.

of these visualisations, for the features *BrowDown*, *BrowInnerUp*, *BrowOuterUp*, *MouthShrug* and *EyeWide*.

One can see that for the three features concerning the eyebrows, there seems to be a division within the conditions: conditions *PosNeg*, *NegPos* and *NeutPos* all behave similarly - they are either all relatively highly engaged (for conditions *BrowInnerUp* and *BrowOuterUp*), or not (for *BrowDown*), whilst the four remaining conditions seem to behave in the opposite way. It is not surprising that this pattern is present for the features describing eyebrow movements. The *BrowInnerUp* and *BrowOuterUp* features are highly correlated (see Table 3.1). Even though their correlation was not so high as to combine them into one feature, this number is still above 0.94. Therefore, it is not strange that these features showcase similar patterns. Additionally, given that one can not simultaneously raise and lower their eyebrows, it is not surprising that the *BrowDown* feature acts as an antagonistic feature to the other two blend shapes. Furthermore, the use of each of the three features over time significantly decreases, the further we get into the target question utterance.

On the other hand, the behavior of the *MouthShrug* and *EyeWide* feature over time seems to be consistent for all conditions. For the first, the participants' use



((a)) Behavior of *MouthShrug* feature for different conditions.

((b)) Behavior of *EyeWide* feature for different conditions.

Figure 6.2: Visualisation of mean movements over time for features *MouthShrug* and *EyeWide*.

increases. In some conditions, this increase is more steep than in others (compare for instance the *PosNeg* and *PosNeut* condition). Furthermore, the increase of use for this feature does not occur at the same time for every condition. See condition *NegNeut*, for which the majority of this increase happens early on, while this happens quite some time later for the *NeutNeut* and *PosNeg* conditions, for instance. In contrast, the behavior of the *EyeWide* feature is the opposite: the participants' use decreases over time. This decrease is quite consistent, apart from the fact that in some conditions the *EyeWide* feature is generally more engaged than others; for instance, compare the *PosNeut* and *NeutNeut* conditions. One remarkable finding is that for the *NegPos* condition, the engagement of the feature peaks at the end of the target question utterances.

Similar visualisations for the four remaining features are presented in Figure D.1. Here, all features are engaged, but such a clear pattern can not be established.

Feature per Condition

The visualised temporal progression of blend shape measurements for each condition, are unfortunately less informative. The gradient of the features' engagement is considerably consistent over time. Two main takeaways from these visualisations are that the *EyeSquint*, and often the *BrowInnerUp*, feature are relatively highly engaged in all situations, in comparison to the remaining seven features, in particular *BrowOuterUp* and *MouthFrown* (see for instance Figure 6.3). Furthermore, the use of the *BrowInnerUp*, *BrowOuterUp* and *BrowDown* feature decreases over time. Lastly, the aforementioned peak of the *EyeWide* feature in the *NegPos* condition is clearly visible, see Figure 6.3. Again, Figure D.2 contains the visualisations for all other conditions, in which one can see that the use of features over time does not contain much variation.

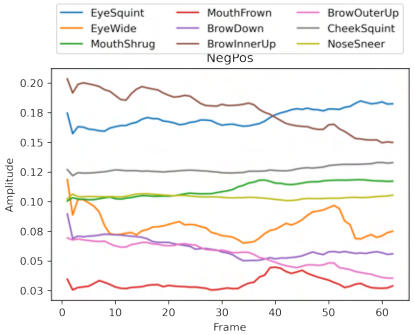
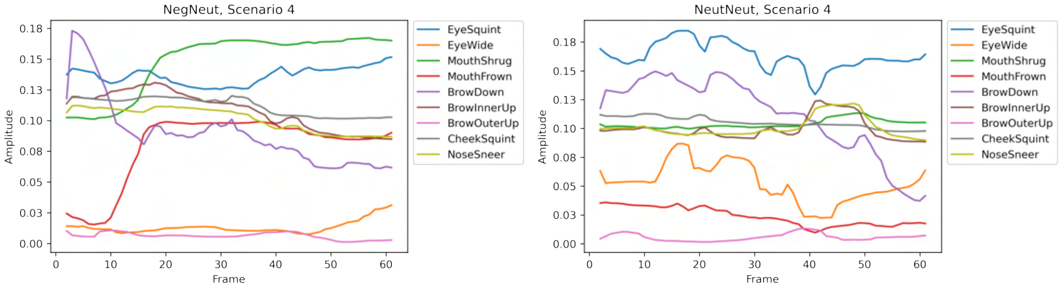


Figure 6.3: Visualisation of mean movements over time for all features in the *NegPos* condition.



((a)) Behavior of all features during scenario 4 condition *NegNeut*.

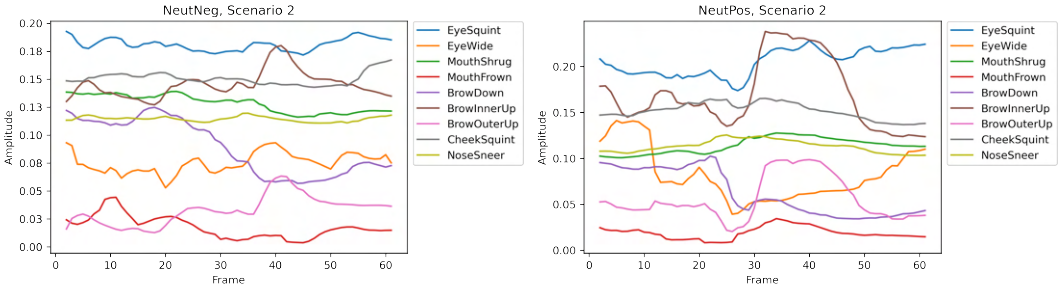
((b)) Behavior of all features during scenario 4 condition *NeutNeut*.

Figure 6.4: Visualisation of mean movements over time for all features in scenario 4 conditions *NegNeut* and *NeutNeut*. A decline in engagement for feature *BrowDown* is clearly visible.

Feature per Condition and Scenario

The behavior of the features over time was not only visualised for each condition, but also for each condition and scenario. This leads to some more insights. First, the use of the *BrowDown* feature decreases over time, similarly to what was found before. In about half of the scenarios, this pattern is clearly visible, while for the other half, it is either slightly visible or the use of the feature does not vary over time. Figure 6.4 provides two examples of this decline in gradient. Figure D.3 presents two more illustrations of this pattern.

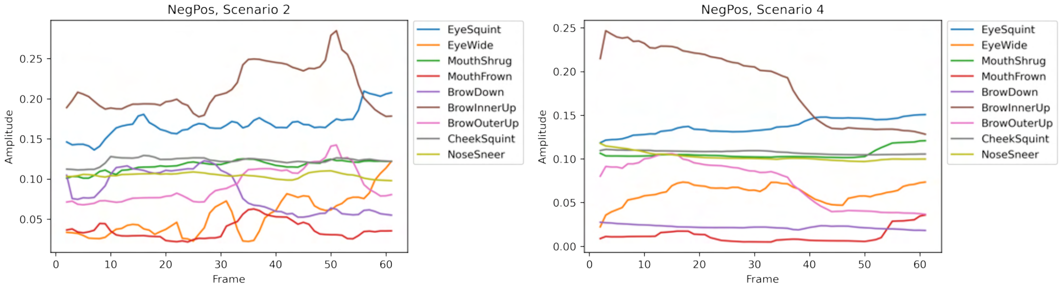
Not only is the behavior of the *BrowDown* feature more visible in these types of graphs, this is also the case for the other features describing eyebrow movements: *BrowInnerUp* and *BrowOuterUp*. We find that generally, when one of these features is engaged, this is also the case for the other, even though *BrowOuterUp* is always less engaged than its counterpart *BrowInnerUp*. See for instance Figure 6.5,



((a)) Behavior of all features during scenario 2 condition *NeutNeg*.

((b)) Behavior of all features during scenario 2 condition *NeutPos*.

Figure 6.5: Visualisation of mean movements over time for all features in scenario 2 condition *NeutNeg* and *NeutPos*. Similar patterns for features *BrowInnerUp* and *BrowOuterUp* are clearly visible.



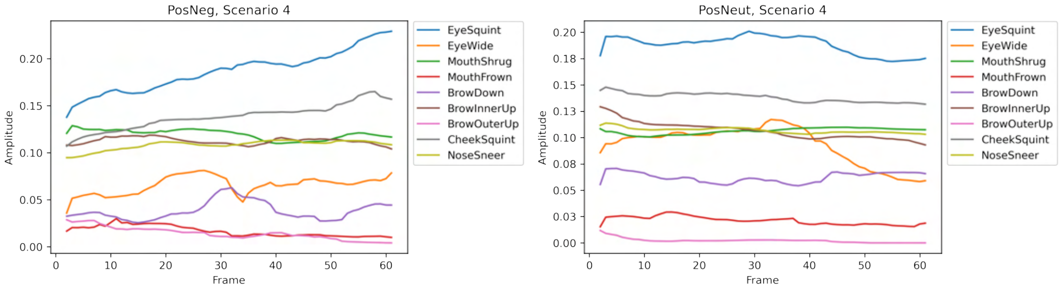
((a)) Behavior of all features during scenario 2 condition *NegPos*.

((b)) Behavior of all features during scenario 4 condition *NegPos*.

Figure 6.6: Visualisation of mean movements over time for all features in condition *NegPos* situation 2 and 4. Feature *BrowInnerUp* is highly engaged.

in which two different conditions for one scenario are presented. One can see that the behavior of the *BrowInnerUp* and *BrowOuterUp* feature over time is similar, even though their amplitude differs. This result parallels the one described by Figure 6.1, in which is shown that these features showcase similar patterns in the same conditions. In addition, similarly to these results found before, the *BrowInnerUp* feature is frequently present in the *NegPos*, *PosNeg* and *NeutNeg* conditions, even though it is sporadically engaged in other conditions as well. 6.5(a) showcases this behavior, as well as Figure 6.6.

Lastly, the presence of the two features *CheekSquint* and *EyeSquint* is consistent for all trials in the data set. In each trial, the engagement of the *CheekSquint* feature is average, in comparison to the remaining features. Furthermore, the gradient of the behavior of this feature never peaks, declines, nor increases in a significant way.



((a)) Behavior of all features during scenario 4 condition *PosNeg*.

((b)) Behavior of all features during scenario 4 condition *PosNeut*.

Figure 6.7: Visualisation of mean movements over time for all features in situation 4 conditions *PosNeg* and *PosNeut*.

Similarly, the *EyeSquint* feature is often relatively relatively highly engaged during the target question utterances. But, particularly in scenario 4, this feature is used even more. Figure 6.7 showcases these patterns, as well as 6.4(b).

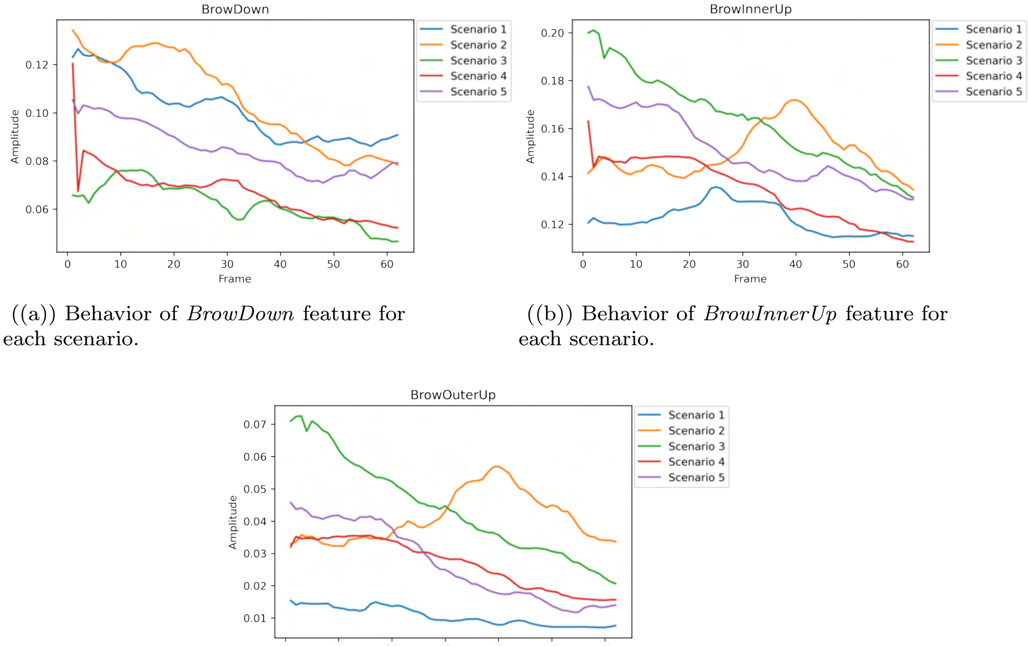
Scenario per Feature

Next, the features were visualised by plotting their temporal progression for each of the scenarios. Quite a lot of similarities are found between the first type of visualisation (conditions per feature) and the current visualisations. This confirms the results reviewed previously in this section.

To start, the *BrowInnerUp* and *BrowOuterUp* features showcase similar patterns to those that were described before. The use of both features decreases over time, with a peak that is seen in scenario 2. Again, the severity of engagement differs between situations: the features are more highly engaged in the third situation compared to the fifth situation. Furthermore, a decrease in use over time can also be found for the *BrowDown* feature, as was further found previously. The visualisations can be found in Figure 6.8.

Not only does the engagement of these features showcase similar patterns as previously described, this is also the case for the *MouthShrug* and *EyeWide* feature. Over time, the participants' use of these features increases and declines, respectively. For the first, this is constant over all situations, except the fifth, in which this feature is engaged to a higher extent. Likewise, for the *EyeWide* feature, the decrease over time is constant for all situations but three. The described visualisations are presented in Figure 6.9.

The graphs corresponding to the four remaining features are presented in Figure D.4. Again, the features are engaged in all scenarios. However, these features seem to be used the least in scenario 4. This can be explained by the fact that scenario 4 contained relatively short target question utterances (*'Is Kim thuis?* (*'Is Kim home?*)).

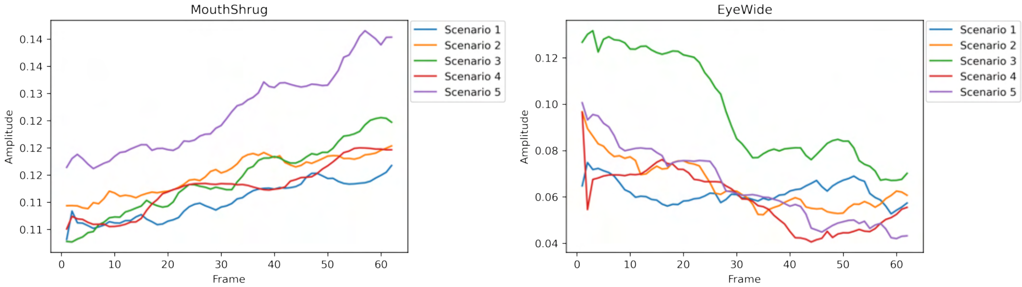


((a)) Behavior of *BrowDown* feature for each scenario.

((b)) Behavior of *BrowInnerUp* feature for each scenario.

((c)) Behavior of *BrowOuterUp* feature for each scenario.

Figure 6.8: Visualisation of mean movements over time for features *BrowDown*, *BrowInnerUp* and *BrowOuterUp* in each scenario. Generally, the use of these features decreases over time.



((a)) Behavior of *MouthShrug* feature for each scenario.

((b)) Behavior of *EyeWide* feature for each scenario.

Figure 6.9: Visualisation of mean movements over time for features *MouthShrug* and *EyeWide* in each scenario.

Feature per Participant

Finally, we have plotted the use of features per condition by participants over time. This was done to establish whether certain participants had a predisposition to use certain features more than others. This did end up being the case. We will now go over each feature briefly and review which participants prefer the use of this feature in comparison to others.

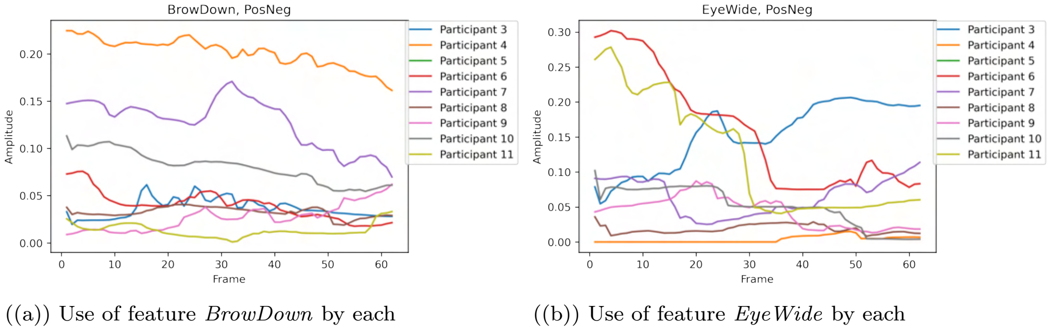
The *BrowDown* feature is primarily used by participants p4 and p7, see 6.10(a). This is consistent with the preliminary analysis of the manual annotations (see Section 4.3.2). Besides, these participants showcase a high engagement of this feature at the start of the target question utterance, whilst this declines over time. This, again, is consistent with the results found as previously described.

Furthermore, p3, p6, p7 and p11 demonstrate the highest engagement of the *BrowInnerUp* feature during target question utterances. Again, this is in congruence with the findings as described in Section 4.3.2. For some participants, the results parallel those described previously: the use decreases over time. For others, this pattern is not visible. See 6.11(a) for a visualisation.

Additionally, the *BrowOuterUp* feature is used by the same participants as the *BrowInnerUp* feature, albeit p7 does not use the current feature to the same extent. Again, this is not surprising, since these features are highly correlated. Further, we can establish that the pattern is the same as before: if the feature is significantly engaged, the use of this feature will decrease during the target question utterance. The gradient of the blend shape measurements over time are presented in 6.11(b) and 6.11(c) for two conditions.

Besides, participants p3, p6 and p11 demonstrate the *EyeWide* feature the most. This visual cue was not observed often during manual annotation (see Section 4.3.2). One can view the gradient of this feature for the *PosNeg* condition in 6.10(b).

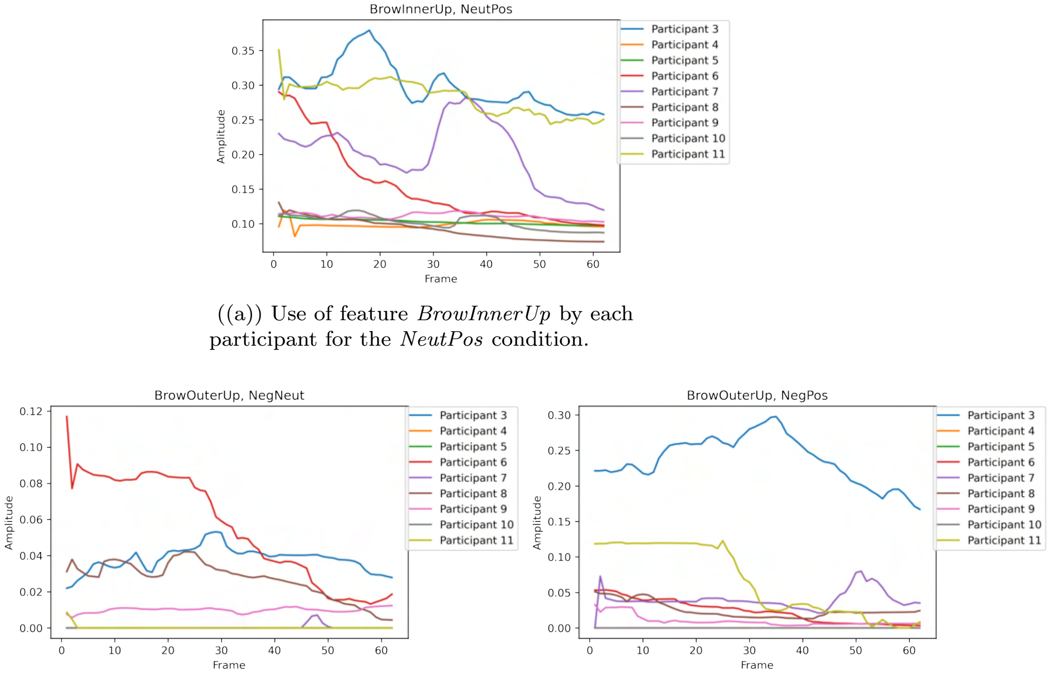
Lastly, four features are primarily used by only one participant. For the *MouthFrown* and *MouthShrug* feature, this is p11, whilst this is p5 for the *NoseSneer* and *CheekSquint* blend shape. Figure 6.12 present an instance of these findings.



((a)) Use of feature *BrowDown* by each participant for the *PosNeg* condition.

((b)) Use of feature *EyeWide* by each participant for the *PosNeg* condition.

Figure 6.10: Visualisation of mean movements over time for features *BrowDown* and *EyeWide* condition *PosNeg* for each participant.



((a)) Use of feature *BrowInnerUp* by each participant for the *NeutPos* condition.

((b)) Use of feature *BrowOuterUp* by each participant for the *NegNeut* condition.

((c)) Use of feature *BrowOuterUp* by each participant for the *NegPos* condition.

Figure 6.11: Visualisation of mean movements over time for features *BrowInnerUp* and *BrowOuterUp*, in conditions *NeutPos*, *NegNeut* and *NegPos*, for each participant.

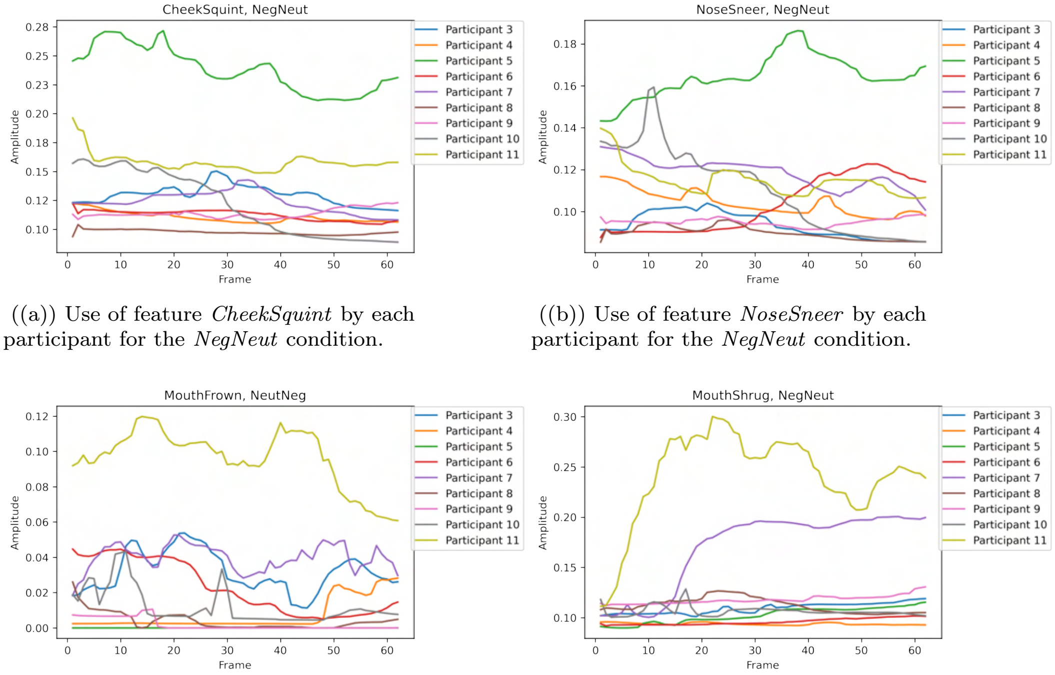
6.1.2 Subsets of Ranged Data Set

As has been previously stated, not only was this type of quantitative data analysis implemented on the complete ranged data set, this was also done for the five subsets corresponding to the most frequent combinations of used word order, polarity markings and sentence type (see Section 4.3.3 and Section 5.2). In a similar vein compared to the previous section, for each of the six variations of visualisations we briefly describe whether similar results were obtained for the five subsets.

Condition per Feature

First as stated, it was found that the *BrowInnerUp* and *BrowOuterUp* feature behave similarly: they are highly engaged in the *NegPos*, *PosNeg* and *NeutNeg* conditions, whilst this is not the case for the other conditions. On the contrary, the *BrowDown* feature acts the opposite; it is frequently engaged in the *NeutNeut*, *NeutPos*, *PosNeut* and *NegNeut* conditions. Similar results are obtained for the *no-pol/no-inv/SR*[2] and *pos/no-inv/SR* data sets, albeit this pattern is more visible

[2] As a reminder, the notation used here refers to the use of polarity markers (positive, negative or no polarity), the used word order (inversion or not) and the sentence type (SR (a sentence radi-



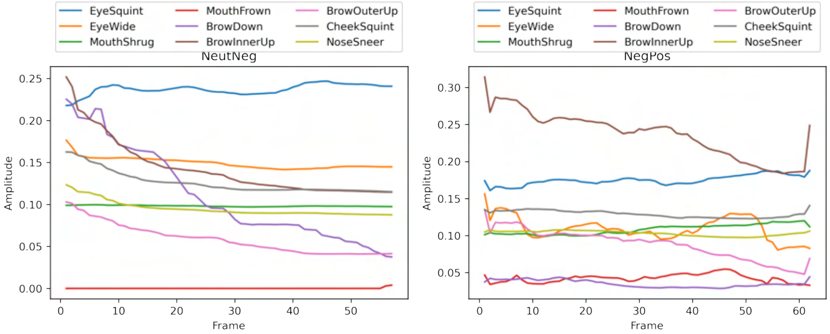
((a)) Use of feature *CheekSquint* by each participant for the *NegNeut* condition.

((b)) Use of feature *NoseSneer* by each participant for the *NegNeut* condition.

((c)) Use of feature *MouthFrown* by each participant for the *NeutNeg* condition.

((d)) Use of feature *MouthShrug* by each participant for the *NegNeut* condition.

Figure 6.12: Visualisation of mean movements over time for features *CheekSquint*, *NoseSneer*, *MouthShrug* and *MouthFrown*, in the *NegNeut* and *NeutNeg* conditions, for each participant.



((a)) Behavior of all features for the *NeutNeg* condition (obtained from the *no-pol/inv/SR* data set).

((b)) Behavior of all features for the *NegPos* condition (obtained from the *pos/no-inv/SR* data set).

Figure 6.13: Visualisation of mean movements over time for all features in the *NeutNeg* and *NegPos* condition.

in the latter subset. The *neg/no-inv/SR*, *neg/inv/SR* and *no-pol/inv/SR* data sets do not show this pattern in a significant manner.

Furthermore, as reported, the *MouthShrug* blend shape measurements generally increase over time for the full data set, whilst the opposite occurs for the *EyeWide* feature. Again, this is also the case for the *pos/no-inv/SR* and *no-pol/no-inv/SR* data set, as well as the *no-pol/inv/SR* data set, albeit to a lesser extent. Additionally, the first showcases this pattern to the most. Further, the *EyeWide* feature demonstrates a peak in the *NegPos* condition, which is slightly visible for the *pos/no-inv/SR* data set. Figure D.5 provides the visualisation corresponding to these findings.

Feature per Condition

For all but the *no-pol/no-inv/SR* data set, the obtained results by plotting the features per condition parallel those of the full data set: the use of the *BrowDown* feature, as well as the *BrowInnerUp* and *BrowOuterUp* feature decreases over time. This was particularly the case for the *no-pol/inv/SR* data set, which 6.13(a) visualises.

Additionally, the *BrowInnerUp* and *EyeSquint* feature are consistently highly engaged during target question utterances for the subsets of the data set, just like is the case for the full data set. Further, in the *pos/no-inv/SR* data set, the peak of the *EyeWide* feature in the *NegPos* condition is again visible. Both of these findings are presented in 6.13(b).

Feature per Condition and Scenario

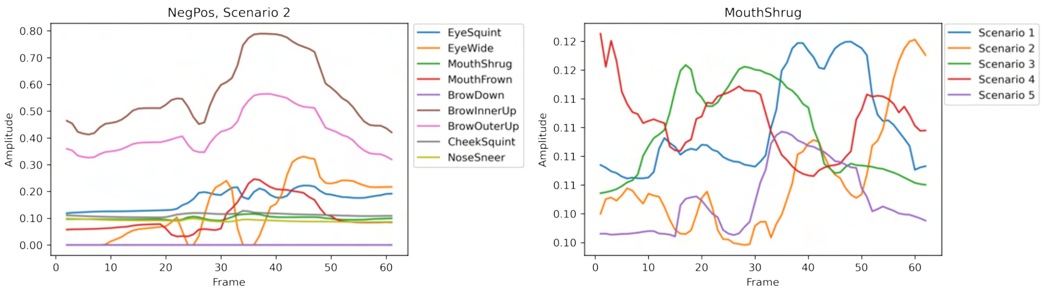
As can be seen in the previous section, for the full data set, the visualisations of the features per trial confirm the results found in the preceding visualisations. One additional observation that is made, is that the *CheekSquint* and *EyeSquint* features are consistently present in the data set without showcasing any in- or decreases, albeit the first to a lower extent than the latter. Similar results are obtained for the five subsets of the full data set: the use of the *BrowDown* feature declines over time, the *CheekSquint* and *EyeWide* feature are consistently present in the data, the *BrowInnerUp* and *BrowOuterUp* feature show simultanuous behavior, and are mainly present in the *NegPos*, *PosNeg* and *NeutNeg* conditions.

A clear example of the behavior of the *BrowInnerUp* and *BrowOuterUp* features over time is presented in 6.14(a). One can see that both features demonstrate the same pattern, even though *BrowInnerUp* is engaged to a higher extent. Furthermore, both features are used significantly more than the remaining seven.

Scenario per Feature and Feature per Participant

Finally, the results obtained from these last styles of visualisations mostly confirm the results found for the full data set. Hence, the use of the features describing eyebrow movements decrease over time, as well as the use of the *EyeWide* feature.

cal) or SR-T1 (a sentence radical plus a tag)). Hence, the *no-pol/no-inv/SR* data set corresponds to that subset containing only those data points corresponding to target question utterances in which no polarity markings were used, inversion was not used and further, no tag was used.



((a)) Behavior of all features for the *NegPos* condition, scenario 2 (obtained from the *pos/no-inv/SR* data set).

((b)) Behavior of the *MouthShrug* feature in each scenario (obtained from the *neg/inv/SR* data set).

Figure 6.14: Visualisation of mean movements over time.

On the contrary, participants use the *MouthShrug* feature more towards the end of the target question utterances. The only difference between the previous results and the current results is that for the *neg/inv/SR* data set, the *MouthShrug* feature does not show a clear increase over time; this feature is active during the entirety of every scenario. See 6.14(b) for a visualisation.

Further, the predispositions participants have to use certain features over others were similarly obtained for the subsets of the full data set. We therefore do not provide these visualisations.

6.2 HDBSCan Clusters

The data analysis for this project includes, aside from visualising the temporal progression of features' activity, the implementation of the HDBScan clustering algorithm, as stated. This section goes over the resulting clusters, for the full normalised and categorised data set, as well as its five subsets (see Section 4.3.3 and Section 5.1.2). Note that all blend shape values presented in this section, as well as Appendix E, are multiplied by 100 for readability.

6.2.1 Complete Normalised and Categorised Data Set

Clusters and their Accompanying Facial Expressions

The HDBScan clustering algorithm, as described in Section 5.3, was first performed on the complete normalised and categorised data set. Table 6.1 shows the seven resulting clusters, as well as the three super-clusters [Esselink, 2022] that were extracted. Furthermore, this table reports on the distribution of the clusters over participants, as well as the mean values of the nine selected features within these clusters. We will examine these results more closely.

First, as can be seen in Table 6.1, 7 clusters were formed, apart from the 'cluster' which contains all data points classified as noise, which comprises 28% of the full

	S_t	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N	28									
C_1	2	35	72	0	52	42	3	18	10	1
C_2	4	40	52	1	12	10	3	11	8	0
C_3	4	26	12	55	6	0	3	9	13	12
C_4	3	25	7	12	6	0	2	0	38	19
C_5	50	12	6	9	4	1	3	1	7	5
C_6	4	37	11	13	5	0	2	0	38	19
C_7	4	38	12	11	6	0	5	1	17	9
N	41									
SC_1	3	21	12	55	5	0	3	7	10	11
SC_2	4	38	11	13	5	0	1	0	37	19
SC_3	51	14	6	8	4	0	3	1	9	6

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (*ES*), *EyeWide* (*EW*), *BrowDown* (*BD*), *BrowInnerUp* (*BIU*), *BrowOuterUp* (*BOU*), *MouthShrug* (*MS*), *MouthFrown* (*MF*), *CheekSquint* (*CS*) and *NoseSneer* (*NS*).

	S_t	S_c **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N	28									
C_1	2	48	0	0	27	4	0	0	0	21
C_2	4	64	0	0	34	0	1	0	0	0
C_3	4	0	21	0	0	56	0	0	0	23
C_4	3	0	0	89	0	1	0	0	9	0
C_5	50	6	17	7	15	8	15	15	18	0
C_6	4	0	0	93	0	0	0	6	1	0
C_7	4	11	0	26	1	4	19	13	10	16
N	41									
SC_1	3	0	26	0	0	66	0	0	0	9
SC_2	4	0	0	96	0	0	0	3	1	0
SC_3	51	6	14	13	12	7	14	15	18	1

((b)) The formed clusters and their distribution over participants.

Table 6.1: Clusters formed by the HDBScan algorithm using categorisation $B2$ (see Table C.1), the corresponding mean feature values and their distributions over participants. Samples assigned to clusters (S_c) are given as a percentage of total samples (S_t).

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	0	56	103	99	96	96	90
C_2		0	71	63	57	58	46
C_3			0	51	50	52	46
C_4				0	36	**14**	**27**
C_5					0	42	29
C_6						0	**23**
C_7							0

Table 6.2: Euclidean distances between the formed clusters as presented in 6.1(a). Small values, corresponding to similar clusters, are marked bold.

set of data points (S_t). All the remaining clusters, except for C_5, do not vary much in terms of size. One can see that certain clusters, for instance C_4, C_6 and C_7, are somewhat similar. Therefore, as stated, the euclidean distances between clusters were calculated, see Table 6.2.[3] In addition, three super-clusters were extracted, which were less discriminatory, using a more strict version of the HDBScan algorithm. Again, one of these clusters contains over half of the data points, whilst the remaining two are relatively small and similar in size. Further, now 41% of data points in S_t was labelled as noise.

First, consider C_5, which is the biggest cluster formed by the algorithm: it makes up 50% of S_t. This cluster corresponds to a neutral facial expression: all features are engaged to a fairly low extent. The corresponding data originated mostly from p10 (18% of the cluster) and p4 (17% of the cluster), but formed on the basis of data produced by all participants but p11. C_5 resembles the super-cluster SC_3, comprising 51% of data points. The distribution over participants parallels C_5's distribution: it is evenly distributed, albeit p11's data is contained in this cluster to a very low extent.

Furthermore, it is apparent that the remaining six fine-grained clusters correspond to facial expressions in which at least one feature is highly engaged. These six clusters will now be described in more detail. Visualisations of the facial expression corresponding to these clusters, using both video stills and a metahuman [Epic Games, 2023], can be found in Appendix F.

Cluster C_1 demonstrates a significant use of the features *EyeWide* (72), *BrowInnerUp* (52) and *BrowOuterUp* (42), next to the lesser engaged features *EyeSquint* (35) and *MouthFrown* (18). This cluster makes up for 2% of the full data set. The participants whose facial expressions this cluster mostly originated from are p3 (48%), as well as p6 (27%) and p11 (21%) to a lesser extent, see 6.1(b). This is compatible with results presented in Section 6.1.1, in which was found that the raising of the eyebrows was primarily demonstrated by p3 and p6. We remark here

[3] As has been mentioned previously, the anatomical correlation between the activity of features has not been taken into account. For instance, think of the *BrowInnerUp* and *BrowOuterUp* feature. Therefore, euclidean distances between clusters in which the eyebrows are and are not raised, might be enlarged compared to the case for which this anatomical relation is intertwined: the distance is now calculated considering both features, instead of only one.

that C_1 is the only cluster formed by HDBScan that contains a high engagement for the features corresponding to raising the eyebrows. Therefore, according to this analysis, raising ones eyebrows seems to always coincide with widening the eyes. Furthermore, since this is the only cluster containing a relatively high engagement for both the *BrowInnerUp* and *BrowOuterUp* features, it can be concluded that if one of these features is engaged, this also holds for the other. This is in congruence with the results described in Section 6.1.1.

Another cluster, corresponding to a facial expression in which a feature describing the eyebrows is active, is C_3 (4% of S_t). Most prominently, this cluster contains the feature *BrowDown* (55), as well as the feature *EyeSquint* (26). As one can see, this cluster sternly contrasts cluster C_1, which corresponds to a facial expression involving wide eyes and raised eyebrows. Therefore, as one can see, their euclidean distance is large; it is even the largest between all clusters that were formed (see Table 6.2). Cluster C_3 is analogous to the super-cluster SC_1, for which the *BrowDown* feature is engaged to the same extent, whilst *EyeSquint* is slightly less active (21). p7's facial expressions were primarily found in C_3 and SC_1 (56% and 66%, respectively), as well as those by p11 (23% and 9%, respectively) and p4 (21% and 26%, respectively). This again confirms the observations made during manual annotation (see Section 4.3.2), for which it was found that these three participants engaged this facial feature the most. Further, this is again in congruence with the results found in Section 6.1.1. Note again that C_3 is the only cluster formed by HDBScan that contains a high engagement by the feature corresponding to frowning the eyebrows. Therefore, according to this analysis, lowering ones eyebrows seems to always coincide with squinting the eyes.

C_4, C_6 and C_7 correspond to similar facial expressions: their euclidean distances all fall below 30, see Table 6.2. These facial expressions are comprised in the super-cluster SC_2. Hence, the overarching facial expression SC_2 describes, comprises the more fine-grained facial expressions, described by C_4, C_6 and C_7. These clusters contain the engaged features *CheekSquint*, *NoseSneer* and *EyeSquint*, all to a somewhat different extent.

C_4, representing 3% of S_t, contains the highly engaged feature *CheekSquint* (38). Further, the features *EyeSquint* (25) and *NoseSneer* (19) were lightly activated. Primarily, this cluster was made up of facial features demonstrated by p5 (89%), as well as p10 (9%), albeit to a much lesser degree. This is in congruence with the results presented in Section 6.1.1. Similarly, the most engaged features in cluster C_6 consists of *EyeSquint* (37), and *CheekSquint* (38), in addition to *NoseSneer* (19). This cluster makes up for 4% of S_t. Furthermore, similarly to C_4, the highest percentage of samples contained in this cluster originated from p5 (93%) (as was found in Section 6.1.1). Because of these findings, it is not surprising that the euclidean distance between clusters C_4 and C_6 is not large.

Moreover, cluster C_7 (also 4% of S_t) corresponds to a similar facial expression, in which the eyes are squinted, and the same slightly holds for the cheeks: feature *EyeSquint* (38) is most active, in addition to *CheekSquint* (17). Samples in this cluster originate mostly from p5 (26%) and p8 (19%), as well as p3 (11%), p9 (13%), p10 (10%) and p11 (16%) to a lesser extent. Hence, the distribution over participants in this cluster is quite evenly spread.

SC_2 exhibits similar results as C_4, C_6 and C_7, comprising 5% of S_t. Here, the *EyeSquint* and *CheekSquint* feature are engaged (38 and 37, respectively), whilst the *NoseSneer* feature is engaged to a lesser extent (19). The distribution over participants for this cluster shows that it mostly comprises data from p5 (96%). Since this super-cluster comprises the three more fine-grained clusters described above, the results are therefore in congruence with the results described in Section 6.1.1, as well as the observations made during manual annotation (see Section 4.3.2).

The last of the six fine-grained clusters is C_2, which contains the most prominent feature *EyeWide* (52) and *EyeSquint* (40), whilst others were not active much. This cluster makes up 4% of all data points. The cluster originates only from facial expressions used by p3 (64%) and p6 (34%), as was also found in Section 6.1.1.

Accompanying Conditions and Temporal Windows

Not only did the HDBScan algorithm form clusters based on the features present in our data set, it also reported on the distribution of the clusters over conditions and temporal windows. Table 6.3 describes this distribution for the seven fine-grained clusters and three super-clusters that were formed. This section reports on this distribution for the ten clusters reviewed above, and compares these results to those described in Section 6.1.1.

The fifth cluster, as well as the third super-cluster, corresponding to the neutral facial expression, obtained their data points from all conditions, which are distributed considerably evenly. Similarly, this is the case for the distribution over temporal windows, even though this expression is found more often at the end of target question utterances (22%), as opposed to the start (16%).

Cluster C_1, describing a facial expression with wide eyes and raised eyebrows, demonstrates a distribution which favours the start of target question utterances, as opposed to the end of these utterances (31% and 25% in comparison to 3%). Additionally, the data points contained in this cluster are most often found in the *PosNeg* (29%), *PosNeut* (17%) and *NegPos* (15%) conditions, the contrastive conditions agreeing with results found in Section 6.1.1.

Regarding the distribution over temporal windows, C_3 and SC_1, corresponding to a facial expression in which the eyebrows are lowered, demonstrate a similar pattern. One can clearly see that for these clusters, most of the data points are contained in the first three windows (27% or 28%, respectively), contrary to the remaining windows (7% or 6%), respectively. Hence, this feature is frequently engaged at the start of question utterances, therefore confirming the results presented in Section 6.1.1. Furthermore, the conditions whose data points this cluster mostly originates from are *NegNeut* (21% and 22%, respectively) and *PosNeut* (21% and 16%, respectively).

C_4 and C_6, corresponding to similar facial expressions, also show similar results concerning the distribution over conditions: both distributions are relatively evenly spread. For C_4, its data points are predominantly found in the *NeutNeut* (23%) and *PosNeut* (22%) conditions, which are the *NegNeut* (23%) and *PosNeut* (19%) con-

	S_t	S_c **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
N	28							
C_1	2	15	11	14	0	13	17	29
C_2	4	8	22	16	2	21	22	9
C_3	4	10	21	9	18	6	21	15
C_4	3	9	15	3	23	12	22	16
C_5	50	13	15	15	17	12	13	15
C_6	4	8	23	10	12	16	19	12
C_7	4	15	19	15	9	17	13	12
N	41							
SC_1	3	12	22	11	16	5	16	19
SC_2	4	8	24	11	9	16	20	12
SC_3	51	13	15	15	18	13	12	15

((a)) The formed clusters and their distributions over conditions.

	S_t	S_c **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
N	28					
C_1	2	31	25	22	19	3
C_2	4	20	25	17	17	22
C_3	4	27	27	24	15	7
C_4	3	20	19	24	23	15
C_5	50	16	18	21	22	22
C_6	4	29	28	19	13	11
C_7	4	16	19	19	21	26
N	41					
SC_1	3	28	26	26	14	6
SC_2	4	30	29	19	13	10
SC_3	51	16	18	21	23	22

((b)) The formed clusters and their distributions over temporal windows.

Table 6.3: Clusters formed by the HDBScan algorithm, using categorisation $B2$ (see Table C.1), and their distributions over conditions and temporal windows. Samples assigned to clusters (S_c) are given as a percentage of total samples (S_t).

ditions for C_6. These findings confirm those described in Section 6.1.1, stating that the *EyeSquint* and *CheekSquint* features were relatively highly engaged during the entirety of target question utterances. However, even though the distributions over temporal windows for these clusters do show similar patterns, they do not showcase these patterns to the same degree: for C_4, the corresponding facial expression is used more towards the beginning and middle of the target question utterances (24% compared to 15%), whilst for C_6, this is very prominently presented towards the start (29% compared to 11%).

In contrast to C_6, the distribution of C_7's data points over time favours the end of target question utterances (26% compared to 16%). Regarding the distribution over conditions: the corresponding facial expression is most frequently found in the *NegNeut* (19%) and *NeutPos* (17%) condition, as well as in the *NegPos* (15%) and *NeutNeg* (15%) condition, albeit less frequently.

The corresponding super-cluster, SC_2, showcases similar patterns. The data comprised in this super-cluster mostly originates from data captured in conditions *NegNeut* (24%) and *PosNeut* (20%), similarly to C_6 and C_7, even though the corresponding facial expression is also found in the remaining conditions. Regarding the temporal progression of this expression; the features contained in this expression are often found at the start of target question utterances (30% compared to 10%), which is analogous to C_4 and C_6.

Lastly, data from C_2 (in which the feature *EyeSquint* and *EyeWide* are most active) is found predominantly in conditions *NegNeut* (22%), *PosNeut* (22%) and *NeutPos* (21%). Moreover, the corresponding facial expression is found during the entirety of target question utterances, which confirms findings presented in Section 6.1.1.

6.2.2 Subsets of Data Set

The HDBScan algorithm was additionally implemented for the five subsets of the normalised and categorised data set, corresponding to the five most frequent combinations of used polarity markings, word order and sentence type: *no-pol/inv/SR*, *neg/inv/SR*, *no-pol/no-inv/SR*, *pos/no-inv/SR* and *neg/no-inv/SR*. Based on decreasing frequency, we will briefly discuss the results found for these data sets one by one.

For the *no-pol/inv/SR* data set, containing data from 84 of 304 trials, the formed clusters by the HDBScan algorithm and their distributions over participants, conditions and temporal windows, are depicted in Table 6.4 and Table 6.5. Again, the distribution of clusters over S'_t is quite even, except for C'_2. Comparing these to the clusters formed based on the full data set, the current clusters mirror the aforementioned super-clusters.

Namely, C'_1, which comprises 5% of the data set, contains the active features *BrowDown* (55) and *EyeSquint* (28), as well as *EyeWide* (18) to a lesser extent. This cluster seems to parallel C_3 and SC_1, taken from the complete data set. Comparing the distribution over participants, one can see that p4 and p7 are solely responsible for the creation of cluster C'_1 (27% and 73%, respectively), as was similarly the case for SC_1. Regarding the distribution over conditions, this

	S'_t	Most engaged features (mean value)								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N'	31									
C'_1	5	28	18	55	6	0	4	9	11	11
C'_2	59	9	4	7	3	1	3	1	4	4
C'_3	6	30	9	14	3	0	4	0	37	19

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (*ES*), *EyeWide* (*EW*), *BrowDown* (*BD*), *BrowInnerUp* (*BIU*), *BrowOuterUp* (*BOU*), *MouthShrug* (*MS*), *MouthFrown* (*MF*), *CheekSquint* (*CS*) and *NoseSneer* (*NS*).

	S'_t	S'_c **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N'	31									
C'_1	5	0	27	0	0	73	0	0	0	0
C'_2	59	2	20	0	18	2	23	23	11	0
C'_3	6	0	0	100	0	0	0	0	0	0

((b)) The formed clusters and their distributions over participants.

Table 6.4: Clusters formed by the HDBScan algorithm for the *no-pol/inv/SR* data set, the corresponding mean feature values and their distributions over participants. Samples assigned to clusters (S'_c) are given as a percentage of total samples (S'_t).

facial expression is found most often in the *NegNeut* (27%), *NeutNeut* (31%) and *PosNeut* (27%) condition, which mirrors the patterns found for C_3 and SC_1. The distribution over time windows further shows that this feature is used primarily at the start of target question utterances (31% compared to 2%).

Moreover, C'_3, comprising 6% of the current data set, contains the active features *EyeSquint* (30) and *CheekSquint* (37). This cluster is analogous to SC_2. The corresponding facial expression is found most frequently in the *NeutNeut* (49%), *NeutPos* (25%) and *NegNeut* (25%) condition, again mirroring the patterns found for SC_2. Note here that p5 is solely responsible for the creation of this cluster, which is in congruence with previous results (p5 is responsible for 96% of data for C_2).

Furthermore, the biggest cluster found for this data set, C'_2, which comprises 59% of S'_t, matches C_5 and SC_3, and corresponds to a neutral facial expression. Similarly to above, this expression is found in all conditions, and is used over all temporal windows, but slightly more towards the end of the target question utterances.

The first cluster C''_1, formed within the *neg/no-inv/SR* data set containing data from 55 trials, comprises 89% of S''_t and corresponds to a facial expression in which all features are engaged to a low extent. A table visualising these results can be found in Appendix E (Table E.1 and Table E.2). This cluster therefore mirrors clusters C_5 and SC_3, corresponding to the neutral facial expression. The distribution over participants is quite even, similarly to the distribution over temporal windows, even though the end of the question utterances were slightly favoured. The only

	S'_t	S'_c **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
N'	31							
C'_1	5	0	27	3	31	0	27	13
C'_2	59	10	26	2	28	13	21	1
C'_3	6	0	25	0	49	25	0	0

((a)) The formed clusters and their distributions over conditions.

	S'_t	S'_c **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
N'	31					
C'_1	5	31	28	25	14	2
C'_2	59	17	18	20	23	23
C'_3	6	21	18	20	21	20

((b)) The formed clusters and their distribution over temporal windows.

Table 6.5: Clusters formed by the HDBScan algorithm for the *no-pol/inv/SR* data set and their distributions over conditions and temporal windows. Samples assigned to clusters (S'_c) are given as a percentage of total samples (S'_t).

difference between C''_1 and its previous counterparts is that the corresponding facial expression was primarily found in the *NeutNeg* and *PosNeg* conditions (44%).

The second cluster formed, which makes up for 4% of the data set, parallels cluster C_1: the features *EyeWide* (83), *BrowInnerUp* (35) and *BrowOuterUp* (34) are highly engaged, as well as the feature *EyeSquint* (39). The distribution of C''_2 regarding conditions shows that the data only originates from conditions *NeutNeg* (40%) and *PosNeg* (60%), which somewhat agrees with the results found for C_1. Furthermore, the samples contained in this cluster originate from data produced by p3 (59%) and p6 (41%), which is analogous to C_1. The distribution over temporal windows reports that the facial expression corresponding to C''_2 is used primarily in the start and middle of the target question utterances (27% compared to 2%). Similar patterns are found for C_1 and SC_1.

Within the *pos/no-inv/SR* data set, containing the data of 30 recordings, only two super-clusters are found, see Appendix E (Table E.3 and Table E.4). The first of these comprises 12% of the data set, and shows high activity for the features *EyeWide* (33) and *EyeSquint* (22), in addition to *BrowInnerUp* (41) and *BrowOuterUp* (41). This cluster clearly parallels C_1, formed within the complete data set. Participants p3 and p11 are responsible for the forming of the cluster C'''_1, which is the similar to cluster C_1. Furthermore, the corresponding facial expression is often used at the end of the target question utterances (19% vs 27%), whilst this is more prominent towards the start for C_1. Lastly, the conditions in which the data of C'''_1 is prominently found, are *NegPos* (49%) and *NeutPos* (51%), which is again somewhat similar to C_1.

The remaining cluster, C_2''', which comprises 55% of S_t''', contains facial expressions in which all features are engaged to a similar extent, see Appendix E. Therefore, this cluster corresponds to the neutral facial expression.

The *neg/inv/SR* data set, containing the data of 25 of 304 recorded trials, produced three clusters (see Appendix E - Table E.5 and Table E.6), two of which do not present clear resemblances with clusters described previously. Within the second cluster (C_2''''), comprising 19% of S_t'''', only the *EyeSquint* feature is slightly engaged (13), whilst other features are not engaged at all. This cluster does not seem to parallel any cluster found within the complete data set. The distribution over participants for this cluster shows that it mostly comprises data from p8 (67%). Additionally, the corresponding facial expression is used most frequently at the end of target question utterances (9% compared to 24%). Further, data from C_2'''' is found predominantly in condition *PosNeg* (39%) and *NeutNeg* (47%).

The third formed cluster C_3'''' (8% of S_t'''') displays a similar result, except for the fact that now the *BrowDown* feature is mildly engaged (16). This cluster also does not parallel any cluster found within the complete data set. The distribution over participants shows that it mostly comprises of data from p4 (94%). Similarly, data from this cluster is found predominantly in condition *NeutNeg* (46%) and *PosNeg* (48%). Finally, the corresponding facial expression is used primarily at the start of target question utterances (24% compared to 10%).

The first cluster, C_1'''', parallels C_5, corresponding to the neutral facial expression.

Finally, for the *no-pol/no-inv/SR* data set, comprising the data of 20 of 304 trials, two clusters are formed, using a slighly more lenient variant of the algorithm. Again, the results are displayed in Appendix E (Table E.7 and Table E.8). One can see that the second cluster C_2''''' contains features with the highest engagement: *EyeSquint* (32) and *CheekSquint* (26), as well as *NoseSneer* (15). This cluster matches SC_2, formed based on the untrimmed data set. First, for both clusters, the data it is comprised of originates from p5 (31% for C_2''''' and 96% for SC_2), although one can clearly see that for C_2''''', this is found frequently for p7 and p9 as well. Furthermore, the distribution over conditions reports that the facial expression corresponding to these clusters primarily occurs in the *NeutPos* (61%) and *PosNeg* condition (25%). Lastly, the distribution over temporal windows favours the middle of question utterances (30% vs. 6%).

Cluster C_1''''' corresponds to a facial expression in which all features are engaged to a low extent, therefore again corresponding to the neutral facial expression, as C_5 and SC_3 further describe.

6.3 Overview of Results

This section summarises the most prominent results, previously discussed in great detail. We first discuss findings resulting from both methods of quantitative data analysis, after which we will describe the remaining results.

For the three features describing eyebrow movements, *BrowDown*, *BrowInnerUp* and *BrowOuterUp*, the engagement of these features declines during the target question utterances (see Figure 6.1 and 6.3(b): C_1, C_3 and SC_1). Furthermore, the *BrowInnerUp* and *BrowOuterUp* feature behave similarly: when one is engaged, so is the other, and they occur most frequently in the *NegPos*, *PosNeg* and *NeutNeg* conditions. Similar results are obtained, for in particular the *neg/no-inv/SR* and *pos/no-inv/SR* data set. Additionally, the *BrowDown* feature is most engaged in the *NegNeut*, *PosNeut*, *NeutNeut* and *NeutPos* conditions, which both the data analysis approach report.

Furthermore, the use of the *EyeWide* feature over time is quite consistent. Whilst for the first method of data analysis, a clear pattern could not be established, the clusters resulting from the implemented HDBScan algorithm show that participants had a preference to use this facial feature at the start of their target question utterances, although it was still also used at the end (see 6.3(b): C_1). The *pos/no-inv/SR* and *neg/no-inv/SR* data set demonstrate similar results for both data analysis approaches.

Besides, the presence of the *CheekSquint* and *EyeSquint* feature is consistent for all trials included in the six data sets, which is similarly found in all subsets of the complete data sets.

Lastly, it is found that some participants prefer to use certain facial features over others, as is the case for both the full data set and its subsets. The *BrowDown* feature is primarily used by p4 and p7, in contrast to the *BrowInnerUp* and *BrowOuterUp* feature, primarily used by p3 and p6. Participants p3 and p11 use the *EyeWide* feature most frequently. Lastly, p5 mainly engages the *CheekSquint* and *NoseSneer* feature, whilst p11 mainly engages the *MouthShrug* and *MouthFrown* feature (see Section 6.1.1 and 6.1(b)). The *EyeSquint* feature is active in data capturing all participants.

In contrast, the use of the *MouthShrug* feature is consistent over conditions, according to the results from the first data analysis method, and increases over time. The *no-pol/no-inv/SR*, *pos/no-inv/SR* and *no-pol/inv/SR* data set demonstrate a similar pattern. Contrasting, this pattern is not found in the results from the implemented clustering algorithm.

Additionally, in the *NegPos* condition, the *EyeWide* feature peaks at the end of the target question utterance. This pattern is found for the full data set, as well as the *pos/no-inv/SR* data set using the first data analysis method.

Moreover, visualising the mean blend shape measurements over time shows that the *BrowInnerUp* feature is highly engaged in all conditions. However, similar results are not found after clustering.

Lastly, the implemented clustering algorithm reports the most prominent facial expressions used in the data in a clear manner. The most prototypical facial expressions, marking polar question in Dutch, involve the following (see Table 6.1):

[1] Raised eyebrows and wide eyes

[2] Frowned eyebrows and squinted eyes

[3] Squinted cheeks, squinted eyes and a sneered nose

[4] Simultaneous squinted and wide eyes

[5] The neutral facial expression

Furthermore, all subsets of the data contain the neutral facial expression (C_5 and SC_3). The *no-pol/inv/SR* data set comprises facial expressions containing lowered eyebrows and squinted eyes (corresponding to C_3 and SC_1). This suggests that lowering the eyebrows always coincides with squinting the eyes. The facial expression corresponding to C_1, in which the eyes are both squinted and widened, next to having raised eyebrows, are found with a high frequency in the *pos/no-inv/SR* and *neg/inv/SR* data set. This suggests that raising the eyebrows always coincides with widening the eyes. Finally, the facial expression containing squinted cheeks as well as eyes, is found in the complete data set, as well as the *no-pol/inv/SR* and the *no-pol/no-inv/SR* data set.

Chapter 7

Discussion

This project aspired to analyse the visual cues marking different types of Dutch polar question utterances, and to compare these to the NMM that mark these utterances in NGT. This chapter proceeds with a comparison between the current results and the findings from previous research.[1] Furthermore, some limitations of the experiment design, pre-processing and data analysis methods are discussed. See Section 7.1 and Section 7.2, respectively.

7.1 Comparison of Results to Previous Research

Given that the topic of interest concerns the visual question markers of *different types* of polar questions, our first step was to determine what these types actually were. The question structures examined in this project, by means of manual annotation, concerned the polarity markings, such as a headshake or the use of a particle, in addition to the used word order and the sentence type. We hypothesised that the inversion of the subject and verb, in question utterances, was the most prominent question marker for polar questions. This did end up being the case: in over half of the target question utterances, participants' target question utterances involved inversion. However, the use of SVO word order is still prominently present in the data. Therefore, using inversion is not the only question marker in Dutch.

Furthermore, we expected that the use of tags was prevalent in the data, as is in line with [Englert, 2010, Esselink, 2022, Oomen and Roelofsen, 2023a, Oomen et al., 2023]. However, for only a small number of experimental trials, the target question utterance contained such a sentence-final particle: 32 out of 252 trials. This strongly contrasts with the BpQ/NGT-experiment, for which around a third of the question utterances involved a tag [Esselink, 2022, Oomen et al., 2023]. A partial explanation for this discrepancy is that in NGT, declarative and interrogative clauses consist of the same word order. Therefore, it is not possible to mark a question utterance using inversion, as is the case for spoken Dutch. Consequently, in comparison to Dutch, NGT may rely on this use of tags more, since it is not the

[1] Again, we remark that apart from the BpQ/NGT-experiment [Esselink, 2022], the previous described studies do not take into account the influence of bias on the presented visual cues. Thus, we keep in mind that these are not substantially comparable.

case that another sentence structural strategy is employed to mark the utterance as a question.

Additionally, particles indicating negative polarity were predicted to be present in the data [Englert, 2010]. Again, our hypothesis is confirmed: in 95 out of 252 trials, the target question utterance contained a particle indicating negative polarity. The number of positive polarity markers used by participants is significantly lower (18/252).

The two latter research questions concerned the ways in which different types of polar questions are visually marked in Dutch, and how these results compare to NGT [Esselink, 2022]. To answer these questions, we have obtained the most prototypical facial expressions from the data, in addition to their temporal progression and how these expressions correspond to contexts.

Using the HDBScan clustering algorithm, seven clusters, corresponding to the prototypical combinations of facial features and their characteristics, were extracted from the data, in addition to three super-clusters. As was discussed in Section 6.2, four particular facial expressions correspond to the ten clusters that were formed, apart from the neutral face (see 6.1(a)). These are the facial expression containing raised eyebrows and wide eyes (C_1), lowered eyebrows and squinted eyes (C_3 and SC_1), squinted eyes, squinted cheeks and a sneered nose (C_4, C_6, C_7 and SC_2) and simultaneous squinted and wide eyes (C_2). Again, visualisations of these expressions, by means of a metahuman [Epic Games, 2023] and video stills, are presented in Appendix F.

The first of these facial expressions, in which the eyebrows are raised and the eyes are wide, is captured within the definition of 'q', as defined by Coerts [1992] (see Section 1.1.1). This 'q', a prominent question marker in sign languages, and specifically NGT, comprises raised eyebrows and a head and body tilt forward, in addition to the optional widening of the eyes. The scope of this project prohibited the analysis of head and body tilts, since these are not captured by the Live Link Face software [Epic Games, 2023]. However, in all other aspects, C_1 resembles 'q'. Similarly to Coerts [1992], Esselink [2022] found that the simultaneous occurence of raised eyebrows and wide eyes is a prominent question marker in NGT. This facial expression was found in all conditions but *PosNeg*, in which positive prior belief is later contradicted with negative contextual evidence. De Vos et al. [2009] further confirm these findings, reporting that raised eyebrows arise in NGT when an element of surprise is present. Further, the contrastive conditions (*PosNeg* and *NegPos*), displayed opposite patterns to the *NeutPos* conditions. The findings by Esselink [2022] and De Vos et al. [2009] strongly contrast with the results found in the current study. As one can see in 6.3(a), in spoken Dutch, raised eyebrows and wide eyes most frequently occur in the contrastive conditions, in addition to the *NeutPos* condition. Hence, in Dutch, these combination of cues pattern alike, whilst this was not the case for NGT [Esselink, 2022].

The finding that participants raised their eyebrows and widened their eyes frequently, differs from results found by Nota et al. [2021] and Nota et al. [2023]. Their research reports that the activity of the eyebrows during question utterances is only significant regarding the frowning of the eyebrows, whilst raising ones eyebrows does not produce similar effects. However, even though this facial expression occurs significantly often in the current data set, note that it only comprises 2% of all data that was collected.

Further, regarding the temporal progression of this expression, the corresponding features are most often displayed at the start of target question utterances, and their use decreases over time. This result is substantiated by both the distribution of the corresponding cluster (C_1) over temporal windows (see 6.3(b)), as well as the gradient of the features *BrowInnerUp* and *BrowOuterUp* (see Section 6.1). Again, dissimilar results were found for NGT: the use of the expression 'q' in NGT spans the entirety of question utterances [Coerts, 1992, Esselink, 2022].

Lastly, clustering subsets of the data regarding the used word order, sentence type and polarity markings, provided us with some more insights: in question utterances involving polarity markers, SVO word order and only a sentence radical,[2] participants raised their eyebrows significantly often. For other subsets of data, this is not the case.

Hence, the current behavior of features active in C_1 resembles the analyses of these features in previous literature to an extent. The fact that this expression occurs quite often overall is fairly surprising, given [Nota et al., 2021, 2023], and disconfirms our hypothesis. However, even though this result therefore seems to resemble those found in NGT, it is clear that the use of this expression in NGT and spoken Dutch is not comparable [Coerts, 1992, Esselink, 2022, De Vos et al., 2009].

In contrast, as had been hypothesised, frowning ones eyebrows, in combination with squinting ones eyes, is a prominent polar question marker in Dutch. The HDBScan algorithm formed both a more fine-grained cluster, as well as a super-cluster, resembling this facial expression (see 6.1(a) - C_3 and SC_1). Therefore, this facial expression can be regarded as an important question marker in Dutch. This finding is in accordance with [Nota et al., 2021, 2023, Zygis et al., 2017], who found that in spoken Dutch and German, frowned eyebrows most frequently occur in question data, in comparison to response data. Further, frowning ones eyebrows leads to a better understanding of question utterances by the addressee [Nota et al., 2023]. Additionally, Miranda et al. [2021] and da Silva Miranda et al. [2020] reported that lowered eyebrows, in combination with squinted eyes, sneered noses and a head tilt to the right, are prominent question markers in Brazilian Portuguese and Mexican Spanish. In this way, Dutch seems to resemble the findings by these studies on spoken languages.

Not only does Dutch resemble many spoken languages, and these results therefore confirm those found previously, this is the case for NGT as well. As Esselink [2022] described, frowned eyebrows are an import biased polar question marker in NGT.[3] Further, Esselink [2022] found that participants' eyebrows were frowned in

[2] Hence, the corresponding subsets of the data are the *neg/no-inv/SR* data set and the *pos/no-inv/SR* data set. Target question utterances that are comprised in these data sets are therefore, for instance: *'Kim is geen vegetariër?'* (*'Kim is not a vegetarian?'*) or *'Dus, Kim is wel een vegetariër?'* (*'So, Kim is a vegetarian?'*)

[3] Note here that in [Esselink, 2022], as well as in the current study, the context was manipulated as to examine the influence of bias on the ways in which questions were asked. The use of this marker in more neutral contexts has not been investigated in both of these studies.

almost twice as many cases as to when they were raised. The same results were obtained currently: C_3 contains twice as many data points as C_1, in addition to a super-cluster SC_1 being formed for this facial expression. However, the temporal progression of this combination of features does not mirror Esselink [2022]. C_3 and SC_1 extract most data points from the start of target question utterances, as was confirmed by the findings in Section 6.1. The opposite effect was found for NGT [Esselink, 2022]. Furthermore, in NGT, the corresponding facial expression most frequently occurred in the *PosNeg* condition, as we similarly hypothesised for Dutch. However, these are the *PosNeut* and *NegNeut* condition for the current data set, thereby refuting our hypothesis.

Finally, in the subset of the data corresponding to target question utterances involving no polarity markings (both visually and spoken), VSO word order and no tag (the *no-pol/inv/SR* data set),[4] the facial expression comprising frowned eyebrows and squinted eyes is frequently present. This is not the case for other reviewed subsets of the data. This is further in accordance with [Nota et al., 2021], who analysed the same types of question utterances as this subset of data describes.

Thus, the features comprised in C_3 and SC_1 resemble the findings reported in previous literature, both on spoken languages, as well as NGT. However, again, the use of frowned eyebrows and squinted eyes in NGT and spoken Dutch is not comparable, with regards to conditions and temporal progression. [Coerts, 1992, Esselink, 2022, De Vos et al., 2009].

The last two prominent facial expressions, containing squinted cheeks and eyes and nose sneers, and simultaneous wide and squinted eyes, are not as frequently discussed in the literature. Therefore, the corresponding clusters of facial features that were obtained from the data are briefly discussed here.

A facial expression corresponding to C_4, C_6 and C_7, as well as the super-cluster SC_2 in which they are comprised, includes squinted eyes and cheeks, as well as a nose sneer to a lesser extent. This expression most frequently occurs in the *NegNeut* and *PosNeut* condition. Furthermore, participants preferred to use this expression primarily at the start of their question utterances. Similar results were obtained by visualising the behavior of these features over time, see Section 6.1. Lastly, p5 is primarily responsible for the creation of these three fine-grained cluster and their corresponding super-cluster.

In NGT, similar patterns were not observed: the *CheekSquint* and *NoseSneer* feature were barely active, especially compared to the remaining features. Furthermore, previous literature regarding question marking in spoken languages has not reported on a combination of squinted eyes and cheeks, and a sneered nose, as being a prominent question marker in Dutch.

However, these active facial features were frequently exhibited in the collected data for this project. In particular, in the subsets *no-pol/inv/SR* and *no-pol/no-inv/SR*[5] of the data set, this facial expression was often presented.

[4] Hence, corresponding target question utterances can be, for instance: *'Is Kim een vegetariër?'* (*'Is Kim a vegetarian?'*)

[5] Hence, question utterances for which the data in this data set was captured, are for instance: *'Is Kim een vegetariër?'* (*'Is Kim a vegetarian?'*) or *'Kim is een vegetariër?'* (*'Kim is a*

Lastly, the facial expression corresponding to C_2 contains simultaneous squinted and widened eyes. This is quite a surprising result, since these features seem to be antagonistic. Consequently, similar results have not been reported in previous literature. Section 7.2 discusses these remarkable results in more detail.

The expression most frequently occurs in the *NeutPos* and *PosNeut* and *NegNeut* condition. Furthermore, the corresponding distributions over temporal windows is evenly spread, which is again confirmed by Section 6.1, in which was shown that the activity of the features *EyeSquint* and *EyeWide* is relatively constant over all trials. Finally, only the data produced by p3 and p6 is responsible for the creation of this cluster. No subsets of the data that were examined more closely in this project, contain a high frequency of these facial expressions.

As stated, the final facial expression found in the data, making up for half of the data points, is the neutral facial expression (see 6.1(a) - C_5 and SC_3). This result is not particularly surprising given that, by manual annotations, it was found that participants rarely exhibited visual cues (see Section 4.3.2), in addition to the distribution of blend shape measurements leaning very much to the lower side of values (see Figure C.1). Furthermore, Nota et al. [2021] found that, when comparing question utterances to response utterances, the response utterances contained an activation of certain facial features in a significantly higher manner than the corresponding question utterances. Hence, the fact that the current data comprises so many neutral facial expressions might not be so surprising after all.

7.2 Limitations

Of course, as holds for every research project, this study had its limitations. The first, and arguably the most critical, is the clusters' distribution over participants.

As was described in Section 5.3 and Appendix B, the data was categorised before it was fed to the HDBScan clustering algorithm. However, after the first implementation (see Table C.1 - $B1$), it became clear that the distribution over participants was not ideal: most clusters were formed on the basis of data produced by $2 - 4$ participants, amongst other shortcomings (as one can see in Table C.3 - which further provides an in depth account of the implemented categorisations and corresponding obtained results). This therefore prohibited us from generalising the data, and consequently drawing conclusions based hereupon. Two additional categorisations of the data were implemented, the results of which can be found in Appendix B. The third implementation produced similar results, to a more extreme extent: all clusters contained data produced by exactly one participant, whilst the second was used to obtain the final results of this project.

However, as stated in Appendix C, this still does not lead to optimal results. As one can see in 6.1(b), we were not able to avoid the above problem for this data set, apart from the cluster corresponding to the neutral facial expression: most clusters were still formed on the basis of data produced by few participants. Especially in

vegetarian?'), respectively.

the case of the super-clusters, this imbalance is found to an even higher extent. Unfortunately, due to the scope of the current project, additional categorisations and their corresponding clustering results have not been implemented. This means that even though some clear patterns were established within the results as presented in Section 6.2, these results need to be interpreted with caution: it could very well be that these are participant-specific. Therefore, this thesis project could only provide a *preliminary* analysis of the 3D data.

Finally, one additional limitation regarding the blend shape values captured in the data set is that the data was not balanced regarding the number of samples for each participant, condition and temporal window, as was similarly the case for [Esselink, 2022]. For instance, when a participant produces relatively long utterances in comparison to another participant, this first participant could end up comprising twice as many data points in the final data set. As a consequence, the clustering results can be skewed towards the first participant's particular facial expressions. This possible influence has not been accounted for within the current data analysis.

Furthermore, as stated, one remarkable finding is that the features *EyeSquint* and *EyeWide* frequently occur together during target question utterances. This was not in line with what we expected, given that these features are antagonistic. To understand the origin of these findings, the joint occurrence of these features in the raw data set was counted. The number of frames, in which for both features a blend shape value was measured above $0.x$, for which $x \in \{1, 2, 3, 4, 5, 6\}$, equals $11.367, 3.967, 1.530, 295, 82$ and 15, respectively. Thus, in over a third of the data set, both the *EyeSquint* and *EyeWide* feature are simultaneously activated, albeit to a mild extent. Especially, for p3 and p11, as well as p6, p7, p8 and p9 to a lesser extent, the *EyeSquint* feature was frequently active in all conditions, even those conditions for which it was not necessarily expected. Section 6.1 further confirms this, see visualisation Figure 6.3: this feature is often highly active in all conditions, compared to the remaining features.

Hypothesising that these results were found because of the lighting in the studio, the temporal progression of the *EyeSquint* feature in some trials was visualised for the declarative condition, see Figure 7.1. As one can see, in the declarative condition, the *EyeSquint* is very highly engaged by many participants, thereby insinuating that the squinting of the eyes is not necessarily a question marker, but something that happened over the course of the entire experimental sessions. Since participants were illuminated from above, and this light was fairly bright, we expect that this influenced the frequency of eye squinting in a significant manner.

Lastly, two participants regularly wore glasses outside of the experiment: p6 and p11. p11 wore lenses during the experimental sessions, whilst p6 did not. We expect that this further greatly influenced the ways in which these participants squinted their eyes, either because they could not see the confederate well, or because their lenses irritated their eyes. Even though the absence of glasses eliminated the influence of glasses and their shimmer on the measured blend shape values, we therefore compromised on the responding behavior of facial features by participants who regularly wear glasses.

One additional and somewhat surprising result is that participants often made use

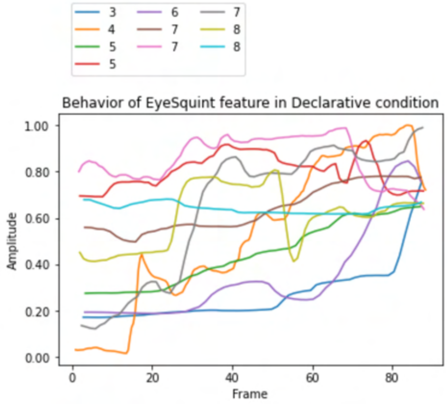


Figure 7.1: Engagement of *EyeSquint* feature during the *Declarative* condition by certain participants. The ten trials in which the feature is engaged most maximally are visualised.

of the neutral facial expression. As one can see in Table 6.1, C_5 and SC_3 both comprise around half of the samples contained in the data set. In comparison to the other formed clusters, this number is extensive. There are a few possible reasons for why this outcome has been observed. The first, and most explicable, is that the neutral facial expression was presented a large amount within the data. Based on the observations made by manual annotation, this could very well be the case. However, during annotation (see Section 4.3.2), it was also found that for p9, the participant exhibiting the least amount of visual cues, these were still presented in over half of the trials. Therefore, it seems like the neutral facial expression was frequently presented during the experimental sessions, but maybe not to the extent as Table 6.1 suggests.

Furthermore, as is presented in 6.3(b), the distribution over temporal windows of the clusters resembling the neutral facial expression, C_5 and SC_3, shows that this expression occurred most frequently at the end of target question utterances, even though it was still found at the start and in the middle as well. This could suggest that the data still contains noise: that participants finished their question utterances, and afterwards relaxed their face, all within the span of the selected video data and corresponding 3D data. This is further corroborated by the fact that the noise was manually removed from the data set; the starting and ending frames of the target question utterances were observed and noted down, which is not the most reliable method.

Another explanation for these findings is the use of the density based algorithm HDBScan. Because the neutral facial expression comprises a lot of the captured data, this results in a large cluster that is relatively dense. Surrounding clusters with a smaller size may now be contained within this large cluster. This prohibits us from observing fine-grained differences between the large and ‘neutral’ cluster, compared to the hypothetical smaller surrounding clusters, even though these differences may be quite informative. Building on this argumentation, one can see that when forming the super-clusters based on the complete data set (see

Table 6.1), informative clusters, for instance regarding the facial expression with wide eyes and raised eyebrows, seem to disappear (and are most probably labelled as noise). This again could result from the use of a density based algorithm, as one can see that the corresponding fine-grained cluster only comprises 2% of the data set.

Lastly, it could have been the case that participants were not completely comfortable during the experiment, for instance, because of the multiple cameras and lights that were facing them. Therefore, their utterances might not be accompanied with the most natural visual features, resulting in the frequent use of the neutral facial expression.

Additionally, one limitation of this project regards the dimension reduction, as described in Section 3.3.1 and Section 5.1.1. In order to obtain only the relevant facial features, the highly correlated features were combined (see Section 3.3.1). In this stage of pre-processing, the combination of features was based on the selection made during the 3D data analysis of the BpQ/NGT-experiment [Esselink, 2022]. Furthermore, in Section 5.1.1, out of the remaining 39 features, nine were selected for data analysis. For each of the 39 features, their behavior over time was compared to their behavior in the baseline recordings captured during the BpQ/NGT-experiment, in which participants mouthed the target questions with a neutral facial expression. In this way, the features sensitive to noise, for instance regarding the mouth and jaw, were excluded.

Whilst selecting the same features here, as in the BpQ/NGT-experiment, provides us with proper means to compare results between the two studies, this may not have lead to optimal results currently. For instance, looking at 6.1(a), the *MouthShrug* feature is barely activated in all of the clusters, therefore suggesting that this feature is not prominently involved in question marking in Dutch. For NGT, on the other hand, this was the case: one cluster contained frowned eyebrows, squinted eyes and a shrugged mouth. Hence, even though similarities have been found regarding the relevance of features in both languages, these similarities are not one-on-one.

Moreover, the combination of highly correlated features has not taken into account the anatomical capabilities of the human face. As was observed in both the current study as well as the BpQ-NGT/experiment, the *BrowInnerUp* and *BrowOuterUp* feature are always simultaneously engaged. Similar patterns are found for the *BrowDown* and *EyeSquint* feature, as well as the features corresponding to raised eyebrows and the *EyeWide* feature. This suggests that this does not only appertain the fact that participants choose to engage these features at the same time, but rather that, because of anatomical constraints, when one feature is active, the other is as well.

Lastly, one limitation is that video data was manually annotated and analysed. As described before (see Section 4.1), this process is categorical and prone to subjective judgements and inter-annotator disagreements. Especially, for the annotations regarding prosodic patterns, these were annotated purely based on observation, which is a difficult task. Furthermore, whether a headshake is inquisitive, or indicates negative polarity [Oomen et al., 2023], can be difficult to determine.

Chapter 8

Conclusion

This project has performed a preliminary analysis of the visual cues that mark different types of Dutch polar question utterances, and has compared these to the NMM that mark these utterances in NGT. We now conclude. Note that, because the distribution of clusters over participants is not representative, the current results and the subsequent conclusions should be interpreted with caution.

Firstly, we aimed to ascertain what the most prevailing types of question structures in Dutch are. The question structures examined in this project, by manual annotation, concerned the spoken and visual polarity markings, the used word order and sentence type. Both VSO and SVO word order were used during the target question utterances, SVO to a slightly lesser extent. Furthermore, for only a small number of experimental trials, the target question utterance contained a sentence-final particle (tag). Lastly, in many cases, the target question utterance contained a particle indicating negative polarity, in contrast to the number of positive polarity markers, which was significantly lower. The most frequent combinations of polarity markers, word order and sentence type are *no-pol/inv/SR*, *neg/inv/SR*, *no-pol/no-inv/SR*, *pos/no-inv/SR* and *neg/no-inv/SR*. Regarding their distributions over conditions, it was found that for the utterances involving polarity, the polarity of the contextual evidence, as provided by the second confederate, was copied.

Secondly, the most prototypical combinations of visual cues in the data were obtained, in addition to their temporal progression and how these expressions correspond to contexts. Five particular facial expressions correspond to the seven clusters and three super-clusters that were formed. These comprise:

[1] Raised eyebrows and wide eyes

[2] Frowned eyebrows and squinted eyes

[3] Squinted cheeks, squinted eyes and a sneered nose

[4] Simultaneous squinted and wide eyes

[5] The neutral facial expression

The first of these facial expressions, in which the eyebrows are raised and the eyes are wide, is found primarily in the *PosNeg* condition, in which positive prior belief is later contradicted with negative contextual evidence. Further, the corresponding features are most often displayed at the start of question utterances, and their use decreases over time. Additionally, in question utterances involving polarity markers, no inversion and no tag, the facial expression containing raised eyebrows and wide eyes is frequently present. For other sentence structures, this is not the case. Finally, raised eyebrows always co-occur with wide eyes.

Frowning ones eyebrows, in combination with squinting ones eyes, is an even more prominent polar question marker in Dutch: participants' eyebrows were frowned in almost twice as many cases as compared to when they were raised. Again, frowned eyebrows always co-occur with squinted eyes. Further, similarly to the contrasting expression containing raised eyebrows, simultaneous frowned eyebrows and squinted eyes were most often displayed at the start of target question utterances, after which their use decreased. Additionally, the data captured in the *PosNeut* and *NegNeut* condition contains this facial expression most frequently. Finally, for the subset of the data corresponding to target question utterances containing no polarity markings (both visually and spoken), VSO word order, and no tag, frowned eyebrows and squinted eyes are often present. This is not the case for other reviewed subsets of the data.

The facial expression containing squinted eyes and cheeks, as well as a nose sneer to a lesser extent, also most frequently occurs in the *NegNeut* and *PosNeut* condition. Furthermore, allthough the corresponding distributions over temporal windows is relatively even, participants still preferred to use this expression at the start of their question utterances. In question utterances containing no polarity and no tag, this facial expression is further often found.

Moreover, the facial expression which contains simultaneous squinted and widened eyes, most frequently occurred in the *NeutPos* and *PosNeut* and *NegNeut* condition. Furthermore, the corresponding distributions over time windows is evenly spread. No subsets of the data that were examined more closely in this project, involve a high frequency of this facial expression.

As stated, the final facial expression found in the data, making up for half of the data points, is the neutral facial expression. The distribution of this expression over participants, conditions and temporal windows is evenly spread.

NGT and spoken Dutch showcase some similar patterns, however, this is not always the case. For instance, in both NGT and in Dutch, the expression containing frowned eyebrows and squinted eyes is an important visual polar question marker. In NGT, this expression occurred mostly in the *PosNeg* condition, while in Dutch, this is the case for the *PosNeut* and *NeutNeut* conditions. Regarding temporal progressions, dissimilarities are found as well: in NGT, frowned eyebrows occur at the end of question utterances, whilst in Dutch, these occur at the start.

Another expression, which arose in a smaller number of cases for both languages, comprises raised eyebrows and wide eyes. In Dutch, one finds this expression mostly in the contrastive conditions, *PosNeg* and *NegPos*, in addition to the *NeutPos* condition. Hence, the patterns of features within these conditions is alike. In NGT, opposite patterns are found for these conditions: the expression is not found in the contrastive conditions at all. Additionally, the gradient of these features differs between languages. In Dutch, a decline in activity is observed, whilst in NGT, the engagement of the features is constant.

Additionally, in NGT, the feature *MouthShrug* is prevalent in the prototypical facial expressions marking polar questions, whilst this is not the case in Dutch. Contrasting, *CheekSquint* and *NoseSneer* exhibit a higher level of engagement here. Furthermore, in Dutch, the *EyeSquint* and *EyeWide* feature are often simultaneously active.

Lastly, the question structures in NGT frequently contain a tag, in around a third of cases. In Dutch, however, similar results are not obtained.

8.1 Future Work

As described by Section 7.2, there are various limitations to the current project. We proceed by suggesting further avenues of research, which in their turn overcome the pertained limitations, as well as additional paths of research one could explore.

We first start by suggesting some tweaks of the studio setup that was used during the current experiment. As was discussed in Section 7.2, the features *EyeSquint* and *EyeWide* are often simultaneously active, as well as the latter being very active overall, which is not in line with our expectations. We hypothesise that the setup of the lighting in the studio contributed to these observed effects. Namely, for the BpQ/NGT-experiment [Esselink, 2022], similar results were not obtained. Therefore, we suggest that in future variations of this experiment, the participants are illuminated from both their left and right side, but with a lesser brightness, as was the case in the BpQ/NGT-experiment [Esselink, 2022].

Originally, for the current project, the aim was to capture the 3D data using two iPhones: one recording the participants' right side of the face, and one recording the left, after which the mean of their captured blend shape measurements is taken and incorporated into the pre-processed data set. The reason for this was that Esselink et al. [2023] found that angle has a small, yet significant influence on the blend shape measurements that Live Link Face captures. The aim was to avoid this effect, however, due to storage issues, this was unfortunately not feasible for the current experiment. When future iterations of this project are executed, we therefore suggest that participants' blend shape measurements are captured from multiple angles.

Furthermore, to avoid a laborious pre-processing task, we recommend the use of synchronising equipment that does work (as we attempted with Tentacle Sync [Tentacle Sync, 2023a]). This process will be even more arduous when additional footage needs to be synchronised.

Lastly, of course, we have only captured data from nine participants. In order to generalise the obtained results to the Dutch language overall, data from more participants need to be analysed. Ideally, in this case, participants are native speakers from a platitude of different regions in the Netherlands, to get the full picture of question markers in the Dutch language.

Regarding the manual annotations that were carried out, it is always sensible to validate these annotations by another annotator, which is what we would therefore suggest. As Oomen et al. [2023] remarked, the process of annotating video footage is prone to subjective judgements and inter-annotator disagreements, in addition to this process being categorical.

Moreover, as stated in Section 7.2, the manual annotations regarding prosodic patterns were performed on the basis of observation. This is not very reliable, considering that intonation often fluctuates, and at a high speed. For the current project, the intonation patterns used by participants were not analysed any further, apart from these robust annotations. However, in future research concerning these patterns, we suggest the use of additional software, such as PRAAT [Boersma and Weenink, 2023], for a more objective and reliable analysis (see Section 4.1 for a more thorough explanation of this software).

Finally, the Live Link Face software [Epic Games, 2023] was developed fairly recently, and not much research has been done on exploring the validity of this software. As Esselink [2022] remarked, initial testing showcases that the measured blend shapes are accurate. However, especially with recent software as this is, validating these measurements against manual annotations would be insightful.

As previously stated in Section 7.2, one limitation of the current project is that the dimension reduction has taken place based on data from the BpQ/NGT-experiment [Esselink, 2022]. However, since the prevailing results suggest that there are some significant differences between the ways in which different types of polar questions are marked in NGT and in Dutch, this therefore also suggests that this selection of features for the Dutch data set was not optimally accurate (for instance, think of the *MouthShrug* feature that is barely engaged in Dutch, but does show significant activation in NGT). In further work, we advice that the correlation of features is calculated based on the current data set, after which those with a high correlation are combined. Furthermore, using a baseline recording of spoken Dutch, and obtaining the features not sensitive to noise based on this Dutch recording, would prove to be insightful. Lastly, it needs to be taken into account that anatomically, some features are prone to be active together (such as the *BrowInnerUp* and *BrowOuterUp* feature). Furthermore, from the current results, it follows that the *BrowDown* feature is always accompanied by the *EyeSquint* feature, and the features regarding raised eyebrows are always accompanied by widened eyes. We here do not present a specific way in which this information can be incorporated into the analysis, however, it is something to consider.

Moreover, as has been described in detail, the current project only carries out a preliminary analysis of the data (see Section 5.3 and Appendix C). Of course, since the data is publicly available, we suggest future studies analyse this data in a myriad of ways. We now go over some key suggestions.

First, we recommend to remove the neutral facial expressions from the data set. As Section 7.2 described, the results presented in Table 6.1 suggest that neutral facial expressions make up half of the data comprised in the data set. This is a very large number, especially in comparison to the remaining clusters that were formed. Since we have used a density based clustering algorithm, HDBScan, small but insightful surrounding clusters may be included in this large cluster. Therefore, we suspect that removing the neutral facial expressions from the data set could lead to more insights. One way to accomplish this is to calculate the mean activation of all features for each participant. Samples in which all features are engaged to an extent lower than this average, or the standard deviation added to this mean, are now discarded. The remaining data set is likely relatively small, but may showcase the patterns present in the data to a better extent.

Secondly, further work could explore the use of different types of clustering analyses on the data. Of course, this project, as well as the BpQ/NGT-experiment, has implemented a density based clustering algorithm on the captured data. However, as was just remarked, the fact that this algorithm is density based does not always work in our favour, especially when considering the super-clusters (see Table 6.1).

Thirdly, as Appendix C has reviewed in much detail, multiple categorisations of the data were implemented in order to obtain the most optimal results. However, due to the scope of this project, only a preliminary analysis was performed, in which the distribution over participants was still not representative. In order to avoid this issue, we suggest a couple of changes to the data normalisation and categorisation.

Firstly, instead of normalising the data based on the target question utterances and the calibration recordings (see Section 3.3), the data can also be normalised based on only these question utterances. The upside to this method is that results are now more comparable to those of the BpQ/NGT-experiment [Esselink, 2022], since this method was also employed in that analysis. However, the downside is that participants most likely did not utilise their entire range for each feature. Therefore, for features with a low engagement overall, the maximum of the captured value is also relatively low, therefore all feature values are subsequently enlarged during normalisation. Aside from this downside, it would still be informative to perform data analysis on this newly normalised data set, since it could lead to new insights.

Furthermore, the third attempted categorisation of the current project could be extended. During this categorisation, for each participant and feature, the upper bounds of the quartiles containing 25% of data were obtained. The replacement value of the quartiles now was the mean value of data points contained in this quartile. However, as can be seen in Appendix C, the HDBScan algorithm now formed nine clusters, each based on all data produced by one participant. This most probably occurred because the values were participant-specific, and the categorised data was very dense: differences between samples produced by participants were very minimal. Therefore, future work could try to categorise the values in the just described manner, but then only for each feature and not for each participant.

Besides, subsequent categorisations and analyses of the data could try to implement a smaller number of bins. The current categorisation employed eight bins, however, this could be reason for participant-specific formed clusters (given that the feature vectors of these frames are then too distinctive, which we try to avoid by using these bins in the first place).

Another consideration for future studies that will use this data, is to remove data from participants p5, p6 and p11. As stated, p6 regularly uses glasses, but was not able to wear these during the experimental sessions. As a consequence, the data produced by this participant may have not been naturalistic and representative. Similarly, this was the case for p11, however, they did wear lenses during the experiment. Furthermore, the fifth participant, as can be seen in Table 6.1, is almost solely responsible for the creation of clusters C_4 and C_6, in addition to super-cluster SC_2. Additionally, as Section 6.1 also described, this participant had a high predisposition of using the *CheekSquint* feature, which was used less by the remaining participants. Therefore, the data produced by this participant could have skewed the results: the noticeable presence of the facial expression containing these squinted cheeks may not have been so prevalent overall.

Besides, during the analysis of the 3D data, in combination with the obtained sentence structure, head tilts and body movements were not incorporated. This is because the Live Link Face application [Epic Games, 2023] does not capture these movements. Given that these features were annotated in ELAN, it would be insightful to integrate these results into the analysis as well.

Lastly, the question structures obtained from the data set, as described in Section 4.3.1, concern the sentence type, polarity markers and word order as used by participants. However, regarding the sentence type, a distinction between different types of tags that were used was not made in this project. In future work, one could investigate the discrepancies between questions containing these different types of tags.

Thus, as one can see, there are various avenues that future studies can explore. The hope is that with the current study, we have taken a small, yet important step, towards the understanding of how Dutch native speakers visually mark different types of polar questions.

Bibliography

N. Nota, J. P. Trujillo, and J. Holler. Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8), 2021. 1017.

N. Nota, J. Trujillo, and J. Holler. Conversational eyebrow frowns facilitate question identification: An online study using virtual avatars. *Cognitive Science*, 47 (12), 2023.

Marzena Zygis, Susanne Fuchs, and Katarzyna Stoltmann. Orofacial expressions in German questions and statements in voiced and whispered speech. *Journal of Multimodal Communication*, 4:1–6, 08 2017.

L. da Silva Miranda, C. da Silva, J. Moraes, and A. Rilliard. Visual and auditory cues of assertions and questions in Brazilian Portuguese and Mexican Spanish: a comparative study. *Journal of Speech Sciences*, pages 73–92, 2020.

L. Miranda, M. Swerts, J. Moraes, and A. Rilliard. The role of the auditory and visual modalities in the perceptual identification of Brazilian Portuguese statements and echo questions. *Language and Speech*, 64(1):3–23, 2021.

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

R. Pfau and J. Quer. *Nonmanuals: their grammatical and prosodic roles (pp. 381-402).* 2010.

L. Esselink. Computer Vision and Machine Learning for the Analysis of Non-Manual Markers in Biased Polar Questions in Sign Language of the Netherlands [Master's Thesis, University of Amsterdam], 2022.

C. Englert. Questions and responses in Dutch conversations. *Journal of Pragmatics*, 42(10):2666–2684, 2010.

J. Borràs-Comes, C. Kaland, P. Prieto, and M. Swerts. Audiovisual correlates of interrogativity: A comparative analysis of Catalan and Dutch. *Journal of Nonverbal Behavior*, pages 53–66, 2014.

A. Gaasbeek. Polar Questions in Sign Language of the Netherlands (NGT) and Dutch [Master's Thesis, University of Amsterdam], 2023.

M. Oomen and F. Roelofsen. Biased polar questions in Sign Language of the Netherlands - Methods description, 2023a. URL: `https://uvaauas.figshare.com/articles/preprint/Biased_polar_questions_in_Sign_Language_of_the_Netherlands_-_Methods_description/21701954`.

Epic Games. Recording facial animation from an ios device, 2023. URL: `http://www.forestry.ubc.ca/conservation/power/` [Accessed: 08/09/2023].

ELAN. ELAN (Version 6.7) [Computer software], 2023. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from `https://archive.mpi.nl/tla/elan` [Accessed: 09/12/2023].

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

U. Zeshan. Interrogative Constructions in Signed Languages: Crosslinguistic perspectives. *Language*, pages 7–39, 2004.

C. Cecchetto. *Sign Language (pp. 292–315).* De Gruyter Mouton, 2012.

M. Oomen and F. Roelofsen. Biased polar question forms in Sign Language of the Netherlands (NGT). FEAST. Formal and Experimental Advances in Sign language Theory. 2023b. DOI: `https://doi.org/10.2436/20.8050.xx.x`.

C. De Vos, E. Van der Kooij, and O. Crasborn. 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, 2009.

M. Oomen, L. Esselink, T. De Ronde, and F. Roelofsen. First Steps Towards a Procedure for Annotating Non-Manual Markers in Sign Languages, 2023. Part of the project Questions in Sign Language (grant number VI.C.201.014, PI Roelofsen).

D. Metaxas, B. Liu, F. Yang, P. Yang, N. Michael, and C. Neidle. 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, 2012.

B. Liu, J. Liu, X. Yu, D. Metaxas, and C. Neidle. 3D face tracking and multiscale, spatiotemporal 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, 2014.

A. Kuznetsova, A. Imashev, M. Mukushev, A. Sandygulova, and V. Kimmelman. 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). . 2022.

V. Kimmelman, A. Imashev, M. Mukushev, and A. Sandygulova. Eyebrow position in grammatical and emotional expressions in kazakh-russian sign language: A quantitative study. plos one, 15(6):e0233731. pages 49–59, 2020.

T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. in 2018 13th ieee international conference on automatic face gesture recognition (fg 2018). pages 59–66, 2018.

D. Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford university press, 1989.

F. Domaneschi, M. Romero, and B. Braun. Bias in polar questions: Evidence from English and German production experiments. *Glossa: a journal of general linguistics*, 2(1):1–28, 2017.

L. Esselink, M. Oomen, and F. Roelofsen. Truedepth measurements of facial expressions: Sensitivity to the angle between camera and face. In 2023 IEEE International Conference on Acoustics, Speech and Signal Processing Workshops (ICASSPW). pages 1–5, 2023.

Tentacle Sync. Tentacle sync studio, 2023a. URL: `https://tentaclesync.com/sync-studio` [Accessed: 13/09/2023].

Tentacle Sync. Tentacle apps, 2023b. URL: `https://tentaclesync.com/tentacle-apps` [Accessed: 09/12/2023].

Adobe. Adobe premiere pro, 2023. URL: `https://www.adobe.com/nl/products/premiere/campaign/pricing.html?gclid=CjwKCAiA1MCrBhAoEiwAC2d64fv9sHO66opvOoN8F66geDxDBavqKGWM_k-yKxmucn0YYJK9O8hghhoCudgQAvD_BwE&mv=search&mv=search&mv2=paidsearch&sdid=G4FRYP7G&ef_id=CjwKCAiA1MCrBhAoEiwAC2d64fv9sHO66opvOoN8F66geDxDBavqKGWM_k-yKxmucn0YYJK9O8hghhoCudgQAvD_BwE:G:s&s_kwcid=AL!3085!3!600767916159!e!!g!!adobe%20premiere%20pro!1441877182!60095930801&gad_source=1` [Accessed: 06/12/2023].

Posit Software. Rstudio desktop, 2023. URL: `https://posit.co/download/rstudio-desktop/` [Accessed: 07/12/2023].

Apple. ARKit blendShapes - Apple Developer Documentation., 2023. URL: `https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation` [Accessed: 18/12/2023].

MPI. Max Planck Institute for Psycholinguistics, 2023. URL: `https://www.mpi.nl/nl` [Accessed: 11/12/2023].

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: a Professional Framework for Multimodality Research. In 5th international conference on language resources and evaluation (LREC 2006). pages 1556–1559, 2006.

M. Sugahara, S. Silva, M. Scattolin, F. Cruz, J. Perissinoto, and A. Tamanaha. Exploratory study on the multimodal analysis of the joint attention. *Audiology-Communication Research*, 27, 2022.

S. Turchyn, I. Moreno, C. Cánovas, F. Steen, M. Turner, J. Valenzuela, and S. Ray. Gesture annotation with a visual search engine for multimodal communication research. In Proceedings of the AAAI Conference on Artificial Intelligence. 32 (1), 2018.

P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program], 2023. URL: `http://www.praat.org/` [Accessed: 11/12/2023].

G. Van Rossum and F. Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

L. McInnes, J. Healy, and S. Astels. How HDBSCAN works. 2016. URL: `https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html` [Accessed: 03/01/2024].

Appendix A

Experiment Design and Data Collection

The information presented in this appendix is mainly taken from Oomen and Roelofsen [2023a], who reported on the design of the BpQ/NGT-experiment [Esselink, 2022]. Their appendix was modified, as to fit the current design.

We created five experimental situations and one practice situation. Confederate responses provide positive ('+'), neutral ('0'), or negative ('–') original speaker bias or contextual evidence for the target question (final participant utterance).

A.1 Practice situation: Is there a metro station nearby?

1. *Original speaker bias*

Context 1: You recently moved to Amsterdam. You are currently at your house, but would like to go the city center. You don't know if there's a metro station nearby. You're meeting Robin, your new neighbor. Ask her.

Participant: "Is there a metro station nearby?"

Confederate A:
+ "Yes, there is a metro station around the corner."
0 "I don't know, I never take the metro."
− "No, there's no metro station nearby."

2. *Contextual evidence*

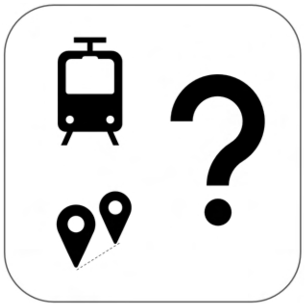
Context 2: On your way, you meet Sam. Ask her whether she knows the best way to the city center.

Participant: "Do you know the best way to the city center?"

Confederate B:
+ "There's a metro station here around the corner. You should take line 51 to Weesperplein, which is close to the city center."
0 "It's best to go by public transport."
− "You can't take the metro, because there's no metro station nearby. You should go by bike."

3. *Target question*

Picture prompt:



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

A.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. Robin 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*

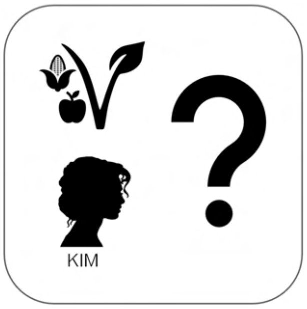
Context 2: You and Sam are cooking dinner together. You're making meatballs. Ask Sam how many meatballs you should make.

Participant: "How many meatballs should we make?"

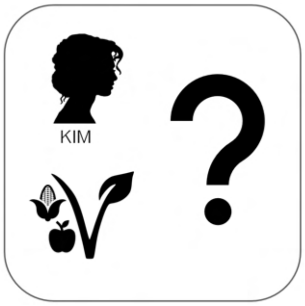
Confederate B:
+ "You don't have to make any for Kim, she is a vegetarian"
0 "Let's make two for everyone, except for the vegetarians."
− "We should definitely make enough for Kim, she loves them!"

3. *Target question*

Picture prompt:



(Version 1)



(Version 2)

Participant: Variation on "Is Kim a vegetarian?"

A.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 Robin, 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***

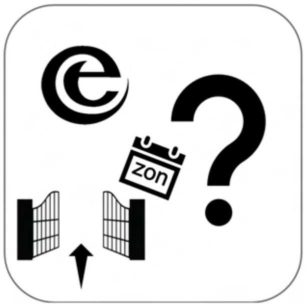
Context 2: Later that day, you meet Sam. She works at the Efteling. You know she has the weekend off. Ask her if she'd like to come to the Efteling with you this weekend.

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

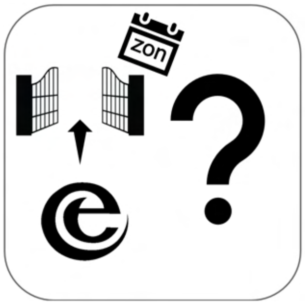
Confederate B:
+ "Fun! Shall we go on Sunday?"
0 "I can't this weekend."
– "The Efteling is only open on Saturday. I'm available then."

3. ***Target question***

Picture prompt:



(Version 1)



(Version 2)

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

A.4 Situation 3: 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. Robin 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?"

A.5 Situation 4: Is Kim home?

1. *Original speaker bias*

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

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*

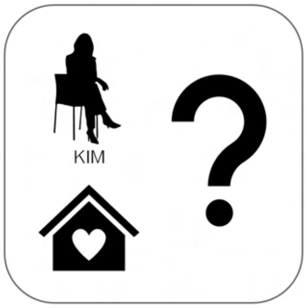
Context 2: On Saturday morning, you unexpectedly have to return home early, but you forgot your keys. On the way home, you call Sam; you can't get a hold of Kim. Ask Sam if Kim could open the door for you.

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

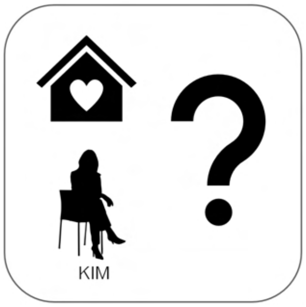
Confederate B:
+ "Yes, I just talked to her and she's there."
0 "I don't know. You should send her a text."
– "Kim is away for the weekend."

3. *Target question*

Picture prompt:



(Version 1)



(Version 2)

Participant: Variation on "Is Kim home?"

A.6 Situation 5: 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. Robin 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***

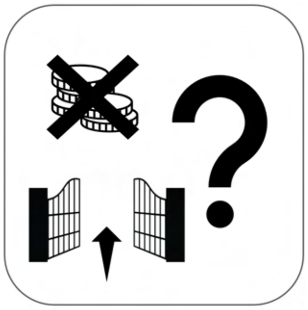
Context 2: A day later, you're at the Veluwe parking lot. You can't find the entrance to the park. At the parking lot, you meet Sam, another visitor to the park. Ask her.

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

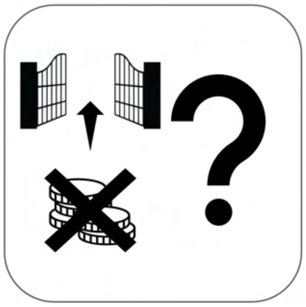
Confederate B:
+ "The entrance is there by the white flag. You don't need a ticket."
0 "The entrance is there by the white flag."
− "The entrance is there by the white flag, but you need to get a ticket at the ticket counter over there first."

3. ***Target question***

Picture prompt:



(Version 1)

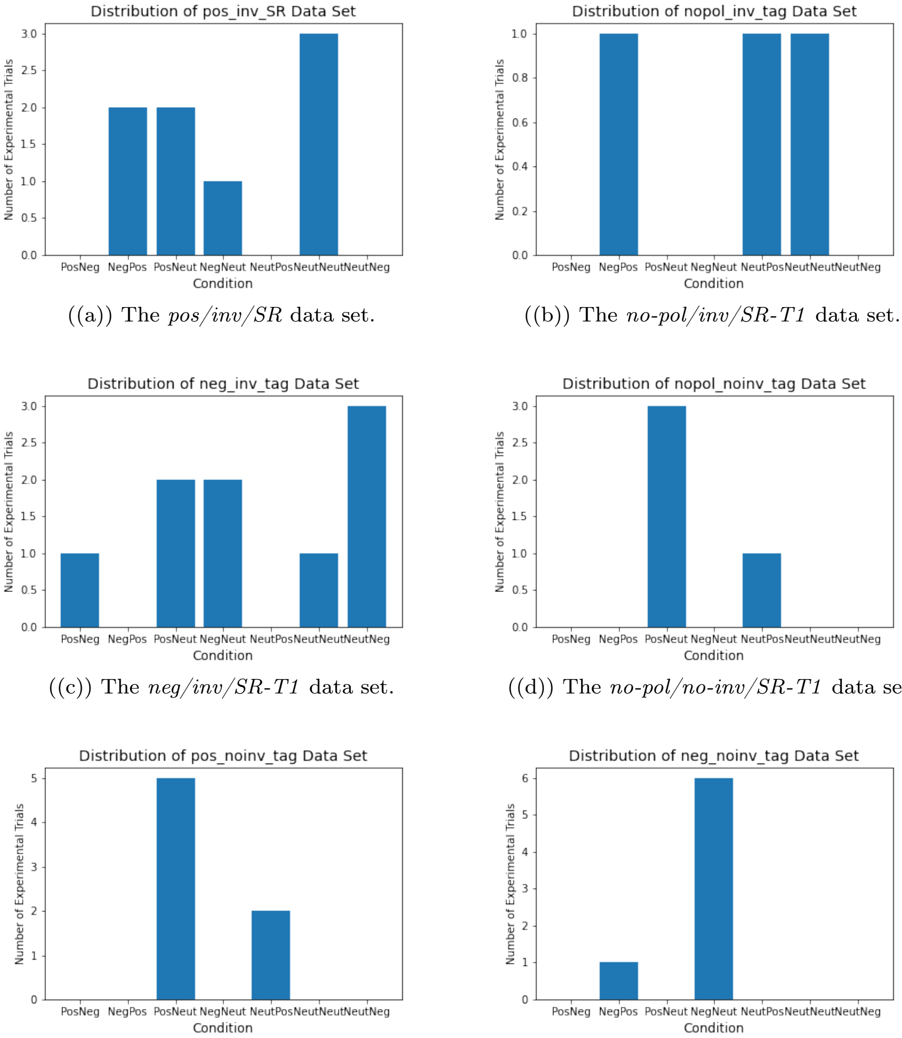


(Version 2)

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

Appendix B

Results: Annotations



((a)) The *pos/inv/SR* data set.

((b)) The *no-pol/inv/SR-T1* data set.

((c)) The *neg/inv/SR-T1* data set.

((d)) The *no-pol/no-inv/SR-T1* data set.

((e)) The *pos/no-inv/SR-T1* data set.

((f)) The *neg/no-inv/SR-T1* data set.

Figure B.1: Distribution of the less frequent question structures over conditions. Note that the distribution for the *pos/inv/SR-T1* data set is not visualised, since this data set is empty.

Appendix C

Analysis of 3D Data

As was previously stated, Esselink [2022] found that performing the HDBScan algorithm on the normalised data set did not provide sufficient results, whilst this was the case for the corresponding categorised data. Accordingly, as was discussed in Section 5.1.1, the data was categorised (after normalisation) in multiple ways; the bins contained various bounds and sizes. Table C.1 provides an overview of two of these categorisations, describing both the bounds and the replacement value of a features' measurement falling within these bounds. Table C.2 presents an example, containing a subset of the normalised data set and the corresponding replacement values for $B1$.

For all categorisations, the replacement value of the first bin equalled 0. Given that there is a high likelihood that participants did not engage certain features, hence their corresponding blend shape value was measured at 0, these values also needed to equal 0 in the categorised data set. Taking the mean of the lower and upper bound, as was the case for most of the remaining bins, would result in a data set containing all data points above 0, which is not representative of the captured data. Furthermore, if a measured blend shape value was relatively small, therefore falling into the first bin, this feature was barely active in the corresponding facial expression. Hence, the engagement of this feature is negligible, therefore the

		Bins							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
B1	*lower bound*	0	10	32.5	55	77.5			
	upper bound	10	32.5	55	77.5	100			
	repl. value	0	21.25	43.75	62.25	88.75			
B2	*lower bound*	0	5	10	15	20	30	45	65
	upper bound	5	10	15	20	30	45	65	100
	repl. value	0	7.5	12.5	17.5	25	37.5	55	82.5

Table C.1: Two implementations of categorisation. Lower bounds, upper bounds, and the replacement value of data points falling within these bounds are presented.

	f_1	f_2	f_3	f_4
v_1	43	68	3	37
v_2	46	69	3	34
v_3	56	70	4	23

((a)) Original Values

	f_1	f_2	f_3	f_4
v'_1	43.75	62.25	0	43.75
v'_2	43.75	62.25	0	43.75
v'_3	66.25	62.25	0	21.25

((b)) Categorised Values

Table C.2: Comparison between original and categorised blend shape values, using categorisation $B1$. Values contained in the left table are rounded to the nearest integer.

corresponding data point was set to 0.

As stated, the decision was first made to categorise the data into five bins, see Table C.1. This data was then fed to the HDBScan clustering algorithm, which yielded the results presented in Table C.3. This table reports on the mean blend shape values of the nine selected features, as well as the distribution of the clusters over participants. One can see that 18 clusters were formed, apart from the 'cluster' N, containing all data points labelled as noise. Here, the implemented clustering was quite lenient, therefore resulting in fine-grained clusters, some of which being fairly similar. There are three main gripes with these results. The first is that the distribution over participants is not optimal: most clusters are formed on the basis of facial expressions exhibited by few participants. Therefore, these results can not be generalised. The second is that contrasting features, such as *EyeSquint* and *EyeWide*, seem to be regularly engaged at the same time, which is not in line with our expectations. These findings could result from bins not being discriminatory enough, therefore producing clusters with skewed mean blend shape values. For instance, when participants eyes are slightly squinted, but wide to a higher degree, it is possible that both of the features are measured between 10 and 32.5. Therefore, in the categorised data set, these are set to the same value, suggesting that these features were equally engaged in the corresponding facial expression. Lastly, none of these clusters contain a feature which is engaged to a high extent, say above 70, which sternly contrasts the BpQ/NGT-experiment for which this was the case. This suggests that the upper two bins were redundant, since barely any values fall within the corresponding bounds. This is further confirmed by the preliminary analysis of the video data, reporting that participants did not exhibit many visual cues.

Thus, from these first formed clusters, the main takeaway was that for the lower values, the bins are not discriminatory enough, whilst for the higher values, these bins are redundant. As a next step, the distribution of values contained in the normalised data set was visualised, and is presented in Figure C.1. As was expected, for each feature, most values fall within the range of $0-20$, and barely any features were engaged to an extent over 60. Therefore, based on this distribution, the data was re-categorised, using the bins ($B2$) Table C.1 describes. In the end, this choice of bins provided the most representative results, hence the resulting clusters and their distributions over participants are presented in Section 6.2.

Another attempt was made to spread the distribution over participants to a

	S_t	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N	29									
C_1	4	44	12	13	0	0	6	0	11	7
C_2	2	21	8	16	0	0	0	0	44	21
C_3	3	44	14	17	0	0	1	0	44	21
C_4	8	23	50	2	21	17	1	13	5	0
C_5	3	44	44	1	13	11	1	9	7	0
C_6	4	19	10	66	1	0	0	5	1	19
C_7	2	12	6	44	2	0	0	2	4	9
C_8	4	19	7	13	0	0	0	0	21	21
C_9	3	21	12	17	1	0	0	0	21	0
C_{10}	2	0	0	0	9	0	1	0	0	21
C_{11}	11	0	0	0	0	0	0	0	0	0
C_{12}	4	0	21	0	0	0	0	0	0	0
C_{13}	5	0	0	21	2	0	0	0	0	2
C_{14}	2	21	5	8	5	0	17	1	21	5
C_{15}	7	21	0	0	0	0	0	0	0	0
C_{16}	3	21	21	0	0	0	0	0	0	0
C_{17}	3	21	0	21	0	0	0	0	0	0
C_{18}	2	21	21	21	0	0	0	0	0	0

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (*ES*), *EyeWide* (*EW*), *BrowDown* (*BD*), *BrowInnerUp* (*BIU*), *BrowOuterUp* (*BOU*), *MouthShrug* (*MS*), *MouthFrown* (*MF*), *CheekSquint* (*CS*) and *NoseSneer* (*NS*).

	S_t	S_c **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N	29									
C_1	4	16	0	15	1	14	25	14	9	5
C_2	2	0	4	82	0	0	0	0	14	0
C_3	3	0	0	82	0	0	0	7	10	0
C_4	8	32	0	0	51	4	2	3	0	8
C_5	3	67	0	0	2	5	7	0	0	19
C_6	4	0	13	0	0	87	0	0	0	0
C_7	2	0	54	0	0	43	0	0	0	3
C_8	4	2	5	60	0	8	0	1	22	1
C_9	3	10	11	0	6	0	7	9	56	1
C_{10}	2	3	2	0	54	28	1	0	11	2
C_{11}	11	1	23	0	24	2	26	9	14	0
C_{12}	4	0	0	0	50	0	1	46	3	0
C_{13}	5	2	67	0	6	9	0	0	15	0
C_{14}	2	3	0	0	0	0	17	58	18	3
C_{15}	7	4	3	0	5	8	44	36	0	0
C_{16}	3	6	0	0	15	10	25	43	0	0
C_{17}	3	17	14	0	4	30	16	0	18	0
C_{18}	2	18	0	0	1	44	6	0	32	0

((b)) The formed clusters and their distribution over participants.

Table C.3: Clusters formed by the HDBScan algorithm using categorisation $B1$, the corresponding mean feature values and their distributions over participants. Samples assigned to clusters (S_c) are given as a percentage of total samples (S_t).

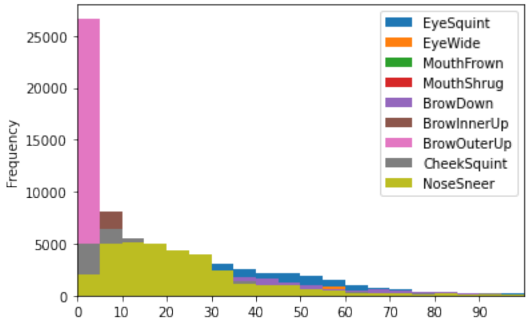


Figure C.1: Distribution of blend shape values for all selected features.

more even extent. For each participant and for each feature, the upper bounds of the quartiles containing 25% of data were obtained. For each of these quartiles, the mean value of data points contained in this quartile was calculated. This was now the replacement value of the data points that fall within this quartile. For instance, for the feature *BrowInnerUp*, a quarter of p3's data points falls below value 7.49. The mean of these data points equals 4.39, hence all values in this quarter now have this as their replacement value. In this way, for each participant, 36 bins were specified. The resulting clusters by HDBScan are presented in Table C.4. As one can see, the clustering algorithm performed in a way exactly opposite to what was expected: all clusters were formed based on data by one participant, and no data points were labelled as noise. One explanation for these findings could be that the clusters are participant-specific, since the blend shape values of features contained in these clusters are specific to each participant as well. Thus, when a participant presents a certain facial expression, during the entirety of their target question utterance, all of the data points have the same value, for each feature. When a different participant exhibits this combination of facial features, the blend shape values are specific to this other participant, and could therefore not be matched with the values exhibited by the previous participant. This is further exacerbated by the fact that participant data is likely very dense: not a lot of differences between captured values of frames are found in target question utterances for each participant. Consequently, the distances between maximal points within these clusters, based solely on the data of one participant, is presumably quite high.

Since the formed clusters, based on this categorised data set, were all formed based on data from only one participant, these results could not be generalised. Unfortunately, the scope of this project prohibited the implementation of more categorisations of the data and their results. Consequently, the decision was made to categorise the data using the bins $B2$ as presented in Table C.1, although these results do not provide a space that is optimal for drawing conclusions. Therefore, the results presented in this project are based on a *preliminary* analysis. See Section 6.2, as well as Appendix E.

	S_t	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N	0									
C_1	13	55	46	32	15	1	12	87	25	44
C_2	11	44	45	16	14	10	26	19	23	31
C_3	12	58	33	26	13	5	10	18	23	33
C_4	11	48	34	26	16	11	31	14	22	29
C_5	10	52	46	29	17	2	9	7	22	23
C_6	10	63	26	16	29	15	14	10	18	35
C_7	9	48	24	19	7	8	35	15	31	18
C_8	12	59	38	26	23	0	11	14	29	35
C_9	12	55	40	22	25	9	20	14	26	34

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (ES), *EyeWide* (EW), *BrowDown* (BD), *BrowInnerUp* (BIU), *BrowOuterUp* (BOU), *MouthShrug* (MS), *MouthFrown* (MF), *CheekSquint* (CS) and *NoseSneer* (NS).

	S_t	S_c **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N	0									
C_1	13	0	0	100	0	0	0	0	0	0
C_2	11	0	0	0	100	0	0	0	0	0
C_3	12	0	0	0	0	100	0	0	0	0
C_4	11	100	0	0	0	0	0	0	0	0
C_5	10	0	100	0	0	0	0	0	0	0
C_6	10	0	0	0	0	0	0	100	0	0
C_7	9	0	0	0	0	0	100	0	0	0
C_8	12	0	0	0	0	0	0	0	100	0
C_9	12	0	0	0	0	0	0	0	0	100

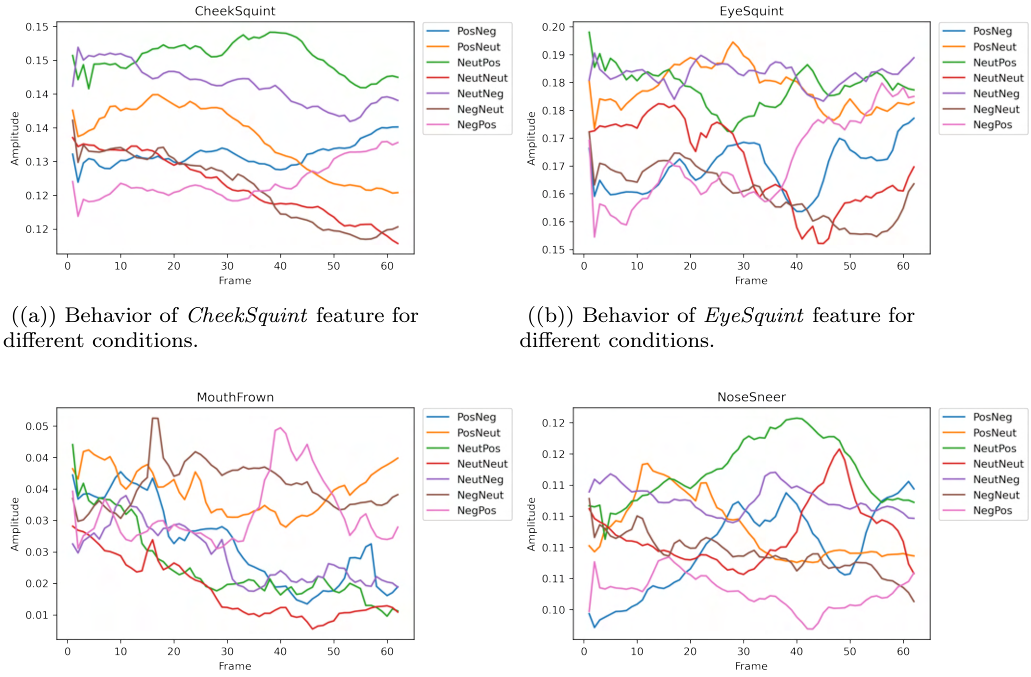
((b)) The formed clusters and their distribution over participants.

Table C.4: Clusters formed by the HDBScan algorithm using participant-specific categorisation, the corresponding mean feature values and their distributions over participants. Samples assigned to clusters (S_c) are given as a percentage of total samples (S_t).

Appendix D

Results: Visualisations

D.1 Complete Ranged Data Set



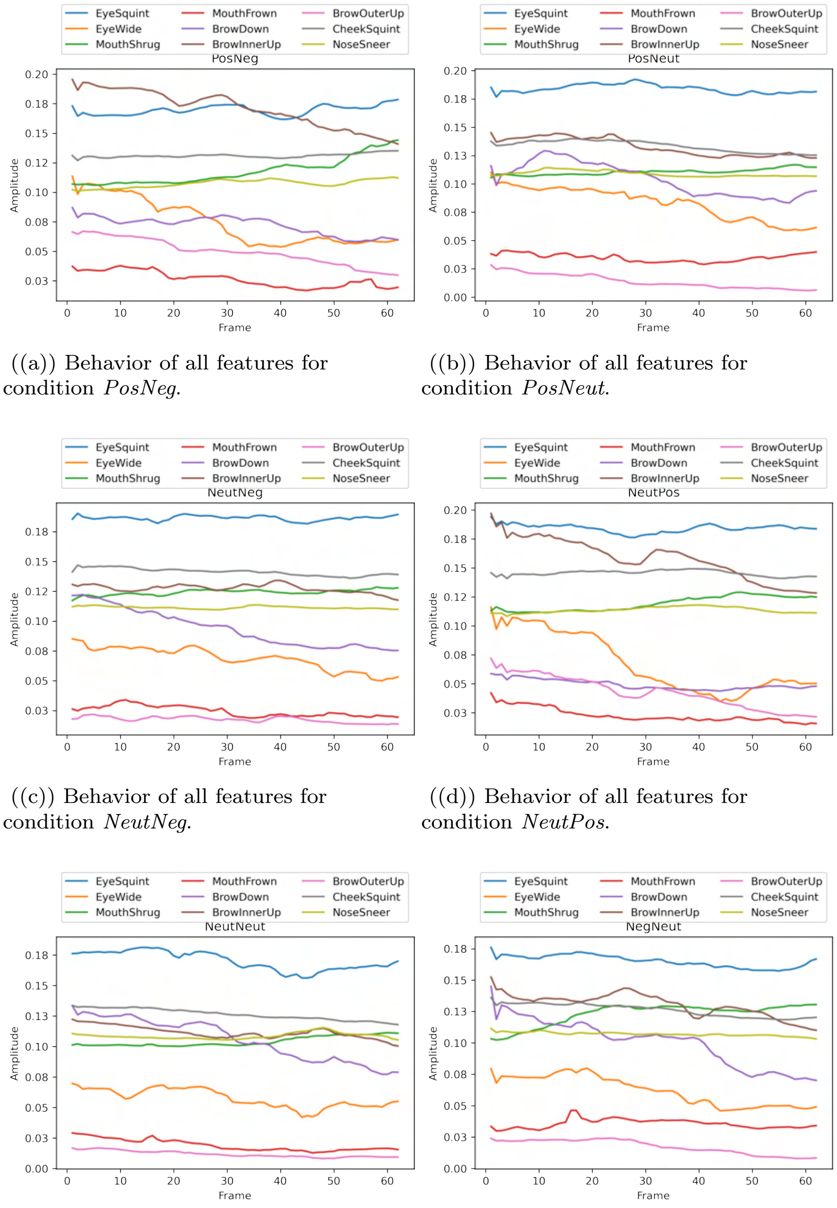
((a)) Behavior of *CheekSquint* feature for different conditions.

((b)) Behavior of *EyeSquint* feature for different conditions.

((c)) Behavior of *MouthFrown* feature for different conditions.

((d)) Behavior of *NoseSneer* feature for different conditions.

Figure D.1: Visualisation of mean movements over time for the features *CheekSquint*, *EyeSquint*, *MouthFrown* and *NoseSneer*.



((a)) Behavior of all features for condition *PosNeg*.

((b)) Behavior of all features for condition *PosNeut*.

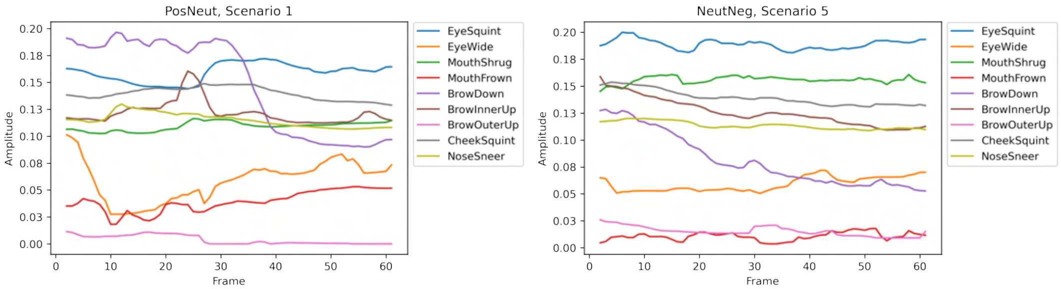
((c)) Behavior of all features for condition *NeutNeg*.

((d)) Behavior of all features for condition *NeutPos*.

((e)) Behavior of all features for condition *NeutNeut*.

((f)) Behavior of all features for condition *NegNeut*.

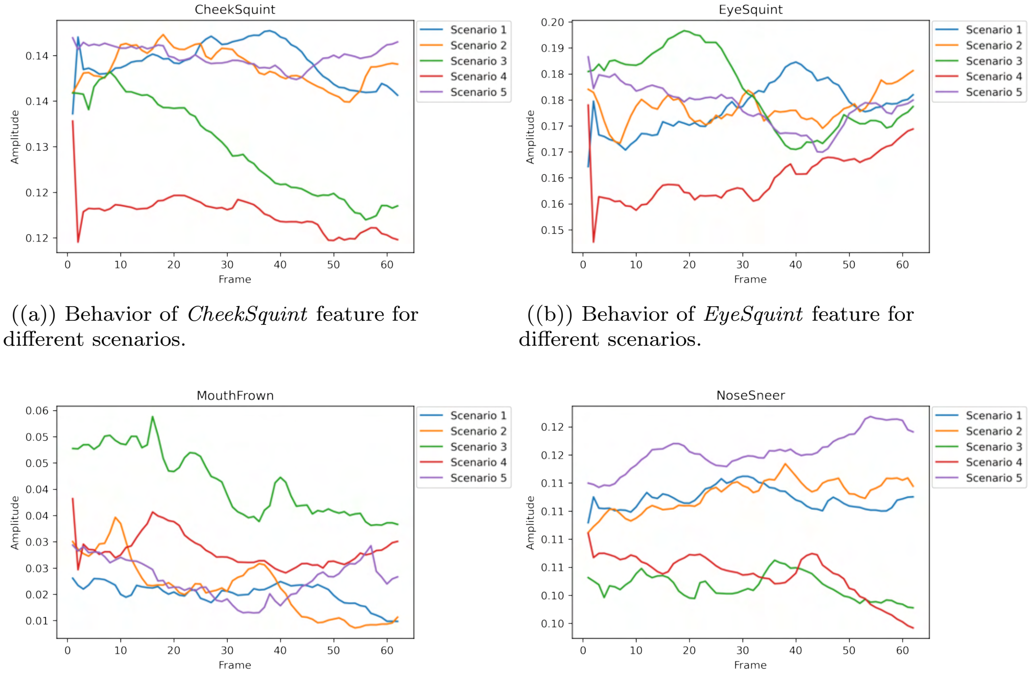
Figure D.2: Visualisation of mean movements over time for each feature in conditions *PosNeg*, *PosNeut*, *NeutNeg*, *NeutPos*, *NeutNeut* and *NegNeut*.



((a)) Behavior of all features during scenario 1 condition *PosNeut*.

((b)) Behavior of all features during scenario 5 condition *NeutNeg*.

Figure D.3: Visualisation of mean movements of features over time for the conditions *PosNeut* scenario 1 and *NeutNeg* scenario 5. The *BrowDown* feature is used less over time.



((a)) Behavior of *CheekSquint* feature for different scenarios.

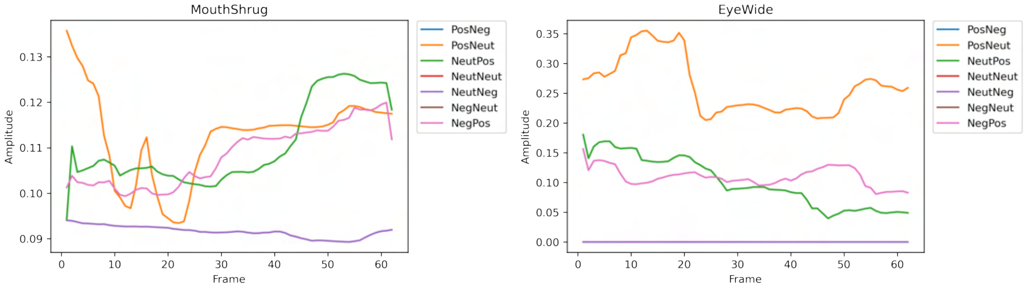
((b)) Behavior of *EyeSquint* feature for different scenarios.

((c)) Behavior of *MouthFrown* feature for different scenarios.

((d)) Behavior of *NoseSneer* feature for different scenarios.

Figure D.4: Visualisation of mean movements over time for the features *CheekSquint*, *EyeSquint*, *MouthFrown* and *NoseSneer* in each scenario.

D.2 Subsets of Complete Ranged Data Set



((a)) Behavior of *MouthShrug* feature for different conditions.

((b)) Behavior of *EyeWide* feature for different conditions.

Figure D.5: Visualisation of mean movements over time for the features *MouthShrug* and *EyeWide*, for each condition. Obtained from the *pos/no-inv/SR* data set.

Appendix E

Results: HDBScan Clustering

	S_t''	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N''	7									
C_1''	89	20	13	8	11	4	7	4	13	8
C_2''	4	39	83	0	35	34	5	17	9	0

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (ES), *EyeWide* (EW), *BrowDown* (BD), *BrowInnerUp* (BIU), *BrowOuterUp* (BOU), *MouthShrug* (MS), *MouthFrown* (MF), *CheekSquint* (CS) and *NoseSneer* (NS).

	S_t''	S_c'' **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N''	7									
C_1''	89	11	2	7	18	6	0	18	22	16
C_2''	4	59	0	0	41	0	0	0	0	0

((b)) The formed clusters and their distributions over participants.

Table E.1: Clusters formed by the HDBScan algorithm for the *neg/no-inv/SR* data set, the corresponding mean feature values and their distributions over participants. Samples assigned to clusters (S_c'') are given as a percentage of total samples (S_t'').

	S_t''	S_c'' **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
N''	7							
C_1''	89	5	4	44	0	0	4	44
C_2''	4	0	0	40	0	0	0	60

((a)) The formed clusters and their distributions over conditions.

	S_t''	S_c'' **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
N''	7					
C_1''	89	19	20	20	20	21
C_2''	4	19	26	27	26	2

((b)) The formed clusters and their distributions over temporal windows.

Table E.2: Clusters formed by the HDBScan algorithm for the *neg/no-inv/SR* data set and their distributions over conditions and temporal windows. Samples assigned to clusters (S_c'') are given as a percentage of total samples (S_t'').

	S_t'''	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N'''	33									
C_1'''	12	22	33	0	41	41	2	11	7	0
C_2'''	55	13	8	7	6	2	3	1	9	6

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (ES), *EyeWide* (EW), *BrowDown* (BD), *BrowInnerUp* (BIU), *BrowOuterUp* (BOU), *MouthShrug* (MS), *MouthFrown* (MF), *CheekSquint* (CS) and *NoseSneer* (NS).

	S_t'''	S_c''' **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N'''	33									
C_1'''	12	96	0	0	0	0	0	0	0	3
C_2'''	55	4	7	7	23	7	0	14	36	1

((b)) The formed clusters and their distributions over participants.

Table E.3: Clusters formed by the HDBScan algorithm for the *pos/no-inv/SR* data set, the corresponding mean values and their distributions over participants. Samples assigned to clusters (S_c''') are given as a percentage of total samples (S_t''').

	S_t'''	S_c''' **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
N'''	33							
C_1'''	12	49	0	0	0	51	0	0
C_2'''	55	48	0	7	0	45	0	0

((a)) The formed clusters and their distributions over conditions.

	S_t'''	S_c''' **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
N'''	33					
C_1'''	12	19	17	18	20	27
C_2'''	55	18	20	20	21	21

((b)) The formed clusters and their distributions over temporal windows.

Table E.4: Clusters formed by the HDBScan algorithm for the *pos/no-inv/SR* data set and their distributions over conditions and temporal windows. Samples assigned to clusters (S_c''') are given as a percentage of total samples (S_t''').

	$S_t^{''''}$	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
N''''	66									
C_1''''	8	0	0	0	0	2	0	0	0	0
C_2''''	19	13	3	4	0	0	0	0	1	0
C_3''''	8	0	0	16	0	0	0	0	0	1

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (ES), *EyeWide* (EW), *BrowDown* (BD), *BrowInnerUp* (BIU), *BrowOuterUp* (BOU), *MouthShrug* (MS), *MouthFrown* (MF), *CheekSquint* (CS) and *NoseSneer* (NS).

	$S_t^{''''}$	$S_c^{''''}$ **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
N''''	66									
C_1''''	8	0	56	0	18	0	26	0	0	0
C_2''''	19	0	20	0	0	14	67	0	0	0
C_3''''	8	0	94	0	0	6	0	0	0	0

((b)) The formed clusters and their distributions over participants.

Table E.5: Clusters formed by the HDBScan algorithm for the *neg/inv/SR* data set, the corresponding mean values and their distributions over participants. Samples assigned to clusters ($S_c^{''''}$) are given as a percentage of total samples ($S_t^{''''}$).

	$S_t^{''''}$	$S_c^{''''}$ **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
N''''	66							
C_1''''	8	0	0	87	0	0	0	13
C_2''''	19	11	0	47	3	0	0	39
C_3''''	8	0	0	46	6	0	0	48

((a)) The formed clusters and their distributions over conditions.

	$S_t^{''''}$	$S_c^{''''}$ **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
N''''	66					
C_1''''	8	35	20	11	22	12
C_2''''	19	9	23	20	23	24
C_3''''	8	24	19	18	28	10

((b)) The formed clusters and their distributions over temporal windows.

Table E.6: Clusters formed by the HDBScan algorithm for the *neg/inv/SR* data set and their distributions over conditions and temporal windows. Samples assigned to clusters ($S_c^{''''}$) are given as a percentage of total samples ($S_t^{''''}$).

	$S_t^{'''''}$	**Most engaged features (mean value)**								
C	%	ES	EW	BD	BIU	BOU	MS	MF	CS	NS
$N^{'''''}$	49									
$C_1^{'''''}$	38	13	7	5	5	2	3	1	5	3
$C_2^{'''''}$	13	32	7	7	6	0	5	0	26	15

((a)) The formed clusters and their corresponding mean values for each feature. Features are abbreviated: *EyeSquint* (*ES*), *EyeWide* (*EW*), *BrowDown* (*BD*), *BrowInnerUp* (*BIU*), *BrowOuterUp* (*BOU*), *MouthShrug* (*MS*), *MouthFrown* (*MF*), *CheekSquint* (*CS*) and *NoseSneer* (*NS*).

	$S_t^{'''''}$	$S_c^{'''''}$ **distribution** (%)								
C	%	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
$N^{'''''}$	49									
$C_1^{'''''}$	38	24	0	0	6	17	13	39	0	0
$C_2^{'''''}$	13	0	0	31	0	27	3	39	0	0

((b)) The formed clusters and their distributions over participants.

Table E.7: Clusters formed by the HDBScan algorithm for the *no-pol/no-inv/SR* data set, the corresponding mean values and their distributions over participants. Samples assigned to clusters ($S_c^{'''''}$) are given as a percentage of total samples ($S_t^{'''''}$).

	$S_t^{'''''}$	$S_c^{'''''}$ **distribution** (%)						
C	%	*NegPos*	*NegNeut*	*NeutNeg*	*NeutNeut*	*NeutPos*	*PosNeut*	*PosNeg*
$N^{'''''}$	49							
$C_1^{'''''}$	58	13	0	0	0	60	5	22
$C_2^{'''''}$	13	11	0	0	0	61	3	25

((a)) The formed clusters and their distributions over conditions.

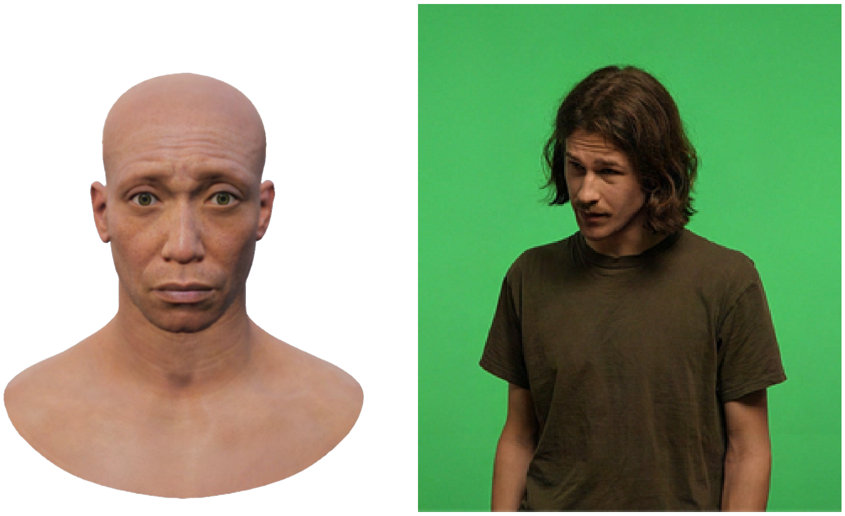
	$S_t^{'''''}$	S_c **distribution** (%)				
C	%	w_1	w_2	w_3	w_4	w_5
$N^{'''''}$	49					
$C_1^{'''''}$	38	19	18	22	18	23
$C_2^{'''''}$	13	15	25	30	24	6

((b)) The formed clusters and their distributions over temporal windows.

Table E.8: Clusters formed by the HDBScan algorithm for the *no-pol/no-inv/SR* data set and their distributions over conditions and temporal windows. Samples assigned to clusters ($S_c^{'''''}$) are given as a percentage of total samples ($S_t^{'''''}$).

Appendix F

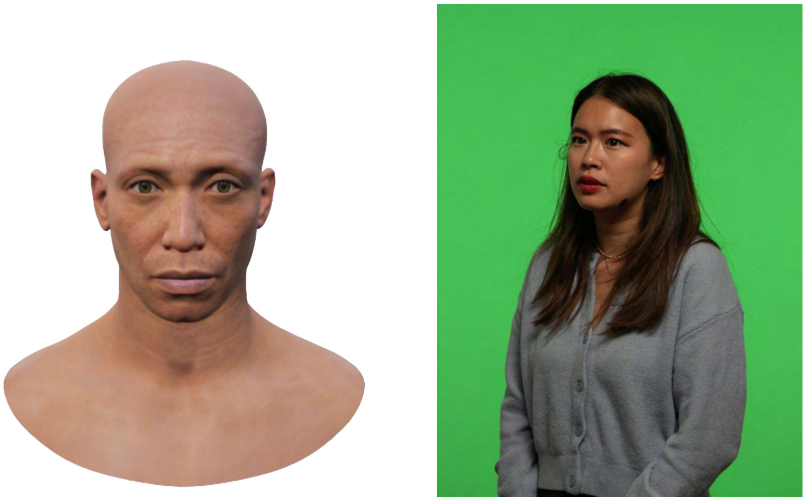
Results: Prototypical Expressions



((a)) Cluster 1 (Metahuman)

((b)) Cluster 1 (scenario 2 condition *Pos-Neut*)

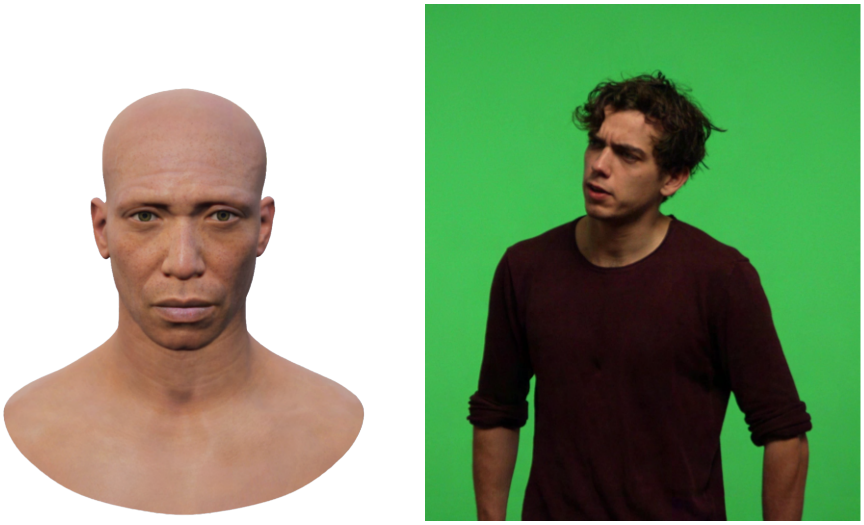
Figure F.1: Visualisation of facial expression corresponding to C_1.



((a)) Cluster 2 (Metahuman)

((b)) Cluster 2 (scenario 1 condition *NeutPos*)

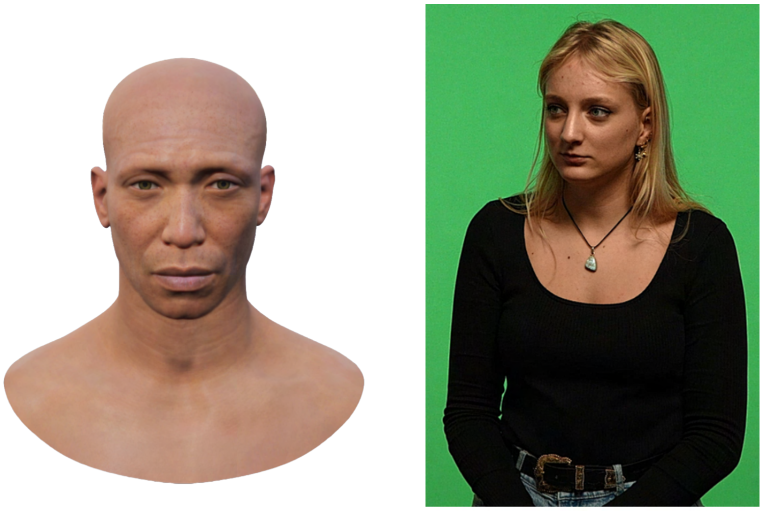
Figure F.2: Visualisation of facial expression corresponding to C_2.



((a)) Cluster 3 (Metahuman)

((b)) Cluster 3 (scenario 2 condition *PosNeg*)

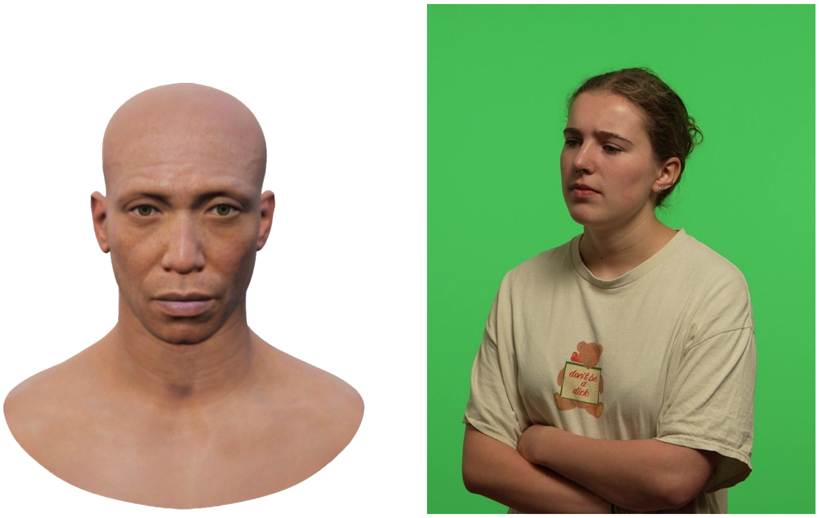
Figure F.3: Visualisation of facial expression corresponding to C_3.



((a)) Cluster 4 (Metahuman)

((b)) Cluster 4 (scenario 4 condition *PosNeut*)

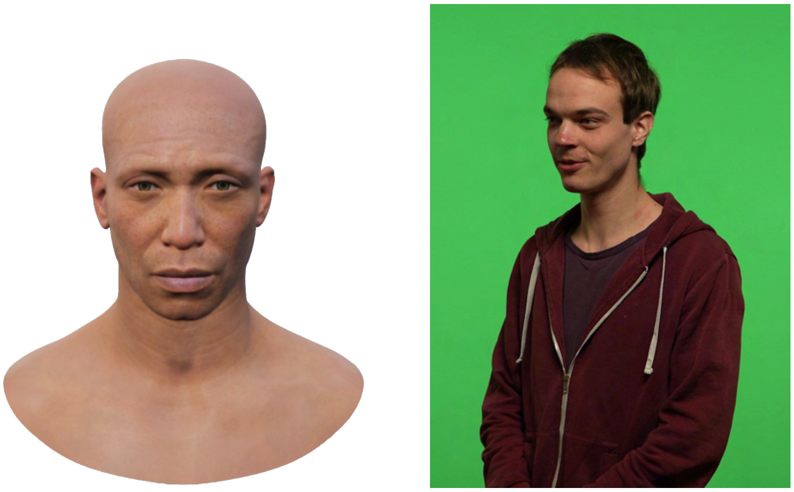
Figure F.4: Visualisation of facial expression corresponding to C_4.



((a)) Cluster 5 (Metahuman)

((b)) Cluster 5 (scenario 3 condition *PosNeut*)

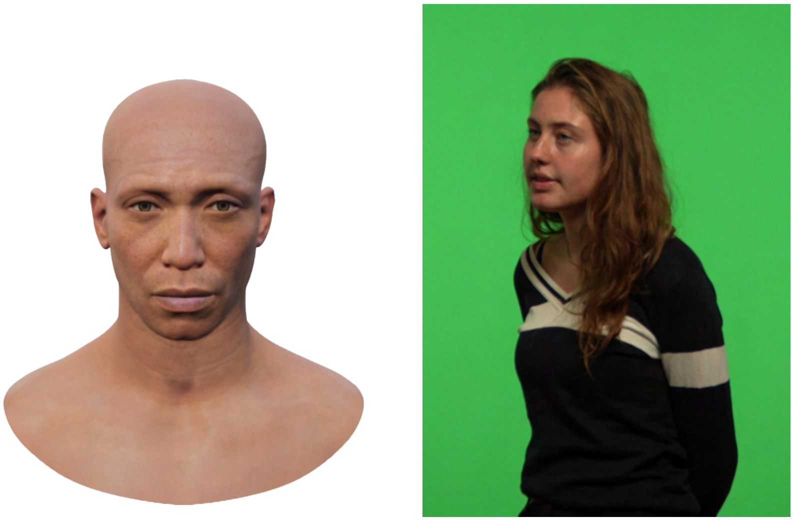
Figure F.5: Visualisation of facial expression corresponding to C_5.



((a)) Cluster 6 (Metahuman)

((b)) Cluster 6 (scenario 1 condition *PosNeut*)

Figure F.6: Visualisation of facial expression corresponding to C_6.



((a)) Cluster 7 (Metahuman)

((b)) Cluster 7 (scenario 5 condition *NegPos*)

Figure F.7: Visualisation of facial expression corresponding to C_7.