UNIVERSITY OF TWENTE.

Towards a Conversational Agent for Learning Dutch as a Second Language

Creating More Speaking Opportunities for Students of the Delft Method

Lukas Elias Paul Hageman

Faculty of Electrical Engineering, Mathematics and Computer Science

Human Media Interaction

Master in Interaction Technology

Enschede, May 2025

UNIVERSITY OF TWENTE.

Towards a Conversational Agent for Learning Dutch as a Second Language

Creating More Speaking Opportunities for Students of the Delft Method

Lukas Elias Paul Hageman

Supervisor: Dr. Khiet Truong
Associate Professor, EEMCS

Co-supervisor: Simone Borsci
Associate Professor, BMS

Faculty of Electrical Engineering, Mathematics and Computer Science

Human Media Interaction

Master in Interaction Technology

I-Tech Final Project

Enschede, May 2025

Acknowledgements

I would like to express my heartfelt gratitude to the people who guided and supported me during the final phase of my studies at the University of Twente.

First and foremost, I want to thank my main supervisor, *Khiet Truong*. I truly appreciated her flexibility - both in scheduling meetings and in responding quickly to my online questions. Our way of working felt more as a collaboration than being supervised, which created a comfortable and productive working environment. At the same time, she gave me the freedom to work independently, which greatly contributed to the development of my scientific skills.

I am also very grateful to my other supervisors, *Simone Borsci* and *Mariët Theune* (although the latter was not officially my supervisor, she still was interested and involved), for their valuable and qualitative feedback. Their enthusiasm for my research topic inspired me and served as a strong motivation throughout the process.

I also had the pleasure of working with *Liesbeth Schenk* and *Astrid van Laar* from the Delft University of Technology. I truly enjoyed our collaboration, and I deeply appreciated their enthusiasm and flexibility. They were always willing to support me - be it with interviews, finding participants, or providing locations for conducting experiments.

A special thank you goes to my friends and family, who consistently showed interest in my work. I am especially thankful to my father, whose support was invaluable during the tougher moments.

Lastly, I would like to recognise my fellow board members of *E.E.E.H.V Cabezota*. Balancing a board year with thesis writing was not the easiest, but *Flinn* and *Rudy* supported me throughout, giving me the space and time I needed. Rudy, in particular, deserves extra thanks - as I am writing these final parts of my thesis on his emergency laptop.

Abstract

Learning a second language comes with its challenges, especially in the Netherlands, where you can comfortably survive with just English. There are few opportunities to practice Dutch, and even if there are, it can be daunting for learners to speak it with natives. This research aims to find a route towards an automatic conversational agent in the Delft method context, a method for teaching Dutch as a second language through a conversational approach. Such an agent would allow for more speaking opportunities and motivate learners to speak Dutch more. Two prototypes were developed through a design lifecycle, ending with an evaluation by potential end users. This evaluation showed that users would show more willingness to communicate in class to a certain extent if they could prepare with a conversational agent, which strategies work well to keep conversations flowing, and what design guidelines conversational agent developers can follow to realise a full-fledged end product for assisting in teaching Dutch as a second language.

Keywords: Conversational Agents, Willingness To Communicate, Second Language Learning, Delft Method

Contents

List of Figures

List of Tables

1

Introduction

1.1 Learning Dutch in the Netherlands

If you are a learner of Dutch, you might relate to the following situation: you ask someone in the Netherlands in your best Dutch where you can find the nearest supermarket. The person responds, but in English. You then ask, again in Dutch, what the supermarket's name is and what it looks like. Once more, you receive a reply in English.

These situations are quite common, as Dutch people speak English relatively well (*EF English Proficiency Index*, 2023) and thus tend to use English when they notice someone approaching them does not speak fluent Dutch. What the exact reasons are for the Dutch to switch to English is outside the scope of this research, but it certainly makes it harder for learners of Dutch as a second language (NT2 - common Dutch abbreviation for "Nederlands als tweede taal") to practice conversation skills. It results in little exposure to the spoken language, on top of the already frequent lack of possibilities for second language (L2) learners to talk to native speakers (Divekar et al., 2021; Hulstijn, 2018), even though it is known that exposure to a target language is important for learning it. A technological solution that could create more exposure could consist of conversational agents (CAs). With such agents, users can practice conversations at any time and any place, extending L2 classes and encouraging self-regulated learning (Moussalli & Cardoso, 2019). Hence, this work focuses on how CAs could assist in NT2 learning.

1.2 Language Anxiety

Next to the lack of conversation opportunities, language anxiety is also a problematic factor for L2 learning. Horwitz et al. (1986) introduced the concept of foreign

language[1] classroom anxiety (FLCA). Students with FLCA experience subjective feelings such as apprehension or worry, show psycho-physiological symptoms such as difficulty concentrating or sweating, and exhibit behavioural responses such as missing class and postponing homework when engaging in language learning. It concerns a situation-specific anxiety, meaning that anxiety arousal mainly occurs when one is faced with a moment when they have to use the target language. The authors also introduced a foreign language classroom anxiety scale (FLCAS) to measure the FLCA variable. From Horwitz (2001), it appears that many works using the FLCAS since then have found a negative correlation between FLCA and L2 performance. According to Horwitz et al. (1986), speaking in class is the most common worry for learners experiencing FLCA. In Dutch class, speaking to the teacher or fellow students in Dutch might be more anxiety-provoking for some students than writing a paragraph. Such an effect was shown for learners of English as a foreign language (EFL) in Indonesia, who experienced a moderate-to-serious level of specifically Foreign Language Speaking Anxiety (FLSA)(Bashori et al., 2022). The EFL learners indicated in interviews that when speaking in class, they felt nervous, and performed head-scratching, leg-shaking, or stammering more than when speaking was not the main focus.

While reviewing the literature on language anxiety, MacIntyre (2017) identified its primary causes and effects, describing it as the most broadly studied emotion in second language acquisition. According to their analysis, the causes and effects of language anxiety can be grouped into three categories: academic, cognitive, and social. Academically, language anxiety negatively impacts learners in classroom settings. Anxious students often achieve lower grades and may develop a decreased self-perception of L2 competence, for example. Cognitively, anxiety interferes with essential mental processes. It hinders the acquisition and retention of vocabulary and affects learners' ability to prepare for and perform in assessments. As a result, more typically nervous students may underperform compared to their more relaxed peers. Socially, language anxiety can reduce linguistic self-confidence and thus motivation to learn. Anxious learners tend to participate less often than relaxed learners, and students may even experience increased apprehension when using their native language.

FLCA partially causes learners to avoid L2 use in specific environments, also beyond the classroom. A person might keep talking in English to their Dutch friends or colleagues. They are then less willing to communicate in the target language in such situations. This, while willingness to communicate (WTC) is an important factor determining someone's language use (MacIntyre et al., 1998). Increasing individuals'

[1] In the literature, the terms foreign language and second language are both used. Strictly speaking, the essential difference is that second language learning is about a language spoken in the learner's local, academic, and/or work environment, while a foreign language is not per se. The current work focuses on learning Dutch as an international person in the Netherlands, i.e. a second language. However, most findings in the literature apply to both types of language learning. Therefore, the terms are used interchangeably in the current work.

WTC can even be seen as *the* primary goal of L2 learning. Whether NT2 learners are willing to communicate in Dutch depends on their motivation and self-confidence. If their WTC is higher, a language learner will get exposed to and use the target language more, which in turn results in a higher L2 level. So, one way to increase the WTC of NT2 learners is by countering or eliminating the causes of language anxiety. MacIntyre (2017) described these causes as follows: causes of language anxiety in the classroom context are errors in pronunciation, unrealistic learner beliefs, instructors who intimidate their students with harsh and/or embarrassing error correction in front of other students, and specific methods of testing. Cognitive causes include fear of losing one's sense of identity, biased perceptions of proficiency, personality traits and/or shyness, and low self-esteem. Lastly, the social causes are fear of being laughed at, embarrassed or making a fool of oneself, a poor-quality accent, misunderstanding communication or using incorrect words, cultural gaffes, competitiveness, and frequency and quality of contact with native speakers. When the latter is not sufficient, it is thus not only problematic in terms of being exposed to the target language itself, but also a cause of language anxiety. Next to that, language anxiety is both an effect and a cause of difficulties in language learning, meaning that this circle has to be broken to improve one's language learning.

1.3 Conversational Agents

One method within NT2 education particularly focuses on solving the academic causes of anxiety: the Delft method (DM)[2] (Van Boxtel et al., 2021). Initially developed at Delft University of Technology (TU Delft) for international students preparing to enrol in Dutch-taught study programmes, this approach focuses on instruction through extensive in-class conversations in Dutch. By offering frequent small tests, correcting speaking students selectively (so only large or important errors are corrected), stimulating and maintaining motivation through interim assessments, and comments on a student's learning process, an attempt is made to reduce language anxiety. Teachers are instructed to deal with anxious learners in a certain way: "Be extra careful with corrections and generous with your praise" (Wesdijk et al., 2021b).

Yet, students might still experience anxiety in and outside of class because of cognitive and social causes. These are harder to solve or reduce by a teaching method, even though the DM is continuously adapted according to new insights from L2 education research. Among others, exploring how technology can be useful plays a big role. By extending the DM technology-wise, more opportunities to reduce language anxiety (and therefore to increase WTC) and increase Dutch language exposure can arise.

Conversational agent technology can offer a solution to the lack of chances for inter-

[2] More details about the DM can be found in Section 2.2.2.

actions in Dutch and the negative effects of language anxiety. CAs can be defined as "spoken dialogue systems that simulate natural conversation with human beings" (Xiao et al., 2024). CAs transcribe people's speech to text with the use of artificial intelligence (AI), determine the underlying meaning/intent of utterances by using natural language understanding (NLU) techniques, and reply by generating proper responses through AI or pre-programmed selection to keep a conversation going. Existing in many different kinds of shapes, sizes, or forms, they can always be available for conversation, will not create the social pressure humans do, are pre-programmed not to laugh at or make fun of their interlocutor, and are used in a "non-competitive" environment. They could, for example, encourage students to practice conversations at home, before heading to their DM class, where they have to speak with people. The technology is then also used to speak to other "native speakers" besides the teacher. As one may notice, CAs can directly influence the social causes of language anxiety. It is hard to influence cognitive causes, especially personality traits, but long-term use might change a user's fears or self-esteem.

However, a common problem for automatic conversations with agents for NT2 learners regards automatic speech recognition (ASR), due to *accentedness*. Accentedness in L2 learning refers to how learners' speech in the target language differs from the local variety of the target language, and how this variation affects both the speaker and the listener (Derwing & Munro, 2009). In the context of the present study, this term refers to how the Dutch speech of learners differs from native Dutch in the Netherlands. Derwing and Munro contributed to the concept of accentedness with the concepts of *comprehensibility* and *intelligibility*. In short, accent is about difference between individuals' speech, comprehensibility is about the listener's perceived effort to understand the speaker, and intelligibility is about how much the listener actually understands. One can argue that for preventing conversations from being disrupted, the latter is the most important. These so-called conversation breakdowns happen because of non-understanding or misunderstandings. If an accent is not in the way of understanding someone in an L2, there is no significant problem. The conversation can continue, and language acquisition can be stimulated.

Most ASR systems are trained on native speech data, which results in ASR systems possibly under-performing on non-native speech. Palanica et al. (2019) compared three CAs, Alexa, Google Assistant, and Siri, on their speech recognition. A variety of people took part in the study, among whom were participants with a foreign accent in English. Based on the pronunciation of medication names, the comprehension performances were evaluated. Palanica et al.'s study showed that Alexa and Siri performed significantly worse on the foreign accent, and that for Google Assistant, there was still room for improvement. McCrocklin and Edalatishams (2020) assessed the intelligibility (i.e., transcription accuracy) of Google's Voice Typing tool. They compared the accuracy of the dictations of English speech samples from L1 English, L1 Mandarin

Chinese, and L1 Spanish speakers. Google's recognition was highest for native English speakers.

Next to comparing native versus non-native speech recognition by ASR systems, it is also important to determine how ASR performance compares to that of L1 listeners. After all, conversations with a CA would be more realistic and, hence, more useful for pedagogical use if the understanding of the non-native accent is comparable to that of native speakers (Derwing & Munro, 2000). The authors of this paper did just that by comparing Google's recognition of non-native speech to that of human listeners. Human listeners followed a similar pattern to Google, where native English speech was best recognised. In a more recent study, Inceoglu et al. (2023) compared the intelligibility of L2 English speech of Google's ASR to native speakers as well. Overall, the ASR performed better on transcribing than L1 listeners, and it also appeared that if a word was correctly identified by many L1 listeners, the ASR was more likely to recognise that word as well. This is a rather positive outcome for ASR, but some findings highlight that ASR approximating L1 listeners' performance may depend on individual speakers and listeners. The researchers give one relevant reason for this: linguistic context is important for speech perception. When a target word is surrounded by other words, it is easier for people to guess what the target word is supposed to be. ASR seems to have more trouble with this. ASR and L1 listeners show similar scores on transcription overall, but they transcribe in different ways.

This ASR issue of having trouble with non-native speech also influences the way L2 speakers experience and interact with CAs. Wu et al. (2020) found that L1 and L2 speakers have different user experiences regarding intelligent personal assistants (IPAs). L2 speakers appeared to feel that the ASR of the assistants sometimes did not recognise what they were trying to say, causing the conversations to break down. Therefore, the participants suggested to the researchers that they would prefer the assistant to understand simpler L2 commands. Especially for NT2 beginners, it might be hard for the ASR to recognise what words they want to say, as they have not mastered the pronunciation of Dutch words yet. Humans can overcome misunderstandings and thus conversation breakdowns because of pronunciation issues by using strategies, whether or not non-verbally (Moussalli & Cardoso, 2019). A DM teacher could apply strategies such as repeating utterances or changing intonation, or they could understand mistakes because of context or experience and therefore keep the conversation going despite mistakes on the student's side. For a CA, this is not as easy. A CA giving simpler commands would then make sense, but the purpose of L2 learning is that the learners gradually improve their pronunciation and language use over time. This requires practice with minimal interrupted conversations. Therefore, it is important to find out how to solve or overcome this issue, to keep conversations flowing correctly, while the chances of ASR errors are high.

Research has been conducted on how to improve non-native speech recognition; however, much like the previously discussed works on ASR, the vast majority of research on technology-enhanced language learning and teaching has focused on English as an L2 (Shadiev & Yang, 2020; Van Doremalen et al., 2016), while Dutch (and other languages) also deserve attention. Søgaard (2022) argues that the dominance of English in the field of natural language processing (NLP) is problematic as it reinforces global inequalities. People fluent in English are granted better access to NLP applications. The author even suggests trying a ban on English NLP for one year, to provide chances for other languages. Such a measure would lead to language models being trained on a broader range of languages, such as Dutch, encouraging the development of more technological applications for these languages and thereby reducing the bias towards English in NLP. A greater variety of accents would become recognisable for ASR systems, which could then also be useful to improve models for non-native English speech. This, in turn, would contribute to enhancing ASR systems on non-native Dutch speech, as more speech data from various types of NT2 learners/speakers could become available and used for model training. While a one-year ban on English NLP would be an extreme intervention, the current work also aims to contribute to linguistic diversity in the NLP field by focusing on Dutch rather than English.

Next to figuring out how to combat ASR problems for non-native Dutch speech, it is necessary to identify the wants and needs of NT2 learners, such that effective CAs could be designed and developed within the DM context. These factors can be incorporated in design guidelines for any researcher or developer who wants to build a CA for this purpose.

Following all previous points, the current work aims to answer the next research questions (more detailed justifications for these are presented in Chapter 3):

- **RQ1**. To what extent can a conversational agent increase the willingness to communicate of learners of Dutch as a second language?
- **RQ2**. Which strategies can be applied to overcome possible conversation breakdowns in a dialogue with a conversational agent in second language learning?
- **RQ3**. Which (conversational) design guidelines can be proposed for a conversational agent in the Delft method context?

Chapter 2 provides an in-depth review of the related literature, which forms the foundation for the general problem statement presented in Chapter 3. The overall approach to address the problem is outlined in Chapter 4. The first step in this approach - conducting interviews with experts - is detailed in Chapter 5. Insights from these interviews served as input for the development of the first prototype, described in Chapter 6. This prototype was then evaluated through co-design sessions, as documented in Chapter 7. A second prototype, introduced in Chapter 8, was tested by target users,

with the results presented in Chapter 9 and discussed in Chapter 10. Finally, Chapter 11 concludes the thesis.

2

Related Work

In this section, relevant works on CAs and L2 learning/teaching are highlighted and discussed.

2.1 Computer-Assisted Language Learning

2.1.1 What is Computer-Assisted Language Learning?

Ever since technology advanced, numerous foreign language (FL) teaching tools have been developed. Hulstijn (2018) described the rise of audio formats and telephony as the only real revolution in language education at the time of writing. The gramophone, tape recorder, and cassette recorder successively contributed to teaching languages by enabling individualisation of listening skill training– skills that Hulstijn considers fundamental to all aspects of language acquisition. More recently, audio formats started to be stored on computers, giving rise to a new field in L2 learning: Computer-Assisted Language Learning (CALL).

Next to digital audio formats, this field includes online dictionaries, grammar checkers, and communication in the target language through digital text platforms such as email, blogs, and wikis (Chapelle, 2010). All developed and/or used to assist in language teaching and learning. More recent examples are online platforms such as *Duolingo*. This is a website and a mobile app that people can use to learn the basic expressions and vocabulary of certain foreign languages. These kinds of CALL systems should be seen as additional tools for language learning rather than the main method to learn how to produce a target language in authentic situations (Teske, 2017). They serve as an extra tool next to or within second language classes.

Golonka et al. (2014) dived into the world of CALL to identify the most effective technologies for FL learning and teaching. Based on over 350 papers published in the period 1993-2009, evidence for the efficacy of various CALL technologies was evaluated. The technologies were divided into five categories: *schoolhouse- or classroom-based technologies*, *individual study tools*, *network-based social computing*, and *mobile & portable devices*. For two technologies, strong support was found: The social computing technology *chatting* and the individual study tool *ASR*. The former technology allows users to think about their messages before allowing their interlocutor to read or hear them. This makes it relatively easy to practice complex structures. However, chatting was only moderately beneficial for speaking skills in the reviewed studies. This makes sense, as written chat does not stimulate speaking and sending audio messages is not as spontaneous as direct spoken conversation.

ASR technology is more relevant to the current work, as the examples given include comparing a student's pronunciation with a target pronunciation and providing feedback, providing an opportunity to work on speaking ability *individually* (self-regulated learning), and allowing the learner to practice a simulated dialogue with a computerised agent. At the time of this review paper, ASR already had an impact on FL learning and teaching, showing the potential of this technology.

More recent work by Shadiev and Yang (2020) also highlights how ASR can assist in improving speaking skills. ASR technology can provide feedback on pronunciation, provide oral practice of word order, or increase learners' performance on comprehension tests. Nevertheless, issues regarding a learner's accent, pronunciation, speed, or volume, causing low accuracy rates of ASR, have been found throughout the research papers in the CALL field; an important factor to consider for the design and development of conversational agents to hold spoken conversations with.

A relevant ASR-based NT2 learning tool was developed by Van Doremalen et al. (2016). Their DISCO system provides feedback to NT2 learners on their speaking performance. Students partake in conversation simulations with computerised characters in which they can choose from words displayed on the screen to use in their response. DISCO aims to improve pronunciation, morphology, and syntax. Experts (teachers) evaluated the system as useful since, during their lessons, they do not have as much time to focus on pronunciation. Additionally, even though they spend more time on morphology and syntax during class, they thought the spoken interaction with DISCO had an added value compared to the written exercises they do during lessons. They also did not know of any real opportunities to practice Dutch pronunciation at home and thought that students would be stimulated to speak more in Dutch and therefore become more comfortable speaking it. DISCO is somewhat similar to a CA but differs in the sense that a user selects texts to read aloud, rather than the CA interpreting the intent of a user's utterance. It offers limited conversational freedom, unlike the CAs described in

the next section.

2.1.2 Conversational Agents in CALL

Divekar et al. (2021) note that the Human-Computer Interaction (HCI) community has not yet given significant attention to language-teaching CAs. However, as mentioned before, these agents can solve the problem of having a lack of practice opportunities for L2 learners and tackle language anxiety problems. Such conversational agents are far from fully replacing L2 teachers or methods and come with unique challenges because of two differences with regular conversational agents: users are not completely proficient in the used language, and, opposite to completing tasks in the most efficient way possible, there is no direct task the CA should solve. The primary aim is language acquisition through the interaction between the user and the agent.

It turns out that of the CA works out there, many of them are on (commercial) IPAs such as Google Assistant, Apple's Siri, or Amazon's Alexa. Xiao et al. (2024) reviewed works on CAs in English as a second language (ESL) learning, for which the majority were on commercial IPAs, and suggested that there are three CA themes in this context: *General communication practice*, *task-based language learning*, and *structured pre-programmed dialogue*.

In *general communication practice*, agents serve as conversation companions by using ASR to understand and respond to the spontaneous speech of ESL learners. For example, researchers investigated how an Alexa-equipped Echo robot could support L2 learning in the classroom (Moussalli & Cardoso, 2016). In class, ESL learners could ask Echo questions from a pre-set list and also ask questions made up by themselves. The findings suggest that a CA such as Echo could offer learners valuable input exposure, output practice, pronunciation feedback and authentic English conversations in a stress-free environment. Furthermore, it was motivating students to learn on their own. It promoted self-regulated learning. However, the ASR was reported to sometimes not understand the requests of the participants. For the theme of general communication practice, Xiao et al. suggest that longitudinal research on students' attitudes towards L2 learning with CAs is needed, to find out whether the enthusiasm of learners remains over time, i.e. whether there is a novelty bias or not. Furthermore, more research must be conducted on in-home autonomous L2 learning with CAs.

Task-based language learning CAs can be used by learners to complete specific tasks. Tai and Chen (2024) used Google Assistant as a means to find out whether CAs are useful for improving ESL learners' English speaking proficiency through task-based learning in class, and a similar study was conducted on oral proficiency outside class (Tai, 2024). The results were promising, with positive contributions such as expanding exposure to English, adding variety and enjoyment, eliminating learners' fear of making mistakes,

and encouraging self-directed learning outside the classroom. In one of Tai's works, the focus was on the impact of the IPA on EFL learners' WTC, particularly interesting for the current work (Tai & Chen, 2023). The results showed that Google Assistant promoted EFL Learners' WTC by reducing speaking anxiety. Learners were displaying higher levels of engagement, motivation, and confidence due to the less threatening environment of IPA interaction. The tasks in all these works were designed based on what the Google Assistant can do, such as playing games, performing music commands, and looking up information on the internet. Most other works on task-based CAs only used off-the-shelf functions of commercial CAs, too, instead of adapting or redesigning the dialogue systems for specific tasks. Also, linguistic improvement has not been evaluated systematically.

Structured pre-programmed dialogue CAs are customised agents, designed and programmed by researchers or educators such that they can hold conversations about specific topics relevant to the targeted learning outcomes. Lee and Jeon (2024) developed a disembodied voice-controlled CA for primary school EFL learners to find out how they perceived it as a language partner. After interactional tasks, the students' perceptions were documented through a drawing task and in-depth interviews. The majority of the children perceived the CA as human-like or something in between an artefact and a human, showing that such a CA is suitable as an interactive language partner for young EFL learners. This theme featured the fewest papers, and those papers presented focused on young instead of adult L2 learners, so there is much room for growth. There was also no paper relating such CAs to WTC, while the current work's objective is to figure out how a structured pre-programmed dialogue can influence WTC of adult L2 learners.

Another paper published after the review of Xiao et al. was about a unique agent-based system (Lee & Lim, 2023): learners give an agent English sentences they have learned as input. This agent analyses these sentences through the Google speech-to-text API and evaluates the pronunciation and intonation of the learners by comparing them to speech data of native speakers. From this, the learner's level is determined. A separate AI teacher starts teaching sentences based on the level of the learner. The learner then feeds sentences to a different agent. This teachable agent builds a knowledge database for the learner, which can be used to generate dialogues suiting the learner's level. This generation of conversations is done by a machine learning model trained on dialogues of various levels and the sentences it learned from the learner. The agent was developed through the game engine Unity, such that it could be used as a mobile app on Android devices and allows for independent language learning.

2.1.3 The Representation of CAs

CAs can differ in their representation. Some are purely digital and perceptible only through their voice (e.g., IPAs such as Google Assistant), others have a virtual representation (e.g., Lee and Lim's teachable agent (Lee & Lim, 2023)), and some possess a physical presence (e.g., NAO (Gouaillier et al., 2009)). The latter category of language-teaching CAs falls within a CALL sub-field known as Robot-Assisted Language Learning (RALL) (Huang & Moore, 2023). A key feature of RALL is the direct physical representation of the agent, which offers a notable benefit: learners tend to achieve better task performance and/or learning gains with physical agents compared to virtual agents or telepresent robots (i.e., live or recorded video feed of a physical robot) (Leyzberg et al., 2012) (Li, 2015). However, virtual agents are flexible and widely accessible as they can dynamically change their appearances, be easily deployed, and be displayed on relatively inexpensive screens. Therefore, it is important to explore whether physical presence is a required design principle or if a virtual embodiment suffices.

Next to embodiment, research has also explored multimodal feedback in CALL. For example, Wu et al. (2020) compared L1 and L2 English speakers' use of Google Assistant via a smartphone versus a smart speaker. L1 speakers preferred the smart speaker, whereas L2 speakers favoured the smartphone. The latter originated from the visual feedback provided on the smartphone screen, which allowed learners to view transcriptions of their utterances and diagnose conversation breakdowns (see Section 2.4.2 for more on this). Similarly, Tai and Chen (2024) found that multimodal feedback worked better than audio-only feedback. Learners preferred interacting with the CA while receiving narrative, on-screen text, and graphical feedback alongside synthesised speech. The use of multiple modalities is hence something to consider when designing a CA for NT2 learning.

2.2 Teaching Methods for Dutch as a Second Language

Since this study focuses on the Dutch language, unlike the predominantly ESL-focused research in CALL/RALL, it is essential to examine the NT2 teaching landscape and determine where CAs fit within it. Therefore, this section covers relevant NT2 teaching methods.

2.2.1 Popular methods

There are several popular methods in practice nowadays to teach Dutch as a second language. Van Boxtel et al. (2021) explain that the *natural method* assumes that L2 is not learned consciously, but is acquired through receptive skills, implicit learning, and sufficiently understandable input. The focus there is thus on meaning and not on form. By learning the meanings of words and sentences, eventually, the form is learned nat-

urally. Another method is the *communicative method*. Learning materials consist of communicative routines and speech acts. The students learn to communicate by doing and imitating existing constructs. A last frequently used method is the *dynamic usage-based* (DUB) method. Within this method, L2 learners learn chunks of language (consisting of form-meaning pairs) in all layers of the target language. Layers are, for example, intonation, morphemes, or collocations (i.e., words that co-occur regularly). By repetitively getting exposed to and using these chunks, learners acquire the language. They need to encounter the target language often for this. Koster (2015) developed and tested a DUB programme for German learners of NT2, based on a popular Dutch film. The students had to first watch a fragment of the film without subtitles and instructions, and later had the task to watch again while paying attention to what the characters were saying. After, the students get to read the lines from the fragments and either read them out loud or hear the teacher reading them out loud. As a last step, an activity based on the fragment is done. Such an activity can be role-playing the scene, for example. From the results, it was shown that the students valued the method positively in terms of learning and motivation, and that their language proficiency rose significantly. They used a WTC questionnaire as a measurement to quantify the motivation of their learners.

2.2.2 Delft Method

Combining all the aforementioned methods forms the basis of the *Delft method* (DM), which Van Boxtel et al. (2021) et al. describe as "a method with a DUB approach in which communication is central and which also contains characteristics from the natural method". Initially, the method was aimed at students[1] who wanted to start a Dutch-taught programme at the TU Delft. The current work focuses on highly educated individuals who are using the Delft method to learn Dutch. The DM provides situations in class that are motivating for both the students and the teachers, as they offer opportunities for engaging, real conversations (Blom, 2006). Especially when the conversations go beyond the strict framework of the text, students will choose positions in a debate or new viewpoints and insights on a topic are shared with the teacher and the rest of the class. It encourages students to speak in the target language more, fostering greater willingness to use it–an essential factor in L2 learning, as discussed in section 2.3.

Before a DM lesson, students are supposed to read a provided text as homework. They ought to study the text profoundly and look up the meanings and/or translations in their L1 of all words. After this, they should listen to the text, first while reading along and later without reading, until they can directly understand and reproduce everything. With the online application, students prepare with listening exercises. In the

[1] Nowadays, the target group of the Delft method is broader than these students: PhD students, employees, refugees, or students who want to learn Dutch for other reasons than their studies.

textbook, the text is also available but in cloze test form, meaning students can fill in the blanks to check if they know the text properly. Other preparations include online fill-in exercises and answering questions from the book with others or alone. The students are instructed to follow these steps:

The Seven Steps of the Delft Method

1. Listen to the text and read along. Repeat each phrase aloud in the pause following the phrase. Look up words and grammar.
2. Same as step 1, without pausing and looking up words/grammar.
3. Listen, understand, and repeat each phrase aloud in the pause after, without seeing the text.
4. Same as step 3, without pausing.
5. Make the exercises.
6. Do the test.
7. Practice with the questions from the text.

Furthermore, students are instructed to repeat each step four to six times, not advance until they are finished with a step, and study a text for two to three hours in total. Then, after preparing, in class, the teacher will talk with the class about the studied text. The challenge for the teacher is to only use the words that were present in the studied text or previously studied texts. Such a text is presented in Table 2.1.

Speaker	Dutch utterance(s)	English translation
Teacher	Hallo, ik ben Inge. Mijn naam is Inge van Dijk. Ik ben de docent.	Hello, I am Inge. My name is Inge van Dijk. I am the teacher.
Teacher	Wie ben jij? Hoe heet je? Wat is je naam?	Who are you? What is your name? What is your name?
Student	Ik heet Paula. Mijn naam is Paula.	My name is Paula. My name is Paula.
Teacher	Hallo Paula.	Hello Paula.
Student	Dag mevrouw.	Hello, Miss.
Teacher	Uit welk land kom je?	Which country are you from?
Student	Wat zegt u?	What did you say?
Teacher	Uit welk land kom je? Waar kom je vandaan?	Which country are you from? Where are you from?
Student	Ik kom uit Frankrijk.	I am from France.
Narrator	De docent heet Inge. Van Dijk is haar achternaam. Zij komt uit Nederland. Meneer Wang komt niet uit Nederland. Hij komt uit China. Hij woont nu in Den Haag. Paula komt uit Frankrijk. Zij woont nu in Rotterdam.	The teacher's name is Inge. Van Dijk is her last name. She is from the Netherlands. Mister Wang is not from the Netherlands. He is from China. He lives in The Hague now. Paula is from France. She lives in Rotterdam now.

Table 2.1: *The text for DM Les 1 - Hoe heet je? (Lesson 1 - What Is Your Name?)*

These are snippets of the first lesson (NT2 Boom, 2019). For each sentence, the reader can click on an icon next to it in the online environment to hear the corresponding audio file. As one can read, the word "leraar" (i.e. teacher) is not present. This holds

for the whole text of lesson 1 and means that the teacher can say ”ik ben de docent” but not ”ik ben de leraar”.

Speaker	Dutch utterance(s)	English (free) translation
Teacher	Hallo, ik heet Christa, hoe heet jij?	Hello, my name is Christa, what is your name?
Student 1	Ehh ... I ... ik naam Mehdi	Ehh ... I ... my name Mehdi
Teacher	Ehm ... Ik heet Christa, hij heet Ivan. Nu jij: ..	Ehm ... My name is Christa, his name is Ivan. Now you: ..
Student 1	Eh ... ja ... Ik heet Mehdi.	Eh ... yes ... My name is Mehdi
Teacher	Oké! En jij? Heet jij Paula? [kijkt nu plotseling cursist 2 aan]	Okay! And you? Is your name Paula? [now suddenly looks at student 2]
Student 2	Oh?? Me? Paula? No, ehm ... no, ik Carmen!	Oh?? Me? Paula? No, ehm ... no, me Carmen!
Teacher	Ik heet Christa, hij heet Mehdi. En jij?	My name is Christa, his name is Mehdi. And you?
Student 2	Ja, o ja, ik heet Carmen.	Yes, oh yes, my name is Carmen.
Teacher	Hallo Carmen! Waar woon je nu?	Hello Carmen! Where do you live now?
Student 2	Espanje ... no ... Delft!	Espain ... no ... Delft!
Teacher	En jij, woon jij in Den Haag? [kijkt nu cursist 3 aan]	And you, do you live in The Hague? [looks at student 3 now]
Student 3	What ... me? Yes ...no Nee, ik in Delft!	What ... me? Yes ...no No, me in Delft!
Teacher	Oké! En hoe heet je? Wat is je naam?	Okay! And how should I call you? What is your name?
Student 3	Mij name ben Sarah.	Me name are Sarah.
Teacher	Mijn naam IS ...	My name IS ...
Student 3	Mij naam is Sarah.	My name is Sarah.

Table 2.2: *An example of how a conversation in a Delft method class can go. The translations to English are freely translated to convey the mistakes students make. Words between curly brackets are words they said in another language than Dutch.*

In Table 2.2, you can read how a conversation within class can go. This example is from the DM guidelines for teachers (Wesdijk et al., 2021b). In there, they advise not to interrupt every mistake, but only for the disruptive ones. This will keep the communication going and will counteract speaking anxiety. Correcting mistakes can be done in an indirect way to keep the conversation flow going (see Table 2.3). Forgoing harsh error correction in front of the student's peers is important to reduce language anxiety. Throughout the course, students are frequently tested and provided feedback on their learning progress, such that motivation is stimulated and maintained. Frequent tests and self-tests allow learners to get confident with the testing methods of the course, minimising anxiety for final tests. A common problem during the conversation classes is the imbalance between talkative individuals and the quiet ones, just as in daily life. However, in a conversation lesson, it is important that every student can talk more or less the same amount. It is the teacher's task to moderate the talkative students and encourage the quieter ones to participate. In some cases, it is difficult to accomplish this. If language anxiety is the cause of this, the teacher is instructed to praise the student frequently and to be extra careful with correcting them.

Speaker	Dutch utterance(s)	English translation
Teacher	Hoe heet mevrouw van Dijk?	What is Mrs Van Dijk's name?
Student 1	Zij naam Inge.	She name Inge.
Teacher	De naam is Inge: ze HEET Inge.	The name is Inge: she is called Inge.
Student 1	Ze heet Inge.	Her name is Inge.
Teacher	Oké! En komt Inge uit Frankrijk? [kijkt nu plotseling naar cursist 2]	Okay! And is Inge from France? [suddenly looks at student 2]
Student 2	Huh? Moi? Non ... de ... eh ... oet Kameroen.	Huh? Moi? Non ... de ... eh ... frum Cameroon.
Teacher	Ja, jij komt uit Kameroen. Maar komt INGE uit FRANKRIJK?	Yes, you are from Cameroon. But Inge from France?
Student 2	Ah, non! Nee, Inge nee out France, Inge out Nederland.	Ah, non! No, Inge no frum France, Inge frum the Netherlands.
Teacher	Inge KOMT uit Nederland.	Inge is from the Netherlands.
Student 2	Inge kom out Nederland.	Inge is frum the Netherlands.
Student 3	[neemt ongevraagd het woord] Inge docent! Dus zij uit Nederland!	[takes the floor unsolicited] Inge teacher! So she from the Netherlands
Teacher	Ja, Inge IS de docent, zij KOMT uit Nederland.	Yes, Inge is the teacher, she is from the Netherlands.
Student 3	Ach ja sorry! Inge KOMT uit Nederland. Ja, want moet de taal spreken goed! Anders	Oh well sorry! Inge is from the Netherlands. Yes, because have to speak the language good! Otherwise ...
Teacher	Je hebt gelijk, maar niemand begrijpt je!	You are right, but nobody understands you!
Student 3	Sorry hoor!	I'm sorry!
Teacher	Kun jij dat aan hém [wijst naar cursist 4] vragen? Komt hij uit Frankrijk?	Can you ask that to him? [points at student 4] Is he from France?
Student 3	Hallo goedemorgen! ... jij uit Frankrijk?	Hello good morning! ... you from France?
Teacher	KOM JIJ uit Frankrijk?	Are you from France?
Student 3	O ja ja ... Jij uit Frankrijk, eh ... Kom jij uit Frankrijk?	O yes yes ... You from France, eh ... Are you from France?

Table 2.3: *A conversation example in which the teacher addresses mistakes. The sentences of the students are repeated with a correction included. The corrections are capitalised in the Dutch utterances.*

The texts in the Delft method are designed to include the most frequent words in the Dutch language. As students progress through the method, they gradually learn more words. In total, the words from the frequency lists count up to around 1000 words. By the end of the method, beginner students who have mastered these words will understand approximately 85% of spoken Dutch(Van Boxtel et al., 2021).

For the implementation of a CA within the DM context, this implies that the CA should be able to have conversations with the students about the text they have studied in preparation for the upcoming class. It is essential to limit its vocabulary to the words known by the student at that point. It can be debated whether a CA should be correcting a student in the same manner as a teacher, or more frequently and strictly. This decision depends on the role and the purpose of the CA. It could be a substitute for

the teacher, a bridge between homework and class, or an additional tool. In addition to the in-class version of the DM, there is also a self-study programme. With the online environment, students can learn at their own pace and place. A CA could be useful in this setting, too, since the technology supports self-regulated learning.

2.3 Willingness to Communicate

As introduced before, willingness to communicate (WTC) is an important factor in L2 acquisition and might even be viewed as the ultimate goal of L2 learning. WTC was first introduced as an instrument by McCroskey and Baer (1985). It refers to the extent to which an individual is willing to engage in communication with another person in a given situation. The authors measured WTC across four different communication contexts: public speaking, talking in meetings, talking in small groups, and talking in pairs. The receiving parties could have been either strangers, acquaintances, or friends. WTC was conceptualised as a personality-based, trait-like construct, meaning that an individual's WTC is influenced by their character. Another key finding was that, in general, WTC tended to be lower when more receivers were present and when there was a greater social distance between the individual and the receivers.

MacIntyre et al. (1998) argue that WTC is not necessarily trait-like, and that various factors can influence it. They therefore conceptualise it as a situational variable, with both temporary and enduring influences. One of the most significant factors that can alter someone's WTC is the language of discourse itself. According to their model, WTC is in turn an important factor in L2 use: when a person's WTC for an L2 is high, they are more likely to use that language. This is important since, as discussed before, more contact with the target language improves a learner's L2 proficiency. Furthermore, as noted in the introduction (Chapter 1), language anxiety is an antecedent of WTC (MacIntyre, 2017). Therefore, a CA could contribute to reducing language anxiety and, by extension, improving WTC, as several studies have aimed to demonstrate.

Ayedoun et al. (2015) developed a CA aimed at improving users' WTC. The prototype featured a virtual on-screen CA role-playing a human waiter in a restaurant. Using a dialogue manager and a multimodal response generator, users were able to engage in conversations with the CA. The authors designed a dialogue script and corresponding intents for restaurant visitors, such as "requesting a table for X people" or "ordering a drink". With the *Wit.ai* speech recogniser, they were able to classify users' utterances based on intent.

The authors identified the following key requirements for the CA to effectively improve learners' WTC: *social conversational contexts*, *conversation smoothness*, and *learners' immersion*. A CA's domain knowledge has to be context-dependent so that it does not give unrelated responses. However, this implies that for a different context, a different

knowledge model is needed. Therefore, a system should allow for changing the system to various conversation contexts. Conversations should be smooth to allow for a natural and dynamic conversation. The CA should apply strategies to prevent or recover from conversation breakdowns (more about conversation breakdowns can be read in Section 2.4) so that a friendly conversational environment is created and a learner's anxiety can be reduced. Finally, the CA environment should be realistic enough so that the learner can experience good immersion. They wanted their CA to be realistic in the sense that it should have an embodiment and be able to engage verbally through speech synthesis and non-verbally through facial expressions and lip-syncs. Through a survey, the authors found that the CA prototype had the potential to improve English learners' WTC.

In a subsequent study, Ayedoun et al. (2019) enhanced the CA by implementing two types of conversational strategies: communication strategies (CS) and affective backchannels (AB). CSs are actions aimed at resolving communication problems. For instance, approximation occurs when a speaker describes an object instead of directly naming it, typically when they cannot recall the word. In their study, ABs were defined as verbal or non-verbal expressions that occur in a conversation's secondary channel, encouraging the speaker. Given the importance of conversation smoothness, the authors designed their CA to take initiative in turn-taking by using these CSs and ABs. This approach allowed the CA to recover from errors caused by failing speech recognition. The specific function of this is further elaborated in Section 2.4.2.

In a third study, Ayedoun et al. (2020) explored how to make interactions with the CA more personalised, taking into account the WTC levels of participants. The method they employed is called *scaffolding* and *fading*. Scaffolding refers to a process in which an expert supports a learner in performing a specific task or achieving a particular objective, often by providing hints during problem-solving exercises. Fading refers to the gradual reduction over time until the learner can complete the task independently. To the best of our knowledge, this was the last update on their CA.

In their studies, Ayedoun et al. used a questionnaire to measure WTC, presenting participants with 30 situations in which a Japanese person would need to use English. Participants rated each scenario, allowing the researchers to calculate a WTC score before and after interaction with the CA. The questionnaire assessed three variables: *Confidence*, *nervousness*, and *desire*. The highest rating for *confidence* was "I think I could do that easily," while the lowest rating was "I absolutely don't think I could do that." For *nervousness*, these were "I would definitely be nervous" and "I wouldn't be nervous." Desire measured to what extent people wanted to give the situation a try.

Tai and Chen (2023) approached evaluating learners' WTC after interaction with a CA similarly to Ayedoun et al. Before and after the interaction with Google Assistant, they

asked participants to fill out a questionnaire regarding their WTC. In the questionnaire before, participants were asked to rate their perception of communication activities in class on a 5-point Likert scale (1 being *strongly disagree*, 5 being *strongly agree*). From the results, it appeared that they were unwilling to communicate in English, had high levels of communicative anxiety, and rated their competence low. The opposite was true after having experience with the CA, indicating that it helped in improving the learners' WTC and some of its antecedents (anxiety and self-perceived competence).

2.4 Conversation Breakdowns

As mentioned before, due to the accentedness, non-native L2 speech is often hard to recognise for ASR software. It is safe to say that this holds especially for beginning L2 learners. This leads to interruptions of conversation, i.e., conversation breakdowns. Whereas for every CA conversation, breakdowns can happen for any reason, it is important to minimise this issue for L2 teaching CAs, especially so that conversations between L2 learners and CAs can keep flowing. To find out how to get this done, it is important to understand what conversation breakdowns are, what accentedness entails exactly, how people deal with conversation breakdowns in human-human interactions, and what work has been done on strategically handling (ASR) issues in human-CA interaction. In the next subsections, related work on these topics is highlighted, respectively.

2.4.1 What are Conversation Breakdowns?

Wrong pronunciations by the speaker leading to non-understandings, but also other conversation aspects like misunderstandings or complex topics, can cause conversation breakdowns. Since ASR systems are primarily trained on native speech, the chance of misclassifying speech of non-native speakers is higher. Therefore, it is essential to find ways of solving or overcoming conversation breakdowns in human-CA dialogues. According to Varonis and Gass (1985), to overcome conversation breakdowns in dialogues with others, people tend to use verbal or non-verbal negotiation strategies. Examples of such negotiation/conversational strategies are modifying speech in terms of form and meaning, changing vocabulary, repeating utterances, elaborating on utterances, adjusting the utterance's syntax, etc.

2.4.2 Conversational Strategies

Strategies by Humans

In human-CA interactions, strategies are also used by either humans or the CAs. Extending their research on the use of Echo as a CA in L2 learning, Moussalli and Cardoso (2019) examined the conversational repair strategies employed by users when communication with Echo breaks down, primarily as a result of pronunciation-related

difficulties. The most common solutions to these situations were to repeat the phrase or to rephrase it. After these two strategies came simply abandoning the attempt. The authors explain that the participants could have been so persistent in repeating and rephrasing as they did not feel judged by Echo - the device can never become irritated, bored, or impatient. Furthermore, they concluded that the Alexa system installed on the Echo speaker is capable of understanding accented English speech, even with a 20% word error rate. Its level of comprehension was comparable to that of human judges, suggesting that Alexa's ASR system achieves human-like intelligibility. In another study regarding human-CA interactions, similar but also different strategies from the human side were identified: self-repetition, clarification requests, explicit correction, confirmation checks, and paraphrasing (Jepson, 2005).

Strategies by CAs

For the current work, it is essential to understand how a CA for NT2 learning can apply conversational strategies to keep conversations flowing. This deployment of strategies is executed by the dialogue manager. The dialogue manager is that part of a CA system that controls the architecture and the structure of a conversation, based on the ASR and NLU input it receives. Jurafsky and Martin (2009) argue that the simplest dialogue manager architecture is a finite-state manager, in which for every question a CA asks, the next step corresponds to the answer of the user. The dialogue manager controls the conversation completely, meaning the system has the *initiative*. When the full initiative lies with the CA, the conversation often becomes too restrictive. This is why most systems use a mixed-initiative approach, in which the initiative can shift between the system and the user during the conversation. However, as Jurafsky and Martin (2009) mention, users with high speech recognition error (like L2 learners) might be better served by more system initiative. A CA can show initiative by applying conversational strategies, just as users do. In general, this is done using two different methods: *Confirming* and *Rejecting*. A system can explicitly confirm by asking a direct question, such as "You want pizza, right?". Implicit confirming, on the other hand, means that the system replies with a new question, repeating the understood information within it; user: " I want to travel to Utrecht."; system: "By which means of transport do you want to go to Utrecht?".

Rejecting, the second method, means that the system tells the user directly that it did not understand their utterance, e.g. "I'm sorry, I didn't quite get that.". The system could start with *rapid reprompting*, by just saying "I'm sorry?". If, then a second time, the system does not understand the user, it can apply *progressive prompting*. This means the system guides the user by saying what the formulation of the desired utterance should look like, e.g. "I'm sorry. Could you please tell me if you want to travel by car, bicycle, or train, for example?".

Benner et al. (2021) provided even more (and sometimes overlapping) categories of conversational breakdown recovery strategies by CAs. They bundled together confirmation and rejection into *Confirmation* and furthermore identified the categories *Information*, *Disclosure*, *Social*, *Solve*, and *Ask*. With *informing*, a CA tries to explain the breakdown situation with helpful messages or feedback on the error. *Disclosing* is similar to informing, but the CA here admits it does not have human intelligence and thus shows its weaknesses and competencies. This lowers the users' expectations and might result in higher acceptance. The *social* strategy aims at adding human-like behaviour by the CA to the conversation. An example can be apologising, which in the previously mentioned book was part of the *rejection* strategy. Next to that, a CA could compensate the user emotionally (e.g., the inclusion of social cues) or non-emotionally (e.g., offering incentives to continue the conversation). A last option is for the CA to use pausing and turn-taking as a means of showing mis- or non-understandings. If a CA uses *solving* as a strategy, it provides a solution to the user to overcome a breakdown. Examples of this are the CA taking an "educated guess" of what the user wants, providing a list of options that may solve the issue, or using pre-programmed utterance templates that a user can choose an utterance from that the CA will understand for sure. The last strategy, *asking*, is quite straightforward. The CA can repeat the question, ask for rephrasing by the user, or use progressive (re)prompting as we have seen before. As one may notice, all these strategies can be used in combination as well.

Looking at a specific application of conversational breakdown recovery strategies relevant to the current work, Ayedoun et al. (2019)'s DiMaCA applies the following CSs to overcome speech recognition errors: simplification/approximation, code-switching (using L1 words with L1 pronunciation instead of L2), asking clarification, suggesting answer pattern, asking confirmation, asking repetition, expressing non-understanding, guessing, and repetition. Apart from these CSs, the next ABs were used: congratulatory AB, encouraging AB, sympathetic AB, and reassuring AB. To determine which of the conversational strategies to use, DiMaCa follows a routine to check a conversation's current state. From start to end, the system checks the following states: whether or not the learner is silent, is asking for help, is able to understand but not to answer, and is not able to understand or to answer. If the problem is not on the learner's side, the system continues to check whether the agent itself is not able to understand or answer, consecutively. If any of the aforementioned states occur, a conversational strategy is applied. If an AB category is triggered, one of the options from that category is selected at random. On the other hand, CSs are selected heuristically in a predefined order, such that it becomes progressively easier for the learner to overcome the current conversation breakdown.

3

Problem Statement

Based on the literature review, CALL seems to contribute positively to L2 students' learning process. Works showed that ASR technology is quite useful, even for NT2 learning. With ASR technology, CAs could offer solutions to language learning problems such as the lack of conversation opportunities with native speakers, speaking language anxiety, and difficulties during self-regulated learning. Ayedoun et al. (2015) and Tai and Chen (2023) demonstrated that CAs can improve learners' WTC in English, indicating that an increase in WTC can be hypothesised for a Dutch-teaching CA as well.

As seen in the related work, the appearance, embodiment, and (multi-)modality of a CA influence an L2 learner's experience of a conversation. Even though these features have been addressed throughout the whole process of the current research, this work mainly investigates what the conversational design of the CA should look like to improve the WTC of NT2 learners. Examples of what we mean by conversational design in this master's thesis are dialogue structures, dialogue topics, types of questions, word use, and turn-taking patterns.

As previous works lack research on L2 dialogue agents for adults and on WTC, the current project aims to establish design guidelines for such CAs. This will be done in the context of the Delft method, which means that specific requirements for the design, such as the build-up of vocabulary, already exist. This vocabulary requirement limits the technical possibilities; this project stuck to pre-programmed dialogues rather than real-time generative solutions such as large language models (LLMs). State-of-the-art LLMs such as ChatGPT cannot (yet) adequately limit themselves to certain words when generating text, violating the vocabulary requirement.

To aim for WTC improvement in the DM context, it is crucial to figure out how to handle conversation breakdowns. This applies to all CAs in general and to CAs in L2 learning especially, since users are not proficient in the language of discourse. There seems to be a trade-off, where on the one hand smooth conversations are necessary to keep language exposure optimal (if a conversation gets interrupted too much, it bothers a learner's learning process), while on the other hand, repetition and mistake (self-)correction help in learning an L2 as well. A perfect ASR system would also cause an unrealistic learning environment, as native speakers also make listening mistakes. A CA should not exceed the speech recognition level of L1 speakers, and learners should also learn how to deal with conversation breakdowns in the L2.

The approach for the current work was thus to start the design of a dialogue with a conversational agent and to figure out what requirements are necessary for this. Ideally, the CA should be suitable for students to use at home for self-regulated learning and testing their skills before or after class. It should motivate them to speak Dutch more, both in class and in daily life situations. In essence, it should serve as an extra 'someone' with whom they can have proper conversations. The aim was not to develop such a full-fledged final product already, but to begin with finding the route towards it. Therefore, again, the main focus is on conversational design, while the side aspects, such as robot appearance and presence, are touched upon briefly. The following problem statement summarises the overall idea that is addressed:

How to create dialogues with a conversational agent, suiting the Delft method, that lead to more speaking opportunities for Dutch learners and motivate them to speak Dutch more?

To break down the problem statement in testable terms, three research questions are central to this project. They are listed again here:

- **RQ1:**

  To what extent can a conversational agent increase the willingness to communicate of learners of Dutch as a second language?

- **RQ2:**

  Which strategies can be applied to overcome possible conversation breakdowns in a dialogue with a conversational agent in second language learning?

- **RQ3:**

  Which (conversational) design guidelines can be proposed for a conversational agent in the Delft method context?

Chapter 4 explains which methods are used to answer these research questions.

4

Approach: Interaction Design Lifecycle

The current work consists of multiple parts that together form the general approach. This approach is based on the simple interaction design model from Sharp et al., 2019 (Figure 4.1). This model consists of the following steps, within a cycle: *Discovering Requirements*, *Designing Alternatives*, *Prototyping*, and *Evaluating*. In general, a final product is finished after reiterating through this design cycle several times. Since a final product is not the objective of the current work, an adapted version of the interaction design lifecycle model is used, visible in Figure 4.2. The *Designing alternatives* node is merged with *Prototyping* into *Designing & Prototyping*) since designing and prototyping happened in parallel and back-and-forth. Furthermore, whereas from *Design alternatives* one can go directly back to *Discovering requirements* in Sharp et al.'s model, this is different for the current project. Following the adapted model, a project starts by identifying the needs and wants of potential users and establishing the design requirements. After defining these requirements, it is time to design and prototype. The resulting prototype then needs an evaluation, which serves as a basis to update the design requirements. Since the model consists of an iterative approach, the previous steps are repeated to enhance the prototype. In the current work, the cycle has been gone through twice in an attempt to seek answers to the research questions. RQ1 is approached quantitatively, while RQ2 and RQ3 are mainly qualitatively approached. In the following, all steps taken in this project are introduced.

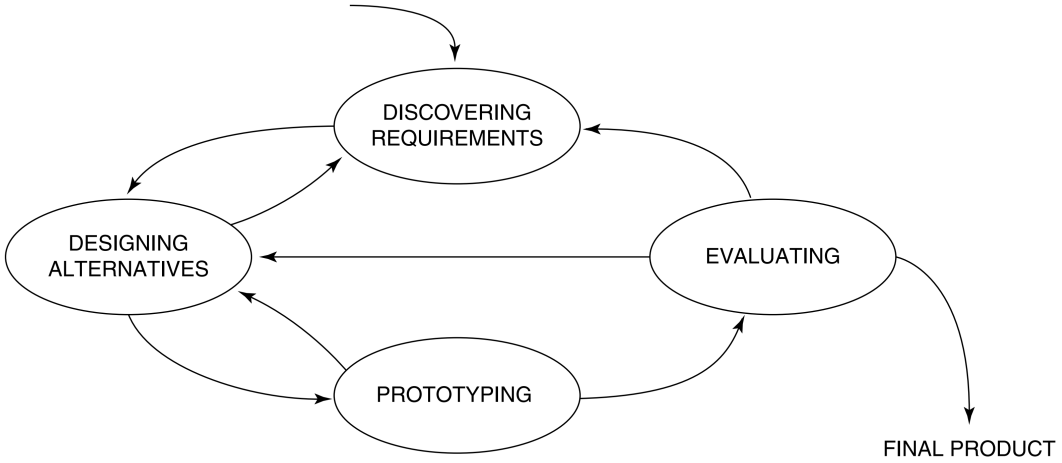


Figure 4.1: *The simple interaction design lifecycle model from Sharp et al. (2019).*

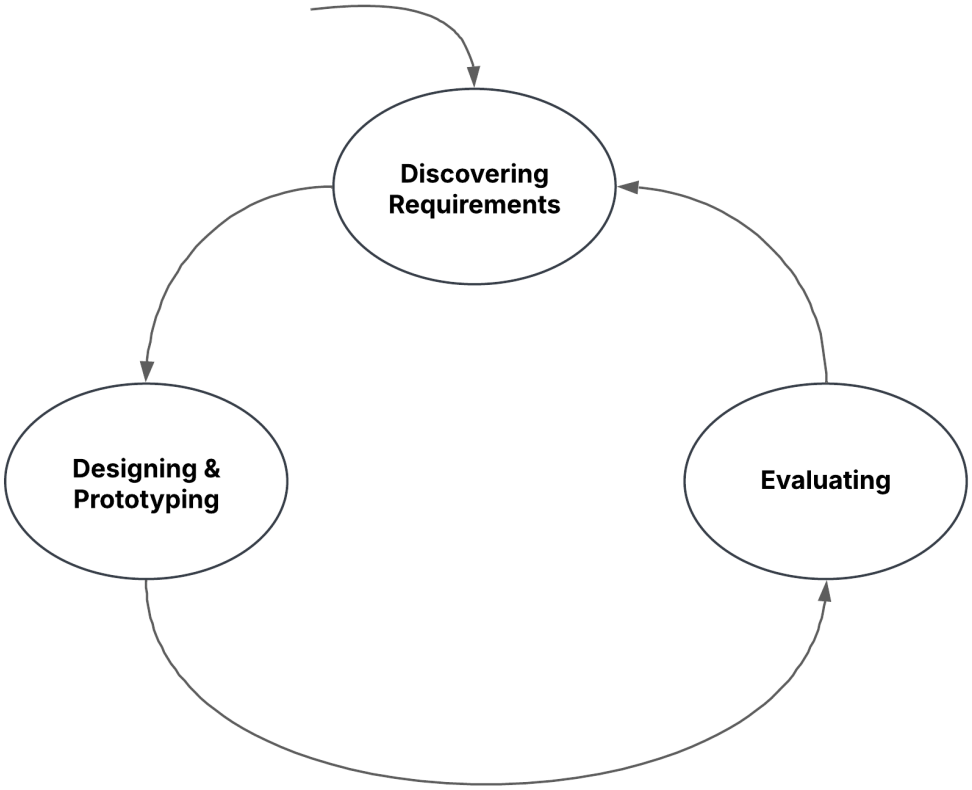


Figure 4.2: *The adapted version of the simple interaction design lifecycle model.*

Discovering Requirements

The first step to identify the needs of the target users and establish the requirements was done through interviews with DM experts. Based on these interviews and the literature review, the first design requirements were established. This was a start for seeking the answer to RQ3. In Chapter 5, the details are explained.

Designing & Prototyping: Prototype I

The design requirements provided a good foundation for the first prototype. Prototype I has been developed through the Furhat social conversational robot platform (Furhat, 2024). The robot is hard-coded, meaning that there has been no use of generative artificial intelligence, such as large language models and that dialogues have been designed by the researcher. More details on the design and development can be read in Chapter 6.

Evaluating: Co-Design

After building this interactive prototype, the evaluation test was performed with the same experts from the first step. The experts were involved in the further design through co-design sessions. They interacted with the prototype, gave feedback, and thought together with the researcher about the next development steps. Here, the search for answers to RQ2 started by testing different strategies. Chapter 7 illustrates how the co-design sessions went and how the design requirements were updated.

Designing & Prototyping: Prototype II

After completing the first step of the design cycle again, it was time to return to the design/building part. The co-design sessions with the experts gave helpful insights for improving the prototype, together with reviewing the related work and revisiting the results of the interviews. Prototype II (see Chapter 8) is an enhanced version of the first prototype and was ready to be tested by target users.

Evaluating: User Tests

The second and last evaluation for this project was performed by conducting user tests with Dutch learners. Participants were asked to interact with the virtual robot and complete questionnaires for feedback on the CA itself and questions regarding WTC (RQ1). In a small interview with each participant, the strategies were also covered (RQ2). Chapter 9 explains the process and the results in detail.

Discovering Requirements: Recommended Design Guidelines

As a last step for this project, the requirements were updated and formulated as design guidelines for CAs in the DM context (RQ3).

5

Discovering Requirements: Interviews

The project started with the first step of the interaction design lifecycle: *Discovering Requirements*. This chapter addresses the interviews that were conducted with two NT2 experts. The method and the results, including the first design requirements, are presented.

5.1 Method

Two separate interviews were conducted with two experts to identify the user needs. These DM experts should know what is important to students and know exactly how the method works, which allows a good foundation for the design requirements.

Participants. The interviewees were NT2 teachers at the Delft University of Technology, hereafter referred to as Teacher I and Teacher II. Teacher I has had around 16 years of experience with the DM, while Teacher II has been teaching DM classes for around 10 years.

Setting. The interviews were conducted online, one-on-one, through video calls.

Procedure. The interviews were semi-structured, which means that there was a set of questions planned to be asked (see Appendix A), but the questions could also deviate from the planned ones to gather more in-depth information. The topics of the questions asked were as follows: the Delft method specifically (e.g., "Can you briefly explain what the Delft method is and how it differs from other methods?" and "Which digital/electronic tools are used within the Delft method?"), conversational agents,

in general and for L2 learning (e.g., "What are your experiences with conversational agents?" and "From which NT2 level could a conversational agent be introduced?"), embodiment ("To what extent do you think/find that [embodied or digital representation][1] would be important in the context of the DM?" and "What do you think of [the Furhat robot]?"), willingness to communicate ("Can you imagine how a CA can contribute to WTC? If so, in which aspects exactly?"), and artificial intelligence ("What are the points of discussion regarding AI within your department?").

Data collection. The responses of the teachers were collected through written notes and video/audio recordings. The audio was transcribed for analysis purposes.

Ethical approval. This method was ethically approved by the CIS Ethics Committee. Both teachers approved the recording of the interviews and agreed that after transcribing, the recordings should be deleted.

Analysis. The interviews were thematically analysed, starting with semantic coding. This means that for each transcribed answer, the essence was highlighted in colour and then summarised in English. Such a small summary is called a semantic code. Examples can be seen in Table 5.1.

Answers	Semantic codes
Een nadeel die cursisten zelf ervaren is, ook juist omdat ze natuurlijk zo gewend zijn, zoals jij en ik ook op school, Engels of Duits of Frans geleerd hebben, dat dat de weg is naar het leren van een taal.	Students are used to learning foreign languages as they did in previous education.
Zeggen natuurlijk onze cursisten heel vaak, ja maar er is zo weinig aandacht voor taalvorm. En ik wil gewoon weten hoe die machine in elkaar zit. Leg het me nou gewoon uit.	Students want to know how the language works linguistically. They want to know theories, reasons, etc.
Misschien is het risico wel meer, doordat we natuurlijk heel veel over inhoudelijke zaken praten, dat pas veel later in dat proces de aandacht voor taalvorm er meer is.	A risk might be that focus on form is way later in the process.

Table 5.1: *Examples of semantic codes for one interview. The answers are to the question: "Zijn er ook nadelen vergeleken met andere methodes?" (Are there disadvantages compared to other methods?).*

After the semantic coding, thematic coding was performed through mapping. Semantic codes were collected and added to corresponding overarching themes such as *Advantages*, *Disadvantages*, and *Experiences with Conversational Agents*. Many semantic codes were connected to others, inside and outside their corresponding themes. Simi-

[1] Squared brackets within quotes/questions in this thesis indicate that something was referred to, but not literally. E.g., an interpretation by the researcher of what an interviewee referred to or a referring question asked by the researcher.

lar codes received a positive link, whereas contradictory codes received a negative link.

5.2 Results

5.2.1 Themes

All themes, their corresponding semantic codes, and the connections between codes are presented in Appendix E. In this section, the questions with important and interesting answers are highlighted per theme, and the first design requirements are presented as well. The teachers also proposed design requirements directly. These have been added to the first design requirements. Below, the most important themes resulting from the thematic analysis are discussed per section.

The Delft Method

The teachers explained the DM's unique aspect of having conversations in class as opposed to grammatical exercises or assignments. The classes are heavily focused on speaking and listening, more than any other NT2 methods. Furthermore, both teachers stressed and confirmed the importance of building up the vocabulary gradually based on word frequency in the Dutch language. They both emphasised the requirement for the CA to use the correct vocabulary, depending on the learner's stage within the DM course of the user.

Advantages and Disadvantages of the Delft Method

The DM comes with both its advantages and disadvantages, which the teachers addressed after being asked ("What are the strongest points of the method? Are there any disadvantages?"). One mentioned advantage was that students learn how to use the learned words creatively to clarify what they want to say. Even though they have a limited vocabulary, teachers notice they figure out how to create sentences and talk about complex topics relatively fast already:

> "When I arrived here [Delft University of Technology], I was very surprised that in a beginner's course with people straight from China, we were already talking, in simple form, but still, about the tax system in the Netherlands and China."

With the Delft method, students learn how to use words in their context by having conversations in every class. This means that they do not necessarily get taught the specific mechanics of the language, but acquire a feeling for speaking Dutch. This is regularly to the dismay of students. The teachers mentioned that students want to know how

the Dutch language works linguistically. They miss explicit explanations of grammar, making them insecure sometimes. Both teachers said that this was a downside of the method.

Since most members of the target group are adults, one of the teachers mentioned that they want to see through something and not just accept things the way they are, like young people tend to do. The feeling of control and understanding is less present when the students are not offered exercises or lists of words to learn by heart. This is how they are used to learn languages in previous education or other language classes. The same loss of control goes for the frequent and specific testing within the DM:

> **"Some do criticise, like 'for such a listening test I have to type exactly what I hear while I want to type what I hear approximately'. So, not all of them are convinced that it is a good indicator of what they can do in the language."**

One teacher mentioned that, from scientific research, it appears that even if you focus a lot on Dutch grammar, it does not mean that students can immediately apply it. Therefore, grammar is not taught specifically in the method. Teacher I also said that even the opposite is true: for learning a language, one does not have to be an expert on that language. From educational science, it is also shown that frequent testing can cause a washback effect: knowing that tests are coming keeps students sharp and on track. They study frequently, rather than studying the majority of the material in the final moments. This appears to work quite well; many students do fine and keep up during the course. This is a plus for the method.

In general, the teachers are positive about the Delft method and are happy to work with it. They notice good results over time and think that leaving out the explicit grammar and typical way of teaching words and phrases is compensated for by how well the Delft method performs. Note, however, that one of the teachers worked on the development of the method and thus might have a partially biased view.

Digital Environment and Testing

To find out how technology is implemented currently within the method, questions about this were asked: "Which digital/electronic tools are used within the Delft method?" and "What are the experiences of the students? Are they positive about it?". The teachers gave extensive explanations and opinions on these tools:

The DM offers a digital environment for the regular course as well as a self-study ver-

sion. The texts from the textbook can be listened to. If a user clicks on a word, the programme shows a translation. There are many different languages for translation available. All sentences from the text can be recorded orally by the student, which the computer can then analyse through speech recognition. Both versions have electronic testing systems, allowing students to do dictation and cloze tests (filling in the blanks). Teacher I deemed these digital tests as "stupid". They missed a form of intelligence, according to her, and were thus not impressive.

To accommodate the self-studying NT2 learners, the teachers made videos in which conversations were demonstrated. An 'interviewer' then faces the camera and asks the student questions about the conversation with pauses, during which the user should answer with the correct words. This, however, does not result in interaction at all, according to one of the experts. Therefore:

> "[A conversational agent] would also be ideal for people who do self-study ... They do not practice that easily with speaking."

Implementation of a Conversational Agent within the Delft Method

In addition to using a CA for self-studying learners, the technology can also fit within the regular course. Both teachers agreed ("At what point within the method do you think such a conversational agent can best apply?") that the best moment within the steps of the DM would be after studying a new text and before coming to class (after or at step 7 of the DM). This way, a student can practice speaking about the text before class and see if they have studied the text well enough. One of the teachers mentioned that it would be nice for students to also use a CA after completing a course, to keep practising what they have learned.

Answering the question of which NT2 level a CA would be useful, both teachers mentioned the beginning level:

> "I think it is very useful for a beginner since we see in the first classes that those people are sitting there with a huge blockage ... Like, you think, I have learned something now, but how on earth does that come out of my mouth? So, it's especially important for beginners to get over that hurdle."

> **"I think, if possible, from the beginning. Maybe especially from the beginning, since people then have few people to practice with. Dutch people are willing to talk to those learning Dutch, but it can be really difficult for them to adjust to that beginner level."**

For advanced students, a CA would also be nice if it could handle more complex topics. This is, however, out of the scope of this project, but still, something to consider for future work.

Experience with Dialogue Systems

The experts were asked about their experience with dialogue systems. During the interview, the examples given (Siri, ChatGPT, Google Assistant, chatbots) are not conversational agents per se, if you look at the definition the current work follows, since these applications do not always perform spoken dialogues. Therefore, this paragraph is about dialogue systems in general. Teacher II did not have much experience with dialogue systems but tried chatting with ChatGPT. They tried to prompt for an example text with only specific words, but the model eventually added non-requested words as well to the sentences – this should be strictly prevented from happening in the DM context. Overall, they were not enthusiastic about OpenAI's application, as it 'hallucinated' and did not show real feelings. Teacher I had some more experience and thought that ChatGPT was a nice big step in the field of linguistics. However, they said that one can recognise the ChatGPT "flavour" after some more chatting experience. Furthermore, the publisher of the DM tried to embed the large language model into a chatbot specifically for DM students. Their experience was not positive; the bot was too pedantic, the conversations were not fun, and it did not stop the conversation when asked to.

Representation and Presence

When asked what Teacher II would think a CA for L2 learning could look like or is capable of, they mentioned that it could be like Apple's Siri:

> **"You can hear a voice from a device, which you can communicate with. You can ask questions. But with a CA to learn a language, I would also imagine the device asking questions to you, which you must answer."**

Both teachers thought that embodiment was not that important ("The previous dialogue systems all have no physical or digital representation representing the agent. To what extent do you think this would be important in the context of the Delft method?"). Teacher I thought that, for her, simply giving the agent a name was already sufficient. They reasoned that this also applies to the DM target group, which consists mainly of theoretically educated individuals. In contrast, a character embodiment might be more appropriate for practically trained learners or those with low literacy skills. Teacher II said that people are already used to talking to devices, and that, therefore, bodily representation was not that important. They were wondering whether a human resemblance lowers the threshold for learners to speak.

However, their opinions shifted to the positive side when they were shown an example video of the Furhat robot as a language teacher. Teacher I found that Furhat came across as natural due to the synchronised mouth and the moving eyebrows. It made her imagine that an embodiment can, in fact, be encouraging. Teacher II also thought it helps when a learner sees a face. They also realised the possibility of presenting a physical representation virtually on a screen, which would be practical for a use case with large student groups.

5.2.2 First Design Requirements

From these interviews and the previous literature, a first selection of design requirements can be formed. Following the MoSCoW (Must Have, Should Have, Could Have, Won't Have) method, evolved from the work by Clegg and Barker (1994), the user wants and/or needs, and system requirements are ranked based on their priority. The priorities are decided based on basic minimal functionality and answering the research questions. In the subsections below, per priority category, the requirements are listed in tables as follows: the requirement number (which is randomly ordered), the users' wants/needs, and the requirement itself. Further explanations per requirement are presented as well.

Must-Haves

ID	Users want/need to ...	So, the CA must have ...
R1	speak in and listen to Dutch.	speech-to-text and text-to-speech functionalities in Dutch.
R2	feel understood to some extent.	the capability to show "understanding" of a student's utterances.
R3	hear the vocabulary learned at the time of interaction.	the possibility for students to indicate their current DM lesson and a database for words taught per lesson.
R4	have a meaningful conversation.	the capability to hold an interesting and logical conversation.
R5	have a conversation going on without too many pauses.	(conversational) strategies in store to overcome conversation breakdowns.

Table 5.2: *First Design Requirements: Must-Haves.*

Table 5.2 presents the design requirements with the highest priority. The following points explain each requirement:

R1. Since the users are learning Dutch, having the CA speak and listen to Dutch is an obvious need. The CA should, at a minimum, have speech-to-text and text-to-speech features to hold conversations in Dutch.

R2. This requirement was proposed by Teacher 1. Even though it makes sense that the users do not speak perfect Dutch, it is better for them to feel understood by the CA, even if only parts of what they try to say are understood. It is therefore important that the CA can show it 'knows' what the conversation is about and what the student tries to say. This can be done through sentences like "I understand" or confirming/repeating sentences from the user.

R3. The user might know and utter words outside of the lesson's scope, but just like in class, they only need to understand the words they learned up until then. It is essential that the CA only uses the words known by the user, just like a teacher does during class. To make sure the CA can be used throughout the whole course, it would make sense to have it follow the lesson structure.

R4. This requirement was proposed by Teacher 1 as well. The conversations should be about the topics from the Delft method. It is important that the conversation makes sense and is not dull. It is crucial that it is interesting and fun for the user to make use of the interaction.

R5. Having too many pauses, i.e. conversation breakdowns, in the conversations is annoying and awkward, thus also not good for practising. If a conversation gets stuck since the CA does not understand the user or vice versa, the CA should apply a strategy such that the conversation can continue. Conversation breakdowns can happen due to, e.g. misunderstandings, wrong pronunciation, or unclear phrases.

ID	Users want/need to ...	So, the CA should have ...
R6	see a human-like face to talk to.	at least a virtual character with human facial features.
R7	have a serious-looking interlocutor.	a form of embodiment adapted to the highly educated target group.
R8	fall back on English when necessary.	speech-to-text for English.
R9	receive implicit and a bit of explicit feedback.	the ability to give a reasonable amount of feedback.

Table 5.3: *First Design Requirements: Should-Haves.*

Should-Haves

In Table 5.3, one can read the requirements that should be met, explained here:

R6. A human face might help a learner to get a better feeling of the language and make it more comfortable to speak. With a human-like representation, a learner can read off facial expressions. A virtual presence might even be preferred over a physical presence, because of the flexibility and prevalence discussed in Chapter 2.

R7. The highly educated people would not want to talk to something like a stuffed animal or a funny-looking cartoon character, according to Teacher I. They want to practice seriously, which means a serious character fits better. Therefore, a serious face and body, with not too many bells and whistles, is required.

R8. Sometimes it is hard for users to come up with a Dutch word, so it helps if they can use the English equivalent to keep the conversation going (code-switching from the literature). Therefore, a speech-to-text feature for English would be convenient. Text-to-speech for English is not necessary, since the CA should reply in Dutch.

R9. Both teachers agreed that just like in class, a user will expect to be corrected indirectly and, if really necessary (but preferably not), directly. If a CA says things like "can you say that again?" or "what do you mean?", it already gives implicit feedback. Next to this, some feedback on word order, word choice, or pronunciation could make the CA a more reliable technology, but this should only happen in extreme cases.

Could-Haves

ID	Users want/need to ...	So, the CA could have ...
R10	have a personalised amount of feedback.	the possibility for students to indicate the desired amount of feedback.
R11	limit their time spending on a conversation.	a way of tracking conversation duration.
R12	feel they directly speak with an entity.	a physical presence.
R13	see, read, or hear feedback besides feedback through the CA's voice.	multi-modal feedback.

Table 5.4: *First Design Requirements: Could-Haves.*

In Table 5.4, one can read the requirements that could be met. These are explained as follows:

R10. Since everyone has their own preferred learning style, it would be convenient if each user could indicate how much feedback they want on their utterances (Teacher I). A slider with different feedback levels could be implemented.

R11. Conversations should end at some point since users do not have all the time to keep going endlessly. Also, at some point, it does not add that much to keep practising the same text. To make sure a conversation will not continue endlessly, Teacher I proposed implementing something like a duration tracker. When the conversation reaches a certain duration, the CA can wrap it up.

R12. Talking to others via a video call, for example, still gives a feeling of distance to people, so a virtual robot will too. Since literature suggests that physical presence adds positive value to the interaction, the agent could also be present physically. This is also dependent on time and place constraints.

R13. From the literature, it is known that multi-modal feedback helps L2 learners. Feedback could consist of displayed text, images, facial expressions, etc.

Won't-Haves

In Table 5.5, one can read the requirement that will not be met.

ID	Users want/need to ...	But the CA won't have ...
R14	use the CA for different Dutch levels.	a possibility for students to indicate their level.

Table 5.5: *First Design Requirements: Won't-Haves.*

R14. Ideally, users would like to use a CA for each DM course. However, even though this is an interesting aspect, it is out of the scope of the current work and will therefore not be considered for now.

These first design requirements and other input from the interviews and literature have been used as a foundation for the prototype. In Chapter 6, the design and technical details are presented in detail.

6

Designing & Prototyping: Prototype I

After discovering the first requirements, the project arrived at the next step: *Designing & Prototyping*. In this chapter, the first prototype is introduced and explained in detail.

6.1 Furhat

The social conversational robot platform Furhat (Furhat, 2024) was chosen for developing, due to its proper documentation and possibility for a virtual version of the robot. The mask (face) chosen is 'Isabel', since the name and face can convincingly come across as Dutch according to the researcher (Figure 6.1). The appearance of the robot was not the focus of the project, but in the evaluation stages, participants were asked about what they thought of this mask (Chapter 7 and Chapter 9). The STT voice used is 'Laura-Neural (nl-NL)', which is created by Amazon Polly.

Kotlin was the programming language in which the prototype was coded, which is cross-platform, high-level, and designed to interoperate completely with Java. Large language models (LLMs) were considered to implement, but due to their tendency to 'hallucinate', it was not fitting to the DM. The robot would use words outside the scope of the target users, which is definitely undesirable for a robot in the DM. Therefore, the choice was made to hard-code dialogues. The bit of artificial intelligence (AI) used in the prototype consists of built-in machine learning (ML) models in Furhat to detect intents and entities.

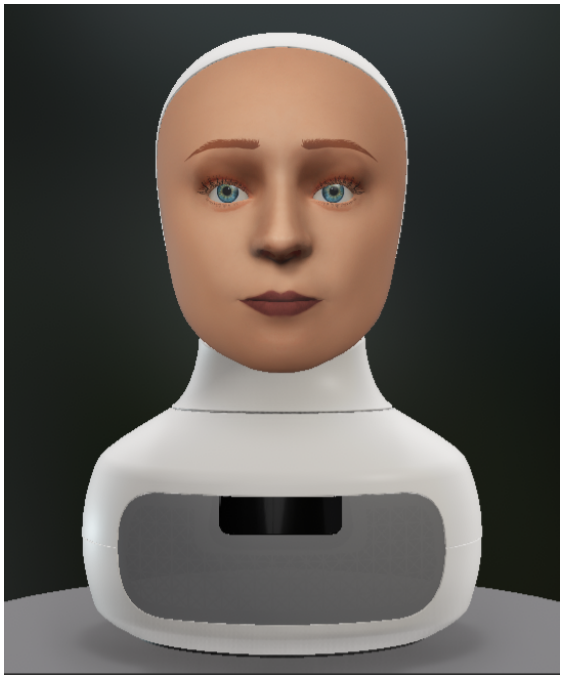


Figure 6.1: *The virtual Furhat used in this project.*

6.2 Dialogue Design

Based on the advice of one of the teachers, the lessons on which the dialogues are based are from the middle of the Green Book: Les 20 – De mooiste reis (The nicest trip), Les 21 – Een dagje uit (A day out), and Les 22 – De Nederlandse bevolking (The Dutch population). Because these texts are in the middle of the book, they allow for longer conversations than at the start, but still offer the challenge of using quite a limited vocabulary. The texts and corresponding questions from the book were studied for inspiration for the dialogue designs. Furthermore, the teacher guide (Wesdijk et al., 2021b) was consulted to create questions, since it contains advice and tricks for teachers to come up with a bunch of questions about the texts. Several conversation patterns and forms are mentioned that can be applied to start conversations or keep them going:

Conversation Patterns and Forms

- **Own experience/opinion**
  - Ask the user about their personal experience, situations, or opinion to keep the conversation relevant and connect class practice with real life. *Example:* "Hoe kwam je hiernaartoe?" ("How did you get here?"), "Wat vond je mooi aan de reis?" ("What did you like about the trip?")
- **Comparing**
  - Compare a situation in the Netherlands with other countries. *Example:* "Zijn het meer of minder mensen dan in Tokyo?" ("Are there more or fewer people than in Tokyo?")

- **Roleplay**
  - Assign character roles from the text to both the user and the robot to act out scenarios.
- **Storytelling**
  - Create a new story using vocabulary from the text.
    Examples: "Lijkt het je leuk om een dagje uit te gaan met mij?" ("Would you like to go out for the day with me?"), "Laten we wat afspreken. We kunnen wat eten en naar de film." ("Let's meet. We can eat something and go to the cinema.")
- **Referring to the text**
  - Ask questions directly based on the text, either as a test or to initiate conversation.
    Example: Questions about the size of Tokyo in lesson 22.
- **Feigned surprise**
  - Express surprise at the user's answer to keep interactions lively.
    Example: "Oh, echt niet?" ("Oh really not?")
- **One-liners**
  - Systematically use short, important expressions from everyday Dutch conversation.

For the first prototype, some of these conversation patterns are applied to the dialogue designs (the ones for which examples are given in the box above), whether intentionally or not. Later, during the development of the second prototype, questions were added specifically using some of these patterns and forms so that the conversations got extended.

Dialogue flows are, of course, crucial in dialogue design. Each of the conversations had its unique dialogue flow, which you can see a (partial) example of in Figure 6.2. The full dialogue designs are presented in Appendix C. As one can see, all dialogues are designed from a robot-initiative point of view. This is done like that because it reduces the possibilities of conversation breakdowns due to the flow control being in the hands of the robot (Jurafsky & Martin, 2009).

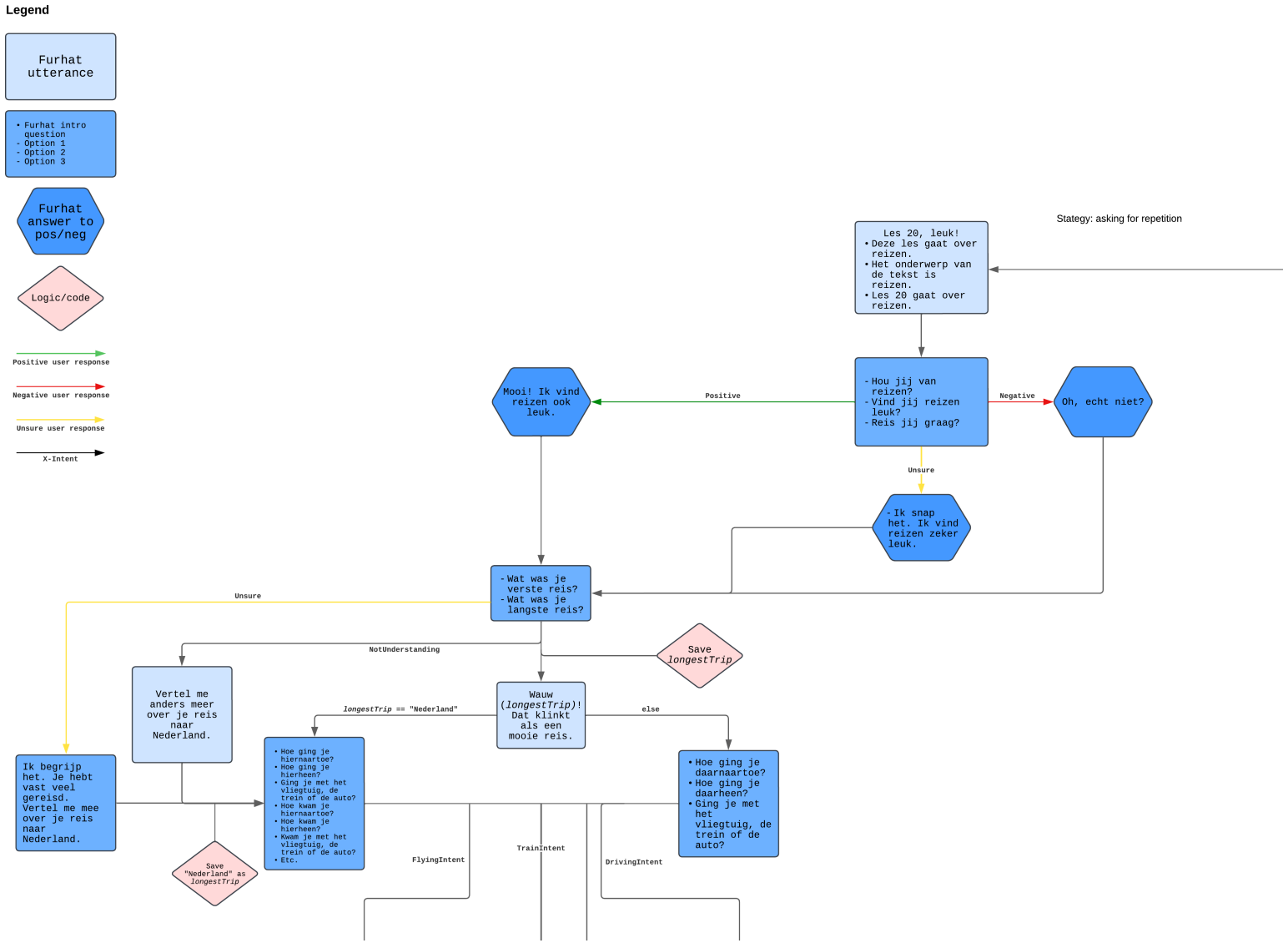


Figure 6.2: *Part of the dialogue design for lesson 20.*

6.3 Strategies

To figure out what strategies could be used to overcome conversation breakdowns (RQ2), the three dialogues each used a different strategy. For lesson 20, the user was asked to repeat themselves if Furhat did not understand them, i.e., *rejecting* (Jurafsky & Martin, 2009). In lesson 21, the user was requested to use more words if the utterance was not understandable for the robot and contained fewer than a certain number of words (typically three, but for some questions, more), or asked to repeat themselves in other cases. This was chosen since it is a way of stimulating the user to use the target language more. Hereafter, this strategy is called *requesting elaboration*. For the final lesson, 22, the robot rephrased its question if it did not understand the user or the other way around, i.e., *rephrasing*. In Section 2.4.2, rephrasing was only addressed as a strategy by humans in human-CA interactions (Jepson, 2005) (Moussalli & Cardoso, 2019), so the current work is a good opportunity to find out if a CA can also use this as a strategy.

For all lessons, the robot used the *repeating* strategy. This means that the CA repeats its question literally when the user indicates they do not understand its utterance or if no response was heard by the CA at all. Initially, it was also attempted to try to use the 'hints' strategy for this lesson, which would mean the robot would start an "aanloopzin". This is a sentence that is not complete and ought to be finished by a student. Since this requires a specific intonation which was not possible to generate with the used text-to-speech model, this strategy was not implemented.

Algorithm 1: LikesTravelling State Logic

`State: LikesTravelling`
`Variable: notUnderstandingCounter = 0`

When this state is entered:

- Prepare a list of questions:
  - "Hou jij van reizen?"
  - "Vind jij reizen leuk?"
  - "Reis jij graag?"
- Choose a random question valid for lesson 20
- Ask the chosen question

If the user responds negatively:

- Say: "Oh, echt niet?"
- Go to the state: `LongestTrip`

If the user responds positively:

- Say: "Mooi! Ik vind reizen ook leuk."
- Go to the state: `LongestTrip`

If the user does not understand:

- Increase `notUnderstandingCounter` by 1
- If `notUnderstandingCounter < 2`:
  - Ask the question again
- Otherwise:
  - Go to the state: `LongestTrip`

If the user is unsure:

- Say: "Ik snap het. Ik vind reizen zeker leuk."
- Go to the state: `LongestTrip`

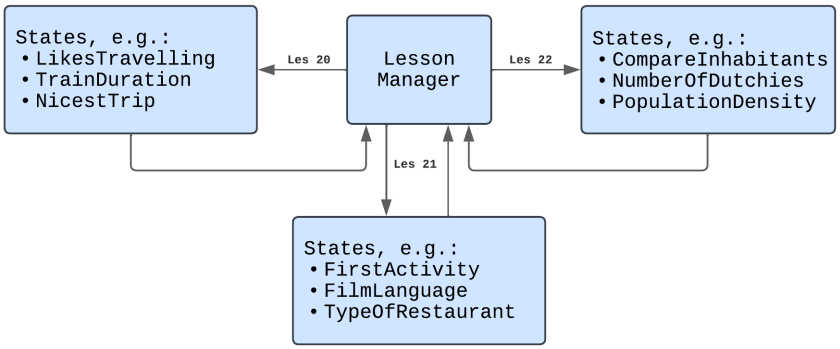


Figure 6.3: *The overarching structure of the prototype. Via the lesson manager, the current lesson is chosen and remembered. For each lesson, there are states, which are entered and walked through depending on the intents of the user's utterances.*

6.4 Technical Implementation

This section describes the technical components of the first prototype. The overall structure is displayed in Figure 6.3. Furthermore, each component of the prototype is described briefly to present a general idea of the application.

Lesson Manager - The lesson manager detects which lesson(s) the user has already done, and depending on that, asks which lesson the user wants to do (again).

States - Each lesson consists of states that can be walked through. For each state, there are intents that can be captured by the robot. Each state has a parent state, depending on the lesson (RepetitionStrat, LengthCheckStrat, or RephraseStrat). In Box "Algorithm 1", the state 'LikesTravelling' is shown. In this state, Furhat asks the user if they like travelling. The user is expected to answer either positively, negatively, that they do not understand the question or that they are unsure about their opinion.

Intents - Intents are labels that can be assigned to a user's utterances, indicating the meaning/intention of the user. There are some standard intents such as 'Positive', 'Negative', 'Unsure', and 'NotUnderstanding'. There are also lesson-specific intents such as 'LongestTripIntent', 'DurationIntent', 'RestaurantIntent' and 'ImmigrationIntent'. In Table 6.1, the example sentences for indicating that the train was used as transport during the user's trip are presented.

User Data - To make the conversation personal, user data can be stored during the conversation. This is done temporarily, meaning that after a conversation, this data will not be saved. The idea is that repeating or coming back to what a user says later in the conversation causes the feeling of understanding from the robot, and that it creates more engagement. Examples of user data are their longest trip, their nicest trip, and the activity they want to do.

Words - To ensure that the robot only uses words fitting the lessons, the DM database is added to the code. This is to make it easy to check for each lesson what vocabulary is available to use by the robot. Furthermore, words that are taught pragmatically to NT2 students are listed as well. These are words such as country names or numbers; words that one will face or need in real life or practical situations, but are not 'really' part of the vocabulary. One feature of the prototype is validating sentences, meaning that it is possible to automatically check whether a sentence is allowed to be said by the robot for a certain lesson. This is particularly useful when creating sentences for dialogue design, since you can input a long list of possible utterances without having to consider whether they are valid. The robot can pick a random sentence and only use it if it is valid. Otherwise, they would go to the next sentence.

Nederlands (Dutch)	Engels (English)
Ik wil stoppen.	I want to stop.
Ik ben klaar.	I am done.
Zullen we stoppen?	Shall we stop?
Eindigen.	End.
Stoppen.	Stop.
Klaar.	Done.
Hou op.	Stop it.
Het gesprek is voorbij.	The conversation is over.
Ik wil niet meer praten.	I don't want to talk anymore.
Het gesprek is over.	The conversation is over.
Laten we het hierbij laten.	Let's leave it at that.
Genoeg nu.	Enough now.
Dat is genoeg.	That's enough.
Stop maar.	Just stop.
Ik ben klaar met dit gesprek.	I'm done with this conversation.
Ik wil afsluiten.	I want to wrap up.
We stoppen hier.	We'll stop here.
Ik wil het niet meer voortzetten.	I don't want to continue this.
Tijd om te stoppen.	Time to stop.
Dit is het einde van het gesprek.	This is the end of the conversation.
Ik ben er klaar mee.	I'm done with it.

Table 6.1: *Dutch and English translations for the intent EndConversation.*

Buttons - Since it is a prototype, so-called wizard buttons were implemented as well to handle technical or conversational issues manually by the researcher (Figure 6.4 and Figure 6.5). Wizarding refers to pretending a functionality of a robot works properly, while in reality, a person is controlling the robot and making it act. There exists a button for stopping the conversation, going to the step of choosing a lesson, requesting a repetition by the robot, and going directly to one of the lessons.

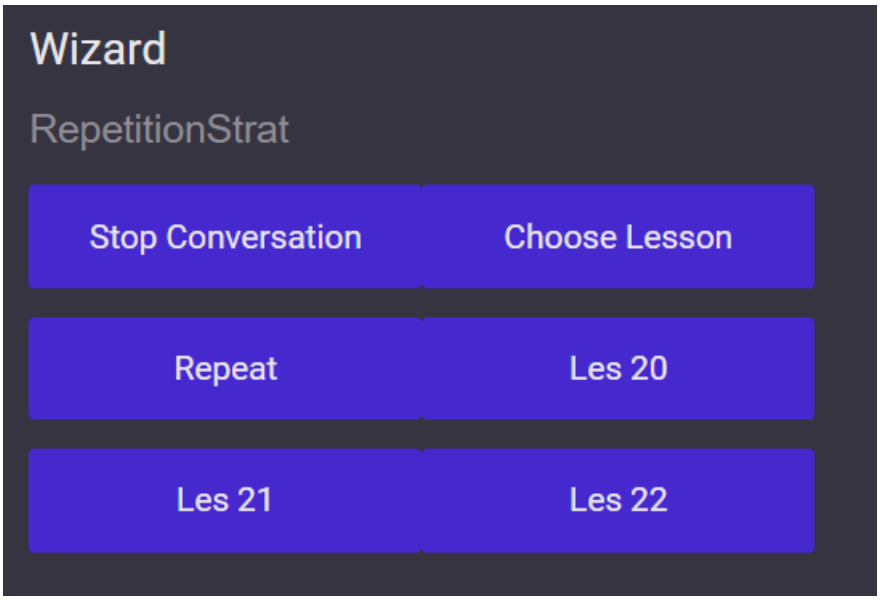


Figure 6.4: *The buttons displayed when Prototype I is started.*

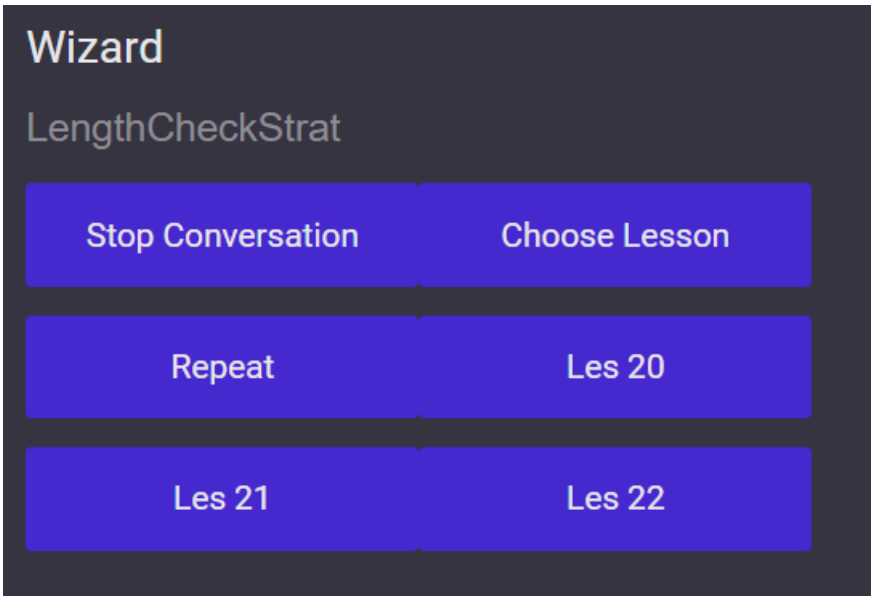


Figure 6.5: *The buttons when Lesson 21 is selected. The 'checking length' strategy is being applied.*

Conversation example - Here, part of the conversation during the co-design session described in the next chapter is demonstrated.

Robot	User
Les 20, leuk. Les 20 gaat over reizen. Vind jij reizen leuk? (Lesson 20, nice. Lesson 20 is about travelling. Do you like travelling?)	Ja. (Yes.)
Mooi. Ik vind reizen ook leuk. Wat was je langste reis? (Great. I also like travelling. What was your longest trip?)	Ooh, mijn langste reis was naar New York. (Ooh, my longest trip was to New York.)
Wauw, New York. Dat klinkt als een mooie reis. Ging je met het vliegtuig, de trein of de auto? (Wow, New York. That sounds like a great trip. Did you go by plane, train, or car?)	Wij gingen met de boot. (We went by boat.)
Dat is interessant. Hoe lang duurde de route? (That's interesting. How long did the journey take?)	Dat duurde drie dagen. (It took three days.)
Ik begreep het niet. Sorry, kan je het herhalen? (I didn't understand. Sorry, can you repeat that?)	De reis duurde drie dagen. (The trip took three days.)
Ik begrijp het nog steeds niet. Kan je het anders verwoorden? (I still don't understand. Can you rephrase it?)	Twee maanden. (Two months.)

Table 6.2: *An example conversation between Prototype I and Teacher II during the co-design session.*

7

Evaluating: Co-Design

Advancing in the interaction design cycle leads us to the next step: *Evaluating*. The evaluation of Prototype I was done through a co-design approach.

7.1 Method

Two co-design sessions were held to find out how to continue the development of the CA dialogue. In this stage, it was too early to test with potential target users. The experts were therefore asked to collaborate on the design and prototype process.

Participants. The same two teachers from the interviews participated, both apart from each other, in a one-on-one co-design session.

Setting. The sessions took place in person, in office rooms at the Delft University of Technology. The researcher ran the robot programme on a laptop, which was connected to a monitor. On the screen, the participant could see Furhat (see Figure 6.1) while the researcher had access to the Furhat studio dashboard (including the buttons from Figure 6.5 and Figure 6.4).

Procedure. Each session consisted of three conversations between the robot and the teacher, one for each lesson. In between, after each lesson, the teacher could indicate their immediate opinions on the conversation they had just had. After, the teachers were explained which different strategies the robot used for each lesson and how other aspects work. To get more in-depth input, semi-structured interviews were conducted as well. Questions regarded their opinions on the conversations (this time they could compare all three lessons since they went through them all), the appearance of the robot, the choice of the specific lessons regarding the DM, the strategies, and multi-

modality. At the end, the researcher went through a survey together with the teachers to receive feedback on several aspects of the robot. First, several options for Furhat's face were shown. The teachers were asked to choose their favourites and explain why, if they wanted to. Next to that, they were asked to rate dialogue-centred conversational design statements like 'the robot understands me well', 'the conversations were interesting', and 'the conversations were long enough' on a scale from 1 to 5 (1 meaning 'completely disagree' and 5 meaning 'completely agree'). Again, they could explain their opinions and ratings. In Appendix D, the survey questions are displayed.

Data collection. The co-design sessions were audio-recorded and transcribed. Furthermore, the survey was an online form conducted through the researcher's laptop.

Ethical approval. The participants gave consent for the recording and transcribing of the sessions. The co-design session approach was approved by the CIS Ethics Committee.

Analysis. Just like for the interviews (Chapter 5), a thematic analysis was performed on the transcripts. Semantic codes were categorised in themes such as *Opinions about the Prototype*, *Teacher Knowledge*, and *Dialogue Design Improvement Points* on a similar map as for the interviews.

7.2 Results

In Section 7.2.1, all themes are discussed. Section 7.2.2 presents the commentary of the teachers for each statement and what the ratings mean for the results. Finally, the updated design requirements are listed.

7.2.1 Themes

The full maps with themes and their corresponding codes are presented in Appendix E. Down below, each theme is discussed.

General Opinions

The teachers had some general opinions about what is important for a CA in the DM context. First of all, it is better to have a good short conversation than a long conversation that does not make any sense. Next to that, it is important to find out how the CA handles grammatically incorrect answers since non-native speakers will make such mistakes frequently and differently than native speakers. The teachers indicated that testing with the target group is therefore essential. When asked about multi-modality and allowing users to read along with what the CA is or was saying, particularly, the teachers differed in opinion. Teacher II proposed as an option to have students read back the conversation after the interaction for reflection. Teacher I disagrees, mention-

ing that the goal is to make the students practice talking purely. Finally, the teachers were asked if switching Furhat's masks for different conversations would add anything to the engagement for the user. They were not opposed to the idea, but it was not that important to them.

Opinions about Prototype I

Teacher I described the mask of Furhat as amicable, though they noted that its intensely blue eyes stood out too much. Other than that, they did not have many comments on the appearance. Teacher II had more to say, even though they said that they generally do not care that much about visual design. Initially, they assumed that the robot was gender-neutral, which they saw as a positive aspect. They said Furhat looked somewhat sterile, which might have influenced that thought. However, later they realised that the feminine appearance and female voice suggested otherwise. At first, they were not a fan of the voice since they thought it was monotone. There was room for improvement in the intonation. However, over time, they became accustomed to the tone and concluded that it was intelligible and neutral. Teacher I agrees with the latter. Additionally, they did not get disturbed by the voice:

> **"For me, there are still so many unnatural things about this conversation, that I am not bothered by [the intonation] yet. I am already aware that there are limitations to the possibilities, and the intonation is not that bad, actually."**

Both teachers appreciated the choice of lessons for this prototype. The lessons offer enough words to play with, but still have their limitations and thus challenge the CA to be careful with its word choice. However, both noted a difference between the first two lessons and the last one. Lesson 22 was more like an examination than an interactive conversation. Teacher I indicated that the questions were too difficult as well, while they liked the initiative taken by the CA in lesson 21. Teacher II commented that they enjoyed the freedom in lessons 20 and 21, where users could share more personal information than in 22. They appreciated that the robot could remember and repeat something the user had said, as well as its ability to handle answers that did not fully align with the question. The number of open questions could be increased, however, to stimulate longer answers from the users.

Other drawbacks addressed by Teacher II were that the robot was too repetitive in clarification requests and that the robot used a word outside the vocabulary. They once again emphasised that this should not happen. A final comment was made by Teacher I regarding the buttons when the researcher suggested that similar buttons

to the wizarding buttons could be handy for users when a conversation is stuck. The teacher commented that in an ideal world, the buttons would not be there, but in this phase of development, it is good to still have them when a conversation cannot continue otherwise.

Teacher Knowledge

Teacher I emphasised that open questions are more effective than closed questions in fostering interactive conversations. They suggested using as many open questions as possible, provided they do not confuse the students. Similarly, Teacher II pointed out that it is difficult for students to give long answers to closed questions. They noted that teachers aim to avoid providing answers themselves, encouraging students to produce as many utterances as possible. They also highlighted the importance of mastering the art of asking questions that evoke the expected responses. During the first interviews, one of the teachers recommended reading through the in-class examples for each lesson to get a sense of how teachers ask questions (Wesdijk et al., 2021a). Furthermore, Teacher II mentioned that students would like to read along with the conversations, but from a didactic point of view, this would be bad. It is better to just listen and answer since they have already practised reading while listening in their preparations. Another point they addressed was regarding the personal conversations they indicated to appreciate earlier: regardless of whether the data is stored, it is, for some people in some situations, not comfortable to answer personal questions:

> "A question like "Why did you come to the Netherlands?" is a [normal question] for many students. But even though it is not stored, it is not that nice for [e.g.] (political) refugees to answer such a question to a computer."

Especially for a robot, this is harder to estimate than for a teacher in class.

Teacher Behaviour

During the conversations, the behaviour of the teachers was observed. What stood out was that Teacher II was confused whether they could already start speaking, that they had to repeat themselves a lot, and that they tended to give short answers to closed questions.

Dialogue Design Improvement Points

Each lesson had its improvement points. Lesson 20 about travelling should have been approached more carefully since the word 'reis' can confuse students. They could won-

der if it refers to transport from A to B, or if it means the whole trip. Therefore, the dialogue design should consider this. Lesson 21 about a "day out" felt like planning a date to teacher II, which might be too personal for practising this topic. When the user does not want to do something, a good follow-up is to have the CA ask for the reason. While this reason does not need to be addressed in detail, it adds an extra step to the dialogue and helps maintain conversational flow. Lesson 22 received the most criticism. It was brief and felt too much like a quiz rather than a conversation, primarily due to the factual nature of the text. However, it remains important to turn such content into a (personalised) conversation to maintain engagement and relevance. Furthermore, the formulation of some questions was too difficult. In general, the art of asking questions to receive answers that the dialogue design expects is important. The CA should therefore also avoid giving answers to questions itself and have the user produce as many utterances as possible.

NLU Improvement Points

From the conversation try-outs, some natural language understanding errors stood out. Points that needed improvement were the CA's understanding of time units, understanding negative answers from the user (e.g., not wanting to go to the theatre), and *vousvoyeren* (using 'u' instead of 'je/jij'). In general, there was a need to add more example sentences to train the built-in ML model better.

Technical/Code Issues

There were also some technical issues or problems with the code that needed solving. The CA did not always count the number of words properly for the length check, the microphone seemed very sensitive since it sometimes picked up what people in the hallway were saying, the CA replied too fast, or there were some very specific issues in the code that caused errors in the dialogue flow.

Strategies

Teacher I was doubting the *requesting elaboration* strategy since they had a hard time determining what a long answer is and what a short answer is. Teacher II thought it was a good strategy, provided that the request itself ("you have to give a longer answer") becomes less patronising. The same goes for the *rejecting* strategy, where "I did not understand you" should change into "What did you say?". The teachers did not have any comments on the *rephrasing* and *repeating* strategies. The teachers also both proposed one strategy each, which could be useful: *asking follow-up questions* and *confirming*. If a conversation gets stuck because the user gives a negative answer, the CA could ask for the reason behind their answer: "Why do you not want to go out with me?". If an answer is unclear, the CA could guess a confirmation, e.g., "Do you want to eat meat?". These strategies were also discussed in the related work in Section 2.4.

7.2.2 Survey

As mentioned in the procedure description, the teachers were asked to give ratings to statements on a 5-point Likert scale from (1: totally disagree to 5: totally agree). In the paragraphs below, for each statement, the comments by the teachers are presented, and it is indicated whether that aspect needs improvement.

The robot hears me well.

Even though both teachers rated this statement a 4, teacher II mentioned that sometimes the CA started talking while they were still speaking. This is an important point, as one will read in Chapter 9 too, since this was a recurring problem in the user tests.

The robot understands me well (improvement needed).

Teacher I and Teacher II assigned a 4 and a 3, respectively, to this statement. The reason for the lower score of Teacher II was that they were doubting a bit whether the CA really understood them. This was mainly due to the conversation for which the CA was supposed to understand time units. The CA did not seem to grasp days and weeks, only hours. Here was room for improvement on the NLU and ML training part.

The robot would also hear/understand international students (improvement needed).

Both teachers rated this as a 3, showing that they did not know to what extent international students would be understood by the robot. Especially since the teachers spoke in a clear ABN (standard Dutch) manner, it was hard for them to determine whether an accented speech would be recognised sufficiently as well. Teacher II hinted at testing it with international users, which, of course, eventually happened as a final evaluation in the project.

The conversations fit well with the accompanying lessons.

The teachers agreed with this statement, assigning both a 4. There was, of course, still improvement possible, but this did not have priority.

The words the robot used corresponded to the correct vocabulary.

In general, the vocabulary used was fine (both teachers rated a 4). However, there were some words that the CA should not have used in both co-design sessions. This is unwanted and was once again emphasised.

The conversations went logically (improvement needed).

The conversations lacked some logic, apparent in the ratings of a two and a three. For teacher II, for example, when they said they did not want to go to the film, the CA did not understand that.

The conversations were interesting (improvement needed).

Teacher I gave a 3 and made some nuances regarding this statement. They said that a conversation 'interestingness' depends on how high you set the bar. To Teacher I, the conversations about lesson 20 and lesson 21 were interesting. Lesson 22 was not, mainly because it felt more like a test to her.

The conversations lasted long enough (improvement needed).

Both teachers disagreed on this statement (both assigned a 2). Teacher I recommended a conversation duration of at least five minutes.

The conversations encourage people to practice speaking.

The teachers thought that the CA definitely stimulated to speak (4 and 5). Teacher I was even speaking too much for the CA since they were interrupted quite a few times, which is again a negative point. Other than that, it means the designed questions and dialogue flow were sufficient.

The conversations encourage practice of listening.

Teacher II also agreed fully with this (5), while Teacher I (4) thought that the structure of lesson 21 was the only one to encourage listening due to the initiative-taking conversational agent.

A student can use the robot to test whether they have mastered the texts (improvement needed).

Teacher II agreed (4), since you could not answer the questions properly if you did not know the content of the text. Teacher I, however, said that at this stage, the quality of the conversations was not good enough to test your knowledge (2). Since Teacher II also said it was a difficult question, there is room for improvement in tailoring the conversations to the lessons.

The robot looked and seemed serious.

The teachers did not have comments on this and rated it quite high (4 and 5). They both indicated from all options the face masks they liked most, but were already positive about the face mask the prototype had at the moment.

The robot suits the target group (improvement needed).

Teacher I had no comments, and rated this a 4. Teacher II did not know whether they should agree or disagree (3) since the DM is not only aimed at university students, while the conversations were suggesting that. Therefore, the dialogue design required some adaptation to a more general audience.

I felt comfortable during the conversation (improvement needed).

Teacher I again had no commentary (4), while Teacher II did not feel comfortable because the CA did not understand them multiple times at the beginning (2). They were wondering how they should say things correctly then.

The robot gave enough feedback.

This statement resulted in quite some discussion for both teachers (both assigned a 3). Teacher II:

> "Implicit feedback is fine. We don't want it to say "the verb is not in the right place" or something. That's not the goal since the goal is more that they're going to talk. I would expect some more reinforcement somewhere, some more positive comments on what you say. I think that could be a bit more present. Also, more enthusiasm in the intonation."

Teacher I:

> "Not that much [feedback], but I didn't think it was necessary either. Sometimes you noticed that a certain answer was already expected. Do I want more feedback? Above all, I want to be encouraged to speak."

7.2.3 Updated Design Requirements

Below, the updated design requirements after the co-design sessions can be read in Table 7.1, Table 7.2, Table 7.3, Table 7.4, and Table 7.5. The tables show the users'

wants/needs, the corresponding requirement, but also the change compared to the previous list of requirements, where the requirements come from, and whether the first prototype met the requirement or not. The latter is based on the ratings/opinions of the expert teachers during the co-design sessions, or is just objectively true or false. Descriptions regarding implementation and/or explanations of new requirements are listed for each requirement (corresponding IDs - N stands for New requirement).

Must-Haves Part 1

R1. *STT and TTS functionalities in Dutch* - This requirement stayed the same and was met before the co-design sessions since the Furhat voice and input settings were set to Dutch. Since evaluating whether this requirement is met is simply true or false, no evaluation was necessary.

R2. *Capability to show "understanding"* - The experts deemed the CA not understanding enough yet, meaning that the example sentences given to the ML model for training were not sufficient, both in quantity and quality.

R3. **Changed**[1] - *Database for words taught per lesson* - During the co-design sessions, once again, the importance of using adequate words for each lesson was discussed. The requirement was split up into this one, 'having a database for the words per lesson', and 'the possibility for students to indicate their level' (a *won't-have*). The latter was far too ambitious for the current project, so it would be a

ID	Users want/need to ...	So, the CA must have ...	Source	Impl.?	Evaluation
R1	speak and listen to Dutch.	STT and TTS functionalities in Dutch.	Common sense	✓	Researcher
R2	feel understood to some extent.	the capability to show "understanding".	Interviews	×	Experts
R3	hear the vocabulary learned at the time of interaction.	a database for words taught per lesson.	Interviews + co-design	✓	Experts
R4	have a meaningful conversation.	the capability to hold an interesting and logical conversation.	Interviews	×	Experts
R5	have a conversation going on without too many pauses.	(conversational) strategies in store to overcome conversation breakdowns.	Literature + co-design	✓ Impr. needed	Experts

Table 7.1: *Updated Design Requirements: Must-Haves part 1. A green background indicates that the requirement is either new, changed, or moved. The "Impl.?" column shows whether the requirement was implemented and if improvement was needed. Details per requirement can be found in the text.*

[1] Old R3: the possibility for students to indicate their current DM lesson and a database for words taught per lesson.

feature more for a final product. The database was implemented before the co-design sessions, along with a validation function that can check if sentences are valid for a specific lesson. Also, a list of countries was implemented, attempting to capture all countries a user could mention.

R4. *Capability to hold interesting and logical conversations* - The experts deemed the designed conversations not meaningful enough yet. They lacked logic and could be more interesting.

R5. (*Conversational*) *strategies to overcome conversation breakdowns* - The three previously mentioned strategies to overcome conversation breakdowns were tested, but needed adjustments for them to work better. Furthermore, new strategies were suggested by the experts.

Must-Haves Part 2

N1. *Conversations of around five minutes* - There was no time goal set for the conversations before the co-design sessions, but they were not nearly long enough yet. Together with the teachers, the goal was set at 5 minutes. If possible, a conversation should not stop before the 5-minute mark, but if a conversation is about nothing or is repetitive, it does not make sense to wait until 5 minutes are reached.

N2. *Challenge students to speak about the topic of the text specifically* - The experts were contradicting each other, where one said that you need to know the texts to answer the questions and the other that at this stage, one cannot test properly if they studied the texts well enough. What is clear is that the dialogues should be designed such that they fit the texts well.

N3. *Wait a while before determining whether the user is done speaking* - The robot was interrupting the teachers quite often during the conversations in the co-design sessions. Especially for Dutch learners, it is important to wait before determining if the user is done speaking. Therefore, a threshold for waiting time must be set.

N4. *Minimum number of words for answers to and use mainly open questions* - Users would be triggered already to speak, but there was room for improvement. More open questions and an important distinction between open and closed questions were necessary.

N5. *Sufficiently long questions and answers* - The robot would challenge students already enough for listening due to the length of the robot's utterances. To keep their attention and practice listening to Dutch, the robot using only short questions and utterances would not help.

N6. *Controlling buttons* - Buttons could help to control when a conversation gets stuck, and even the CA cannot repair it. The buttons should be a backup modality, and thus not needed in an ideal scenario.

N7. *Set of questions with the same meaning/intentions at each conversation state* -With varying options of utterances, a CA will repeat itself less. At each moment, a random utterance can be chosen, which keeps the conversation diverse in general.

ID	Users want/need to ...	So, the CA must have ...	Source	Impl.?	Evaluation
N1	have a conversation that takes long enough yet is not too long.	conversations of around 5 minutes (as long as they stay meaningful)	Interviews + co-design sessions	×	Experts
N2	be able to test if they studied the texts well.	ability to challenge the student to speak about the topic of the text specifically	Interviews	✓ Impr. needed	Experts
N3	have enough time to answer.	to wait a while before determining if the user is done speaking.	Co-design sessions	×	Experts
N4	be encouraged to speak.	to request a min. number of words for answers to open questions, to use mainly open questions & to avoid giving answers to its own questions.	Literature + interviews + co-design sessions	✓ Impr. needed	Experts
N5	be encouraged to listen.	utterances and questions that are of sufficient length.	Literature + interviews + co-design sessions	✓	Experts
N6	control the conversation if it somehow stops working	buttons to go to certain points in the conversation or repeat Furhat's utterances, etc.	Co-design sessions	✓ Impr. needed	Researcher
N7	hear diverse utterances from a CA.	a list of questions with the same meaning/intentions at each conversation point.	Co-design sessions	✓	Researcher

Table 7.2: *Updated Design Requirements: Must-Haves part 2. A green background indicates that the requirement is either new, changed, or moved. The "Impl.?" column shows whether the requirement was implemented and if improvement was needed. Details per requirement can be found in the text.*

Should-Haves

ID	Users want/need to ...	So, the CA should have ...	Source	Impl.?	Evaluation
R6	see a human-like face to talk to.	at least a virtual presence with human facial features.	Literature + interviews	Yes	Researcher
R7	have a serious-looking interlocutor.	a form of embodiment adapted to the educated target group.	Interviews	-	-
R9	receive implicit and reinforced feedback.	the ability to give a reasonable amount of implicit and reinforced feedback.	Interviews + co-design sessions	✓ Impr. needed	Experts
N8	personal yet private conversations.	built-in dialogues that are not too sensitive and a temporary user data memory.	Co-design sessions	✓ Impr. needed	Experts

Table 7.3: *Second Design Requirements: Should-Haves. An orange background indicates that the requirement is either new, changed, or moved. The "Impl.?" column shows whether the requirement was implemented and if improvement was needed. Details per requirement can be found in the text.*

R6. *Virtual presence with human facial features* - The researcher picked a human-like face mask for the Furhat CA.

R7. *Embodiment adapted to target group* - The experts found it hard to say whether the appearance of the Furhat fit the target group. Therefore, this was something to evaluate during the user tests.

R9. **Changed**[2] - *Reasonable amount of implicit and reinforced feedback* - If a CA says things like "can you say that again?" or "what do you mean?", it already gives implicit feedback. Implicit feedback was implemented, but reinforced feedback, such as compliments and encouragements, was not as present as desired.

N8. *Not too sensitive dialogues & temporary data memory* - During the conversation, the data can be saved. But after that, it should be removed. This feature was already active for some states, but could be used in more states to achieve more personal conversations.

[2] Old R9: the ability to give a reasonable amount of feedback.

Could-Haves

ID	Users want/need to ...	So, the CA could have ...	Source	Impl.?	Evaluation
R8	fall back on English when necessary.	speech-to-text for English	Literature	×	-
R12	feel they directly speak with an entity.	a physical presence.	Literature	×	-
N9	have variation in terms of speaking partners.	different appearances and voices per lesson.	Co-design sessions	×	-

Table 7.4: *Second Design Requirements: Could-Haves. A purple background indicates that the requirement is either new, changed, or moved. The "Impl.?" column shows whether the requirement was implemented and if improvement was needed. Details per requirement can be found in the text.*

R8. **Moved from Should-Haves** - *STT for English* - Speech-to-text for English was not implemented and could therefore not be evaluated.

R12. *Physical presence* - There was no physical Furhat used, so this requirement could also not be evaluated.

N9. *Different appearances and voices* - Different characters for different conversations would make the application more engaging and fitting. This requirement was proposed to the experts, but not implemented yet. Therefore, an evaluation was not possible.

Won't-Haves

ID	Users want/need to ...	So, the CA won't have ...	Source
R10	have a personalised amount of feedback.	the possibility for students to indicate the desired amount of feedback.	Interviews
R11	limit their time spending on a conversation.	a way of tracking conversation duration.	Interviews
R14	use the CA for different Dutch levels.	a possibility for students to indicate their level.	Interviews
N10	practice all lessons from the courses.	possibility for students to indicate their current DM lesson.	Interviews

Table 7.5: *Second Design Requirements: Won't-Haves. A red background indicates that the requirement is either new, changed, or moved. The "Impl.?" column shows whether the requirement was implemented and if improvement was needed. Details per requirement can be found in the text.*

R10. **Moved from Could-Haves** - *Indicate desired amount of feedback* - A slider with different feedback levels could be implemented, but it is not a priority for this project. The CA already would give implicit feedback, the amount of which cannot be changed.

R11. **Moved from Could-Haves** - *Tracking conversation duration* - To make sure a conversation will not continue endlessly, it would be smart to implement a duration tracker. When the conversation reaches a certain duration limit, the CA can wrap it up. This can be considered for future work.

R14. *Indicate Dutch level* - Even though this is an interesting aspect, it is out of the scope of the current work and will therefore not be considered for now

N10. *Indicate current DM lesson* - The prototype is now based on only three lessons. Implementing all lessons is a lot of work. For future work, considering using LLMs can be helpful for this.

In Chapter 8, it is explained how the design and development of Prototype II aimed at meeting these (new) requirements.

8

Designing & Prototyping: Prototype II

8.1 Dialogue Design

As it appeared during the co-design sessions that lesson 22 brought too many complications, it was decided not to implement it in Prototype II. This came with the advantage that a bigger pool for user test participants was possible: lesson 22 is not part of the Elementary 1 course, so those students would not have been able to participate in user tests if lesson 22 were to be included. Furthermore, the dialogues for lessons 20 and 21 received more attention and were expanded in an attempt to meet the five-minute requirement.

New states were added, based on the previously mentioned conversation patterns and forms:

Conversational Strategies in Prototype II

- Own experience/opinion
  - Examples from Prototype II are ”Wat vind je interessanter op vakantie? De cultuur of de natuur?” (”What do you find more interesting on vacation? Culture or nature?”) and ”Heb jij ook wel eens vertraging?” (Do you also have delays sometimes?).
- Comparing
  - An example from Prototype II is ”We kunnen de rekeningen delen. Ik kan tikkies sturen. Zo doen we dat hier. Hoe gaat dat in jouw land?” (We can split the bills. I can send tikkies. That’s how we do it here. How

is that in your country?)

- Storytelling
  - Prototype II still had the same story about Furhat asking the user to go out, but this time the story was that they were going with a group of friends instead of just the two of them.
- Feigned surprise
  - Furhat had more cases in which it could express surprise, for example, if a user says they are from Mexico after Furhat explaining that its favourite trip was to Mexico.
- Oneliners
  - Furhat systematically ended lessons with "Klaar is Kees!". Not all words in this sentence are part of the vocabulary for all lessons, but the sentence can be viewed as "pragmatic" just like country names.

8.2 Strategies

Instead of the separation of strategies from Prototype I, this prototype used *rejecting* (asking for repetition), *rephrasing*, and *requesting elaboration* (*minimum number of words*) for both lessons. The latter only for (some) open questions, as advised by the teachers. Just like Prototype I, *repeating* was used for both lessons, sometimes also instead of rephrasing, since the questions by Furhat were selected randomly from a list per state. The new recommended strategies *asking follow-up questions* and *confirming* were implemented: e.g., respectively "Waarom vind je reizen niet leuk?" (Why do you not like travelling?) and "Wil je eerst naar de film?" (Do you first want to go to the cinema?).

8.3 Technical Implementation

Lesson 1 Vocabulary

les, de, hoe, heet, heten, je, hallo, ik, ben, mijn, naam, is, docent, wie, jij, wat, dag, mevrouw, uit, welk, land, het, kom, komen, zegt, zeggen, u, waar, vandaan, Frankrijk, en, meneer, uw, komt, China, woont, wonen, in, welke, stad, woon, nu, Den Haag, straat, centrum, op, nummer, telefoonnummer, ook, nee, hij, dichtbij, haar, achternaam, zij, Nederland, niet

Lesson 2 Vocabulary

voornaam, adres, postcode, plaats, telefoon, mobiel, werk, Engeland, heb, hebben, Engelse, Nederlandse, Nederlands, nationaliteit, lang, hier, zes, jaar, leeftijd, wanneer, geboren, maart, alleen, bij, geen, vrienden, vriend, of, familie, getrouwd, trouwen, met, een, Nederlander, zijn, we, jullie, huis, buiten, kinderen, kind, ja, twee, jongen, van, drie, meisje, vijf, maanden, maand, ze, man, zoon, dochter, hun, moeder, vader, heeft

States - In Box Algorithm 2, the state template is presented. This is the foundation of each state, where a developer can add the following: utterances the CA can choose from, next states per response type, open questions, and closed questions. The Kotlin code can be found in Appendix G.

Algorithm 2: State Template

```
State: StateName
Parent: LesXY
Variables:
  • notUnderstandingCounter = 0
  • previousUtterances = empty list

When this state is entered:
  • Prepare a list of options:
      – "Optie 1"
      – "Optie 2"
  • Choose a random utterance
  • If a valid utterance for lesson XY is found:
      – Add it to previousUtterances
      – Ask the utterance
  • If no valid utterance is found:
      – If there are previous utterances, repeat the last one
      – Otherwise, say: "We gaan naar de volgende vraag." and go to
        NextState

If the user wants to end the conversation:
  • Go to the state: Stopping

If the user requests repetition:
  • If there are previous utterances, repeat the last one
  • Otherwise, say: "We gaan naar de volgende vraag." and go to NextState

If the user does not understand:
  • Increase notUnderstandingCounter by 1
  • If less than 3:
      – Re-enter the state
  • Otherwise:
      – Say: "We gaan naar de volgende vraag." and go to NextState

If the user answers a closed question:
  • Say nothing and go to NextState

If the user answers an open question:
  • If the answer is too short:
      – Say: "Kan je je zin wat langer maken?" and re-enter the state
  • Otherwise:
      – Say nothing and go to NextState
```

Intents - New intents were created, and old example sentence lists were extended with the support of the LLM ChatGPT. For each intent, the prompt included a small set of examples for a new intent or the old intent from Prototype I, when it was aimed to extend the sentence list. Table 8.1 shows the updated version of the intent from Table 6.1.

Nederlands (Dutch)	Engels (English)
Kunnen we stoppen?	Can we stop?
Kunt u stoppen?	Can you stop?
Kun je stoppen?	Can you stop?
Zullen we hiermee ophouden?	Shall we stop this?
Het is genoeg geweest.	It has been enough.
Ik heb geen zin meer om te praten.	I don't feel like talking anymore.
Ik heb hier genoeg van.	I've had enough of this.
Dit gesprek hoeft niet verder te gaan.	This conversation doesn't need to continue.
Ik vind het wel goed zo.	I think it's fine like this.
Het is tijd om af te sluiten.	It's time to wrap up.
Ik wil het gesprek beëindigen.	I want to end the conversation.
We kunnen nu stoppen.	We can stop now.
Ik ben klaar met dit onderwerp.	I'm done with this topic.
Ik wil hier niet meer over praten.	I don't want to talk about this anymore.
Zullen we het beëindigen?	Shall we end it?
Mag ik nu stoppen?	May I stop now?
Kunt u dit gesprek afronden?	Can you conclude this conversation?
Laten we hier een punt achter zetten.	Let's put an end to this.
Het is klaar.	It is done.
Voor mij hoeft het niet meer.	I don't need it anymore.
We kunnen hiermee stoppen.	We can stop with this.
Ik vind het gesprek klaar.	I think the conversation is done.
Ik wil niet verder praten.	I don't want to talk further.
U mag stoppen.	You may stop.
Stop alstublieft.	Please stop.
Kun je ophouden?	Can you stop?
Kunt u ophouden?	Can you stop?
Zullen we stoppen met praten?	Shall we stop talking?
Ik stel voor om nu te stoppen.	I suggest we stop now.
Ik zie geen reden om door te gaan.	I see no reason to continue.
Dit lijkt een goed moment om te stoppen.	This seems like a good moment to stop.
Ik beëindig dit gesprek.	I'm ending this conversation.
Het is tijd om te stoppen.	It's time to stop.
Ik zou graag willen afsluiten.	I would like to conclude.
Dit gesprek kan stoppen.	This conversation can stop.
Ik wil niets meer zeggen.	I don't want to say anything more.
Er is niets meer te bespreken.	There's nothing more to discuss.

Table 8.1: *Added sentences for the EndConversation intent from Table 6.1.*

Words - Boxes Lesson 1 and Lesson 2 show how the vocabulary of the DM was implemented as a database. Each word was assigned to the lesson it appeared for the first time. In Appendix G, the code implementation of the word database is presented.

Buttons - In Figure 8.1 and Figure 8.2, the enhanced buttons are shown. One can see that for almost every state, there exists a button in Prototype II. This is done such that the researcher can control the conversation more precisely by switching between states directly. The pragmatic word list, which for Prototype I only contained countries, was updated with languages, food & drinks, numbers, months, etc.

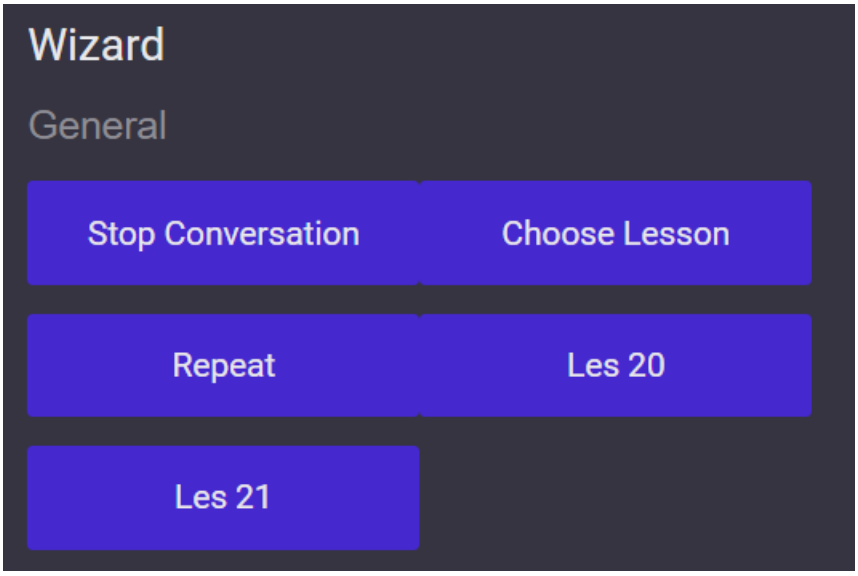


Figure 8.1: *The buttons displayed when Prototype II is started.*

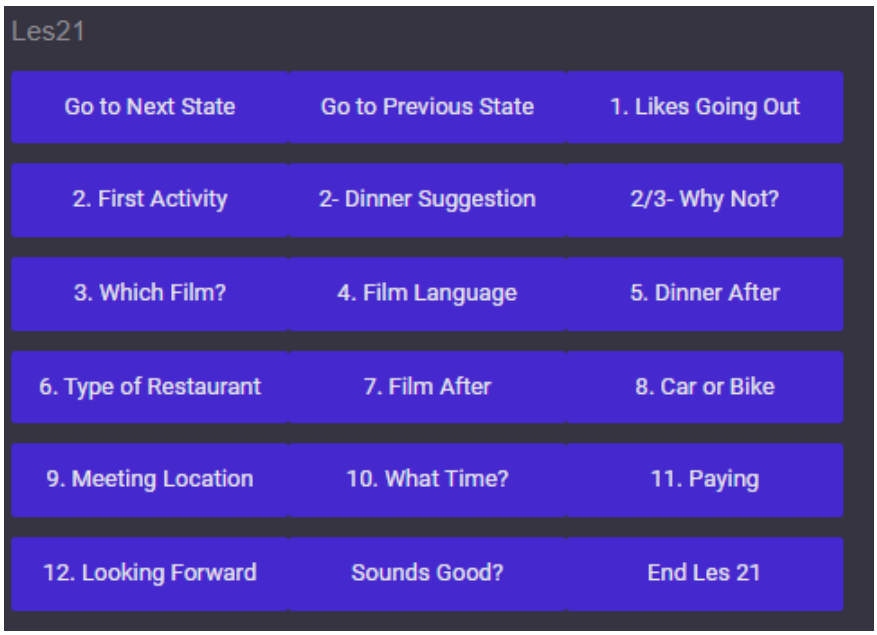


Figure 8.2: *The buttons when lesson 21 is selected.*

Conversation examples for Prototype II are presented in the next chapter about the user tests.

9

Evaluating: User Tests

Next to experts, it is essential to test a prototype with potential target users. This can result in different perspectives on the same aspects and in new insights. Participants were recruited at the TU Delft for the user tests with Prototype II.

9.1 Method

Participants. In total, 11 participants took part in the user tests. Each participant completed lesson 20 and 21 of the Delft method Green Book, in one way or another: they either followed the Beginners Intensive course (BI), slower-paced course Elementary 1 (E1), which goes up until lesson 21, additionally Elementary 2 (E2), or the digital self-study. The most important demographic data per participant are displayed in Table 9.1. Here one can see which course(s) each participant completed, that the majority was woman, the origins were quite distributed geographically, the most common mother tongue was Spanish but other than that each participant had a different L1, the most common L2 being German, and that many participants were unemployed (explainable by them following an intensive Dutch course).

Setting. The user tests took place in a classroom at the TU Delft. Participants were to sit down at a table across from the researcher. The researcher's laptop was given to the participants for two tasks (see measurements and procedure) and connected to a monitor during the interaction with the CA. This monitor was presented to the users, with the virtual Furhat displayed. The speakers and microphone of the laptop were used for playing and receiving audio.

Measurements. To collect the data, multiple measurements were prepared. The conversations with Furhat were recorded on audio (if given consent by the participant)

DM course(s)	Gender	Origin	L1	L2(s)	Occupation
BI	Man	Brunei Darussalam	Malay	-	Unemployed
BI	Woman	Brazil	Portuguese	Spanish	Unemployed
BI	Woman	Colombia	Spanish	German (A1)	Unemployed
BI	Woman	Ukraine	Ukrainian	Russian	Unemployed
BI	Woman	United States	English	Some Spanish, some Japanese	Employed
BI	Woman	Iran	Persian	-	Unemployed
BI	Woman	Turkey	Turkish	-	Unemployed
E1, E2, I1	Man	Mexico	Spanish	Italian, German, Mandarin	Master university student
E1	Man	Norway	Norwegian	Beginner level German, Swedish	Master university student
BI, II	Woman	Mexico	Spanish	-	Bachelor university student
BI, E1	Man	Thailand	Thai	Some Japanese	PhD student

Table 9.1: *Participants overview. The abbreviations for the courses are as follows: BI (Beginners Intensive), E1 (Elementary 1), E2 (Elementary 2), and II (Intermediate Intensive). L2s are second languages other than English and beginner-level Dutch.*

and automatically logged. Before the conversations, the participants were asked to complete a so-called pre-questionnaire. The questions were open, multiple-choice, Likert scale questions, or ratings. The first questions were direct demographic questions, the results of which are visible in Table 9.1. Additionally, questions regarding experience with chatbots and alike were asked: e.g., "Which of the following chatbots/conversational agents/assistants do you have experience with?" and "How would you describe your experience with chatbots?". The Likert scale questions contained WTC-related statements. These questions were to be answered on a Likert scale ranging from 'never' (1) to 'very often' (5) and are presented in Table 9.2. These statements are an adapted version (to the DM) of the statements from Menzel and Carrell (1999). This questionnaire was evaluated with a Cronbach's alpha of 0.92 and was validated to investigate classroom L2 WTC by Yu (2015). After the conversations, the participants were asked to complete a post-questionnaire. Again, they had to rate the same statements, but this time keeping in mind or imagining that they could use the CA to prepare for their DM lessons. In addition to the WTC statements, the participant rated (almost) the same statements from the co-design sessions regarding direct feedback on the CA. This, on a scale from 1 being 'totally disagree' and 5 being 'totally agree'. The answers to the questionnaires were automatically saved through the digital forms.

The pre- and post-questionnaires can be found in Appendix F. The last measurement used was a semi-structured interview to receive more in-depth answers regarding the CA's features and usefulness. The following are the main questions asked:

1. What are the improvement points for this robot?
2. Would you use this robot to prepare your lessons?
3. Would you use this after your lessons?
4. Would you think the robot would also help you to speak more in daily life situations in which you have to use Dutch?
5. Do you think a physical version would be better?
6. The robot used some strategies to solve when you didn't understand the robot, or the robot didn't understand you. Have you noticed any specific strategies it used?

Since it was semi-structured, follow-up questions such as "Which situations would you prepare for?" or "What do you think of this strategy?" were asked. Each semi-structured interview was audio-recorded as well. All recordings were transcribed for analysis purposes.

Procedure. In total, the experiment procedure consisted of three tasks for the participant. First, the participants were asked to complete the pre-questionnaire on the researcher's laptop. After the pre-questionnaire, the participants interacted with the CA. Just before the interaction, the participants had the opportunity to read through lessons 20 and 21 in either their own Green Book or a digital version on the laptop. This, to refresh their memory about the lessons and prepare them for the conversation. Furthermore, the participants were assured that their Dutch level was not being judged. Through speech, the participants could indicate which lesson(s) they wanted to practice. The conversations mainly went autonomously, but there have been cases where the researcher had to intervene through the wizarding buttons. Lastly, the participants were asked to complete the post-questionnaire and were asked questions in the semi-structured interview.

Ethical approval. This user test approach was approved by the CIS Ethics Committee. All participants agreed on the recording and transcribing, where applicable.

9.2 Analysis

Eight of the 11 conversations have been tracked through logging, allowing for an analysis of the errors and the speech-to-text performance. Unfortunately, something went wrong with the three other conversations. The logs were transformed into transcript tables, which contain per CA-user turn pair what the CA said/asked, what the STT interpretation was of the user's response, the real response, whether there was a con-

#	Statement
1	When the class is engaged in an open discussion.
2	When the class size is small.
3	When the teacher asks for a response from the class.
4	When the topic is interesting.
5	When my views differ from my classmates' views.
6	When I am sitting far away from the teacher.
7	When I followed all steps of the Delft method to study the text of the current class.
8	When almost the whole class is actively participating.
9	When the class is engaged in a heated debate.
10	When I am comfortable with the subject matter.
11	When an assignment is being discussed.
12	When no one else is talking.
13	When I am sitting close to the teacher.
14	When my views differ from the teacher's views.
15	When I am angry about a topic.
16	When I know the correct answer.
17	When I can really help clarify the discussion.
18	When I dislike my classmates.

Table 9.2: *WTC statements from Menzel and Carrell (1999) adapted to the DM context.*

versation breakdown, which type of technical error occurred, and specific remarks. These logs were used for answering both Section 3 regarding conversational strategies and Section 3 proposing guidelines, partially.

Conversation breakdowns occur when the conversation (flow) is influenced negatively by errors or when they cause an unpleasant, unexpected, or illogical response from the robot. If, e.g. not all words get captured but the conversation still goes on according to how it should, then a conversation breakdown has been overcome. There was still a technical error, however. A conversation breakdown is always caused by a technical error, but a technical error is not always the cause of a conversation breakdown. The robot takes initiative, so the pronunciation and grammar errors from the user also do not count as conversational errors. Communication strategies aim at overcoming and/or preventing conversation breakdowns. All technical error types are described in Table 9.3. To get a grade for each conversation, rating how well a conversation went in terms of conversational breakdowns happening on a scale from 1 to 10, the following formula was applied:

$$\left(\frac{\text{Breakdowns}}{\text{No breakdowns} + \text{Breakdowns}} \times 9\right) + 1$$

where **Breakdowns** represents the number of times a conversation breakdown occurred, and **No breakdown** represents how often no conversation breakdown occurred for all CA-user turn pairs.

Technical Error Type	Description
Robot interruption	When the robot interrupts while the user is still busy replying or thinking. Usually happened because the robot incorrectly detected that the user finished or completed their answer.
Speech-to-text	The speech-to-text result did not correspond to what the user said, even when the user did not make a pronunciation error, and this led to an illogical conversation.
Dialogue flow	The coded dialogue flow might not be completely robust. This leaded to the robot not being able to understand unexpected answers or handle certain questions from the user.
Robot didn't hear anything	When the robot did not hear any speech, while the user was talking.
Classified wrong intent	When the model classified the wrong intent (so an intent that can be captured). This could happen because too few example sentences were provided or intents being quite similar to one another.
Not all words captured	When not all words by the users are captured, sentences might be incomplete and lead to misunderstandings.
Other	Special errors that did not happen frequently.

Table 9.3: *All technical error types and their descriptions.*

The pre- and post-ratings enabled the performance of statistical analysis. First, the normality of the differences for each statement was assessed using the Shapiro-Wilk test. If the differences were normally distributed, a paired t-test was conducted; otherwise, the Wilcoxon signed-rank test was used. The analyses were performed using Python, with the significance level set at 0.05. For the feedback from the survey, the average score for each statement was computed. The semi-structured interviews were thematically analysed by again using semantic coding.

9.3 Results

9.3.1 Conversations

Table 9.4 and Table 9.5 respectively show parts of the highest-graded and lowest-graded conversation. Table 9.6 is an overview of the occurring conversational errors and the grades for each conversation.

Robot	User (STT)	User (real speech)	Conversation breakdown?	Technical error
Welke les wil je doen? Je kan kiezen uit 20 of 21.	Hallo ik kies best 21	Hallo ik kies **les** 21.	No breakdown	Speech-to-text
Is goed! Les 21 gaat over een dagje uitgaan. Mijn vrienden en ik willen een dagje uit gaan. Lijkt het je leuk om een dagje mee te gaan met ons?	Ja ik kan	Ja ik kan.	No breakdown	-
Dat is goed om te horen! Laten we wat afspreken. We kunnen wat eten en naar de film. Wil je eerst een film kijken of eerst uit eten?	Ik denk eerst uit eten is beter	Ik denk eerst uit eten is beter.	No breakdown	-
Gezellig! Laten we beginnen met ergens avondeten. Ik ga een restaurant zoeken. Wat eet je het liefst: vlees, vis of vegetarisch?	Ik eet lieverd vis	Ik eet liever vis.	No breakdown	Speech-to-text
We kunnen naar een visrestaurant. Ben je het ermee eens?	Ja dat klinkt goed	Ja dat klinkt goed.	No breakdown	-
Super! Dan doen we dat. Ik zal reserveren. Wil je na het eten naar een film kijken?	ja	Ja.	No breakdown	-

Table 9.4: *Part of the highest-graded conversation. "User (STT)" stands for what the speech-to-text model interpreted the user said.*

Robot	User (STT)	User (real speech)	Conversation breakdown?	Technical error	Remarks
Welke les wil je doen? Je kan kiezen uit 20 of 21.	maar ik	Ik keus ehh.	Breakdown	Robot interruption	-
Sorry, ik begrijp je niet. Welke les wil je doen? Je kan kiezen uit 20 of 21.	N/A	Ja ik begrijp het. Ik keuze een ehh de les 20.	-	-	Researcher intervention.
Is goed! Les 20 ging over reizen. Reis je vaak?	N/A	Soms. Soms-	Breakdown	Robot didn't hear anything & Robot interruption	Participant laughing (nervously?)
Ik hoorde je niet. Wat zei je?	soms maar eh ik	Soms, maar eh ik uhh.	Breakdown	Robot interruption	Participant was thinking too long about their answer. Participant comments that the robot went very fast and they needed to get more time for thinking, so lesson 20 got restarted.
Is goed! Les 20 ging over reizen. Vind jij reizen leuk?	Ja ja ik ik heb eh een reis dat we dan ook	Ja ik heb een reis dat was leuk.	No breakdown	Speech-to-text	-
Mooi! Ik vind reizen ook leuk. Wat was je verste reis?	mijn	Mijn verste? Ehh mijn verste reis was.. (interrupted) Australië.	Breakdown	Robot interruption & Speech-to-text	-

Table 9.5: *Part of the lowest-graded conversation. "User (STT)" stands for what the speech-to-text model interpreted the user said.*

Part. #	Robot interruption	STT	Dialogue Flow	Robot didn't hear anything	Wrong Intent	Other	Cv. grade
1	4	16	6	5	7	2	6.1
3	8	19	0	7	1	1	5.0
4	5	17	1	7	5	0	5.6
5	12	15	2	2	3	0	3.9
6	5	4	5	0	6	0	6.4
7	2	17	5	0	8	0	7.2
9	7	19	7	0	4	0	4.0
10	3	10	5	0	4	0	6.4
Total	**46**	**117**	**31**	**21**	**38**	**3**	**-**

Table 9.6: *Here is presented how frequently which technical error occurred and the conversation grade (formula introduced in Section 9.2) for each conversation. The headers of the columns match the rows in Table 9.3.*

The best conversation was with participant 7 (7.2), while the worst conversation was with participant 5 (3.9). The most frequent error was "speech-to-text", while "robot didn't hear anything" happened the least (next to "other"). These error frequencies are important for the guideline ranking in Chapter 10; recovering from more frequent errors has more priority than errors with a lower chance of happening. Therefore, proper speech-to-text functionalities and giving users enough time to speak are among the highest-ranked guidelines.

Participant #	CB prevented	CB	% CB prevented
1	9	20	31.03
3	9	21	30.00
4	14	15	48.28
5	6	19	24.00
6	5	15	25.00
7	16	16	50.00
9	9	26	25.71
10	7	14	33.33
Total	**75**	**146**	**33.94**

Table 9.7: *Overview of conversation breakdowns (CB) that were prevented. The percentages reflect how many breakdowns were avoided out of total breakdown risks.*

In Table 9.7, it is shown how effective the combination of strategies was in preventing conversation breakdowns. In total, 33.94 % of breakdown risks were prevented. There is, hence, still some improvement possible here. Note, however, that breakdown prevention is not equal to overcoming breakdowns. A breakdown can also be overcome by conversational strategies after it has happened, not only by preventing it.

9.3.2 Willingness to Communicate

Statement	Pre	Post	Normality	P-value	SMD
1. Open discussion	4.27	4.36	≤ 0.05	1.000	0.16
2. Small class size	4.55	4.73	≤ 0.05	0.625	0.31
3. Response request teacher	4.36	4.55	≤ 0.05	0.625	0.24
4. Interesting topic	4.64	4.64	> 0.05	1.000	0.00
5. Different view classmates	3.82	4.18	> 0.05	0.167	0.45
6. Far away from teacher	3.55	4.00	≤ 0.05	0.004	0.60
7. Followed all DM steps	4.09	4.64	≤ 0.05	0.031*	0.71
8. Active participation	3.82	4.45	≤ 0.05	0.0623	0.82
9. Heated debate	3.64	4.09	≤ 0.05	0.125	0.58
10. Comfortable subject	4.55	4.36	≤ 0.05	1.000	-0.27
11. Assignment discussion	3.91	4.27	≤ 0.05	0.250	0.45
12. No one talking	4.00	4.36	≤ 0.05	0.219	0.49
13. Close to teacher	3.82	4.45	≤ 0.05	0.046*	0.85
14. Different view teacher	3.73	4.00	> 0.05	0.341	0.32
15. Angry about topic	2.73	3.36	≤ 0.05	0.063	0.71
16. Knowing correct answer	4.00	4.27	≤ 0.05	0.500	0.39
17. Able to help clarify	4.09	4.55	≤ 0.05	0.250	0.55
18. Disliking classmates	2.36	2.91	≤ 0.05	0.063	0.72

Table 9.8: *This table presents the (abbreviated) WTC statements rated by the participants. Pre: average rating of the statement for the pre-questionnaire. Post: average rating of the statement for the post-questionnaire. Normality: whether the ratings for this statement were normally distributed. P-value: the p-value of the statistical test (P-values < 0.05 were considered statistically significant and are marked with an asterisk). SMD: Standardised Mean Difference indicating practical effect size.*

Table 9.8 shows the average ratings, normality distribution, and the significance of each WTC statement. Two statements showed statistically significant increases in post-

questionnaire ratings compared to pre-questionnaire ratings: "When I followed all steps of the Delft method to study the text" ($p = 0.031$, $SMD = 0.046$) and "When I am sitting close to the teacher" ($p = 0.046$, $SMD = 0.85$). These results suggest that the interaction with Furhat had a positive effect on participants' willingness to communicate in these contexts.

9.3.3 Feedback Survey and Semi-Structured Interview

Statement	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	Avg
Robot hears me well	3	2	3	2	2	4	3	2	4	4	4	3.00
Robot understands me well	4	3	3	2	2	4	3	1	3	4	3	2.91
Conversations fit lessons	4	5	5	5	5	5	5	2	5	5	4	4.55
Correct vocabulary	5	4	5	4	5	5	5	5	5	4	5	4.73
Logical conversations	4	4	5	4	4	4	5	4	5	4	5	4.36
Interesting conversations	3	4	4	4	4	4	5	2	5	4	5	4.00
Sufficient length	4	2	2	4	1	4	4	5	4	5	3	3.45
Encourages speaking	5	4	3	5	2	4	4	5	5	5	5	4.27
Encourages listening	5	5	4	4	5	4	5	5	5	5	5	4.73
Tests text knowledge	5	5	4	4	5	4	4	4	5	5	5	4.55
Robot looks serious	3	4	2	3	3	4	2	3	4	4	2	3.09
Fits target group	5	5	4	3	3	5	5	2	5	3	3	3.91
Comfortable conversation	4	3	3	3	2	4	5	4	5	5	5	3.91
Enough feedback	4	1	3	4	4	5	3	2	4	2	3	3.18

Table 9.9: *The results of the ratings of feedback statements regarding the robot (task 3).*

The worst score was for the statement "Robot understands me well". The best scores were for both "Correct vocabulary" and "Encourages listening". In the next section, the themes resulting from the interviews are explained, including explanations for some of these scores.

Themes

General usefulness of CA

In general, all participants were positive about the idea of using a CA for the Delft method. Participant 6, for example, thinks they would benefit a lot from such a CA:

> "For me, it can be really helpful because when I study the lessons, I need to talk about the lessons out loud to somebody. Because of this, I asked my classmates to do such a thing, but they were [too] busy to do this ... It can be helpful for me to be more confident and speak more about the lesson in class."

Participant 7 shares the opinion on confidence:

> "If I'm not confident enough to speak in public, then I think [the CA is] a good option to practice."

However, participant 8 has his doubts about the CA solving the full problem of confidence. He thinks this is more of a task for the teachers.

When to use the CA

Most participants indicated that the CA could be used both before *and* after class, except for participant 1, who thinks it is useful only after, and participant 4, who prefers only before. Participant 7 indicated that they would prefer to use the robot more after the whole course, since then they have fewer chances to practice, whereas participant 5 thinks they would use it more before classes to prepare, since they would only use the robot when they need to review lessons. Participants 8, 9, and 10 would like to see some improvements before they would use the CA after classes. Participant 8, for example, would like to see the agent work like Siri, on a mobile phone. Participants 9 and 10 would like to use it for writing exercises, so a writing feature would be needed.

CA features

It was very clear that users would like to have more time to think of and give an answer to the CA's questions: Five out of the eleven participants indicated this. This was also

clear from the conversational errors in Table 9.6, where one can see that in total the participants were interrupted by the robot 46 times.

Even though the appearance of the CA was not the main focus of the current work, the topic has been discussed during the co-design sessions, as mentioned before, but it also came up during the interviews with the users. Three participants (8, 9, and 11) would have preferred the looks to be more artificial or cartoony rather than so human-like. Both participants 9 and 11 mentioned the characters from the language learning app Duolingo as examples of humanoid but not too human-like representations. Whereas participant 8 came up with a lot of ”better” characters, such as V-Tubers or Pokémon, participant 9 ended up saying that a black box also would suffice. He did not care that much about the appearance. Participant 11 did care about it, even though he is very familiar with using bodiless ChatGPT (even in an L2 learning context). He emphasised he thought human-like robots are creepy to him, but also explained that human-like faces would be better for facial expressions. The conclusion was that it is very tough to decide between a human or non-human-like face. Looking at the ratings, there was neither much agreement nor disagreement with the statement ”The robot looks serious” (average = 3.09). The participants leaned more towards agreement for ”The robot fits the target group”, but there is still room for improvement.

By the average ratings regarding hearing and understanding (3.00 and 2.91), participants indicated that the CA is not hearing and/or understanding them as well as desired. Furthermore, users would like to see an option to slow down the speech (1, 9, 11), possibly in real time (9). Participants 8, 10, and 11 would like to receive explicit feedback, although participant 11 also mentioned that implicit feedback also makes sense:

> **"I think when you speak to the robot, and they don't understand, you know by your heart that your pronunciation is still not correct. So, you will start to have to repeat several times until they understand you. And I think that's also quite similar to what's happening in real life when people don't understand you as well."**

Participant 9 saw differences between the CA and teachers but did not make a strong judgment about it. What was important to him was that he would like to have the robot use his answers more noticeably. His answers should determine the conversation flow. Other ideas for features were: An option to turn on captions (participant 1), voice can

be softer and friendlier (participant 6), conversations can be longer (participant 7), option to see how much of the lesson is left (participant 9), use silence to give hints (participant 8), option to write to the CA (participants 9 and 10), best thing is to make robot work together with the lessons, not replacing them (participant 9), use user's answers more noticeably (participant 9).

Daily life situations

All participants deemed the CA useful for practising daily life situations, except participant 10. They stated that robots cannot predict human behaviour that well just yet. Participant 2 initially had her doubts, since they thought the CA would have to deviate too much from the DM for that. However, if a similar robot could do conversations that are like daily life situations, this would be useful to her. Participant 1 mentioned that for this purpose, it would not be important to him, but would be to people who are nervous to speak in such daily life situations.

Most participants mentioned any type of service situation, such as in the supermarket, ordering in restaurants, or going to the doctor. Three participants would also practice chit-chatting with the CA, and two of them also mentioned asking for directions.

Language use in the Netherlands

Both participants 2 and 10 mentioned that the DM teachers speak in a very general and clear Dutch, while they want to practice with more varied Dutch too. Participant 10 hesitated to fully rely on a robot as they thought it was too clean-cut and not representative of all different kinds of Dutch speakers. They, however, liked the researcher's idea of multiple characters within Furhat that they could practice with. Participants 2 and 9, however, thought that the CA was very suitable for practising with varying Dutch. Another point, which the current work started with in the introduction, was that it is quite easy to default to English in the Netherlands since Dutch people tend to speak Dutch back to non-fluent speakers. For participant 9, the CA would be useful in the sense that he could practice situations which are a bit more complex, for which he normally uses English in the Netherlands. For participant 10, the robot not switching to English like Dutch people tend to is a positive point.

Conversation strategies

A majority of the participants noticed that the CA used strategies to overcome conversation breakdowns. Participants 2, 3, and 5 remarked that they heard the robot rephrasing its questions. Participant 3 emphasised that they thought this was a good strategy. For participants 4, 6, and 9, the repeating requests from the CA to them stood out positively. Furthermore, participant 10 thought that if the CA repeated a question,

it was slower. This feature was not implemented, so it must have been just perceived as such by the user.

Participants 1, 7, 8, 9, and 11 also answered with comments on dialogue design rather than the implemented strategies, like changing the subject, using words from only the lessons, and continuing the conversation even though the CA did not interpret correctly what the user said.

Physical vs. virtual CA

Five participants thought it would be interesting to try a physical version of the robot (participants 1, 2, 8, 9, and 10) as opposed to four others who indicated they or others might be scared of a physical CA (participants 2, 5, 7, 11). Three participants believed that a physical body would allow for a better placement of the microphone, meaning that it would hear better. An advantage of the virtual robot, however, is that it can be used almost anywhere. Four participants explicitly said that the virtual version would be better, while two others thought a physical version would be.

Classroom context

Two participants discussed some interesting points regarding the classroom environment. Participant 2 addressed the problem of talking with internationals: when talking, both speaking partners form English sentences with Dutch words. This leads to direct translation from English to Dutch. Participant 8 found that many classmates are very shy and self-conscious in the classroom, while he has a higher willingness to communicate.

10

Discussion

Overall, participants were quite enthusiastic about the concept of a DM-CA. They could see it being useful and not only for the DM itself but also for practising daily life situations. For them, it would be a good addition to the set of people they can practice spoken Dutch with currently. This is due to the earlier-mentioned cases throughout the current work: Dutch people switching to English, teachers speaking differently compared to other Dutch speakers, and classmates using anglicised Dutch. The participants could see how a CA will not switch to English and how it could have different voices and accents. The permanent availability (for the virtual version) contributes to the positively perceived usefulness as well. The rest of this chapter answers the research questions proposed in Chapter 3, addresses the limitations of this study, and gives recommendations for future work.

10.1 RQ1 - Willingness to Communicate

To what extent can a conversational agent increase the willingness to communicate of learners of Dutch as a second language?

All 18 statements from the WTC questionnaires, except statements 4 and 10, had an increase in the average score. However, for only two statements, this increase was significant:

7. "When I followed all steps of the Delft method to study the text"
13. "When I am sitting close to the teacher"

For statement 7, it means that if students prepare classes well and could interact frequently with a conversational agent as a last step in their preparations, they think it

would stimulate them to speak more in class and thus result in an improvement of willingness to communicate. For statement 13, it means that, especially when sitting close to the teacher, students think they would speak more during class if they could practice often with a conversational agent. Overall, no statements were rated significantly worse on average, indicating that there is at least no significant negative effect on WTC in class by using a DM-CA for preparing lessons.

The qualitative results of the user tests show the overall potential of improving WTC by a CA. Language anxiety appeared to be an important factor that prevented people from speaking up in class. Participants indicated that practising with a CA could increase their confidence so that they would speak more during class. Others indicated that they can imagine how a CA can help others who are nervous to speak (rather than themselves).

10.2 RQ2 - Strategies

Which strategies can be applied to overcome possible conversation breakdowns in a dialogue with a conversational agent in second language learning?

All participants were asked if they had noticed any strategies used by the CA to overcome situations in which one or both interlocutors did not understand the other party. The *rejecting* and the *rephrasing* strategies stood out most. Below are given examples of conversation breakdown solutions per strategy type. Per table, the utterance of the robot and the speech-to-text interpretation of the user's utterance are presented.

Robot	User (STT)
Was de reis voor je werk of voor het plezier? (Was the trip for your work or for fun?)	plezier (fun).
Ik begreep het niet, sorry. Kan je het herhalen? (I don't understand it, sorry. Can you repeat it?)	eh het was plezier te zien mijn family (eh it was fun see my family).
Dat klinkt als een mooie vakantie! (That sounds like a nice vacation!).	-

Table 10.1: *The rejecting strategy overcoming intent classifying issues in the conversation with participant 10.*

In Table 10.1, Participant 10 answered the question of the CA correctly, but it was not possible for the CA to classify it. Therefore, the CA rejected the participant's answer and asked for a repetition. This time, the participant gave a longer answer, which was correctly classified, leading to a continuation of the conversation. Similarly, the CA indicated it did not understand the user in Table 10.2, but this time the question was worded differently (two times after the original question). The difference in approach here is due to the dialogue design. For the question in the conversation with participant 10, there was no special action implemented to take when an answer could not be

Robot	User (STT)
Heb je liever veel natuur of veel cultuur? (Do you prefer a lot of nature or a lot of culture?)	Ja ja ik ik heb over Australië Cultuur en de (Yes yes I I have about Australia culture and the).
Ik begrijp het niet. Heb je liever veel natuur of veel cultuur op reis? (I don't understand. Do you prefer a lot of nature or a lot of culture during travelling?).	Ehm oké ik ik heb ik het liefst (Ehm okay I I have I prefer).
Ik begrijp het niet. Heb je liever veel natuur of veel cultuur op vakantie? (I don't understand. Do you prefer a lot of nature or a lot of culture on vacation?).	Oké eh ik heb liever van cultuur (Okay eh I prefer of culture).
De cultuur is inderdaad interessant ja (Culture is indeed interesting).	-

Table 10.2: *Rephrasing strategy overcoming speech-to-text issues in the conversation with participant 5.*

Robot	User (STT)
Hoe ging je daarnaartoe? (How did you go there?).	Wat zegt u (What are you saying).
Hoe ging je daarnaartoe? (How did you go there?).	met een vlakte vlechten (with a plain braids).
Dat is interessant! (That is interesting!).	-

Table 10.3: *Repeating strategy to accommodate for the user not understanding the robot in the conversation with participant 7.*

Robot	User (STT)
Hoe ging je daarheen? (How did you go there?)	Ik begin (I begin).
Ik begrijp het niet helemaal. Kan je meer woorden gebruiken? Ging je met het vliegtuig, de trein of de auto? (I do not fully understand. Can you use more words? Did you go by plane, train, or car?)	Ik begin met vliegtuig en treinen (I begin with plane and trains).
Dat is interessant! (That is interesting!).	-

Table 10.4: *Requesting elaboration strategy overcoming speech-to-text issues in the conversation with participant 3.*

Robot	User (STT)
Ik hoorde je niet. Wat zei je? (I didn't hear you. What did you say?)	misschien eerst naar die film (maybe first to that film).
Wil je eerst naar de film? (Do you want to go to the cinema first?)	Ja en na eh we gaan naar de restaurant voor eten (yes and after eh we go to the restaurant for food).
Oké, eerst de film. Dat is goed! (Okay, first the cinema. That is alright!).	-

Table 10.5: *Confirming strategy overcoming speech-to-text issues in the conversation with participant 1.*

Robot	User (STT)
Eet je graag vlees, vis of vegetarisch? (Do you like meat, fish, or vegetarian?).	ik (I).
Wat eet je normaal? (What do you eat normally?)	Ik eet naar mijn vis of vlees (I eat to my fish or meat).
We kunnen naar een Argentijns restaurant gaan (We can go to an Argentinian restaurant).	Ja ik (Yes I).

Table 10.6: *Asking follow-up questions strategy overcoming speech-to-text issues in the conversation with participant 4.*

classified. Therefore, it used the default strategy: *rejecting*. The question for participant 5 had the *rephrasing* strategy specifically implemented. Table 10.3 presents how the CA replies when a user indicates they want a repetition of the question. This is more of a reaction to a user's approach to overcoming a conversation breakdown, but still a strategy since the CA follows specific rules to aim at conversation flow maintenance.

Table 10.4 shows how the robot applied the *requesting elaboration* strategy. The CA interpreted too few words from participant 3 and asked them if they could use more words. After that, the answer contained enough words according to the CA. However, it still did not make sense in terms of content. Nevertheless, the CA was able to continue the conversation due to the dialogue design by just answering: "That is interesting!". It did that because the conversation was designed to make this happen. It served as a fallback so that the conversation could go on.

The same can be said for Table 10.5 and Table 10.6: the strategies are applied, but within the dialogue design at those specific states in the programme. They are preprogrammed and not dynamic. Therefore, a distinction could be made between dynamic strategies, static strategies, and designed fallbacks. The first type of strategies can be applied anytime during the conversation when a condition is met, such as when there is no intent classification is possible or when an open question is being asked. The second type can be built within the dialogue at specific states that require such a strategy, for example, states that need a specific follow-up question in case a negative answer was given. Finally, fallbacks are needed for when the strategies cannot overcome the conversation breakdowns. It is supposed to be a last resort since, as it is hard-coded, it might not always make complete sense in terms of content. The following can be categorised as dynamic strategies: rejecting, rephrasing, requesting elaboration, and repeating. The others are static strategies: confirming and asking follow-up questions. Fallbacks can be neutral answers to a user's utterance, such as "that is interesting!" or "good to hear". It does not go into the answers of the user, but still maintains the conversation flow.

10.3 RQ3 - Design Guidelines

Which (conversational) design guidelines can be proposed for a conversational agent in the Delft method context?

The input from the participants of the user tests was used to establish general design guidelines, which can be used as a checklist while developing a DM-CA. Below, Table 10.7, Table 10.8, and Table 10.9 present all proposed guidelines, the reasoning behind them, and to which previous established design requirement each guideline corresponds. Most guidelines are focused on conversational design, but some focus (partially) on other aspects: GL5, GL13, and GL15. The guidelines are categorised as follows: basic functionality, breakdown prevention, experts' wishes, users' wishes, and others. These categories can be used as a priority system for development as well, basic functionality being top priority (green) to ensure the system operates. After comes breakdown prevention (orange), since this is required for maintaining conversation flow. Then the DM experts' wishes (purple) have priority over the users' wishes (red) because they are well-informed about the Delft method and know what is best for their students. Nevertheless, since it regards user-centred design, the ideas and opinions of the users are important to include. Starting with implementing these can happen when there is a stable foundation for the conversational agent. In the text, more details per guideline are discussed.

ID	Guideline: the CA has ...	Rationale: so that users ...	Req. ID
GL1	proper STT and TTS functionalities in Dutch.	can speak and listen to Dutch.	R1
GL2	at least a virtual presence.	can use the agent almost anywhere.	R6
GL3	(verbal) strategies and fallbacks in store such that it can overcome conversation breakdowns.	have conversations going on without too many pauses or restarts.	R5
GL4	to wait a while before determining if the user is done speaking.	have enough time to answer.	N3

Table 10.7: *The proposed design guidelines with the highest priority. Each row shows, from left to right, the ID of a guideline, the actual guideline, the reasoning behind the guideline, and to which requirement ID from previous chapters they correspond.*

GL 1. *STT and TTS functionalities in Dutch* - This guideline is important for basic functionality. Without properly transcribing what the user is saying, the conversational agent cannot function well. Certain parameters for the text-to-speech can vary, such as voice, gender, and accent.

GL 2. *At least a virtual presence* - Another basic functionality guideline is that there must be some form of presence of the agent for the user to be able to interact with it. Virtual should be the bare minimum, while physical is an option too.

GL 3. (*Verbal*) *strategies and fallbacks to overcome conversation breakdowns* - As discussed previously, the design of a dialogue can be such that conversation breakdowns are overcome (fallbacks). That is why this guideline is an updated version of requirement R5. Furthermore, non-verbal strategies can be used to overcome conversation breakdowns. These have not been tested during this study, but participants indicated that facial expressions of the CA can help with that and that they themselves also indicate conversation breakdowns through (facial) gestures.

GL 4. *Wait a while before determining whether the user is done speaking* - As was seen in the results of the user tests, robot interruption occurred quite often. Users need extra time to think about their answers, especially since they are not native speakers. The tests confirmed this requirement, and therefore, it is an important guideline.

ID	Guideline: the CA has ...	Rationale: so that users ...	Req. ID
GL5	the capability to show "understanding".	feel understood to some extent.	R2
GL6	a database for words taught per lesson.	hear the vocabulary learned at the time of interaction.	R3
GL7	the capability to hold interesting and logical conversations.	can have meaningful conversations.	R4
GL8	conversations of around 5 minutes per DM lesson (as long as they stay meaningful).	have conversations that take long enough yet are not too long.	N1
GL9	ability to challenge students to speak about the topic of the text specifically.	are able to test if they studied the texts well.	N2
GL10	to request a minimum number of words for answers to open questions.	are encouraged to speak.	N4
GL11	to use mainly open questions.	are encouraged to speak.	N4
GL12	to avoid giving answers to its own questions.	are encouraged to speak.	N4
GL13	utterances and questions that are of sufficient length.	are encouraged to listen.	N5
GL14	a set of questions with the same meaning/intentions at each conversation state.	will hear a variety of utterances from the CA.	N7
GL15	dialogues that are not too sensitive.	can have personal conversations yet maintain their privacy.	N8
GL16	a temporary user data memory.	can have personal conversations yet maintain their privacy.	N8
GL17	the ability to give a reasonable amount of implicit and reinforced feedback.	receive implicit and reinforced feedback.	R9
GL18	a way of tracking time.	limit their time spent on a conversation.	R11

Table 10.8: *The proposed design guidelines with medium priority. Each row shows, from left to right, the ID of a guideline, the actual guideline, the reasoning behind the guideline, and to which requirement ID from previous chapters they correspond.*

GL 5. *Capability to show "understanding"* - Users made clear that they want to feel understood. This holds both for understanding what the users say and what they mean. Understanding can be shown by repeating the information the user gives or asking logical questions, for example.

GL 6. *Database for words taught per lesson* - A crucial part of the DM is the correct use of vocabulary. This cannot be underestimated and needs to be followed strictly by the CA.

GL 7. *Capability to hold interesting and logical conversations* - Users want meaningful conversations that make sense. The conversations should be about the studied texts but not just random questions. The CA should make nice connections between previous questions and users' answers for a proper conversation.

GL 8. *Conversations of around five minutes* - The conversations per lesson with Prototype II were not fully five minutes long, and one was shorter than the other. Five minutes is still a good goal for each conversation, as long as the conversations remain meaningful.

GL 9. *Challenge students to speak about the topic of the text specifically* - The CA is especially useful if a user can test with it whether they prepared well enough for the upcoming class. Therefore, the design of the dialogues should match the lessons properly.

GL 10. *Request a minimum number of words for answers to open questions* - The minimum number of words requested helps students think more about their answers and produce more in the target language.

GL 11. *Use mainly open questions* - Open questions stimulate users to speak more, which is the goal of the interaction. From solely answering "yes" or "no", the production of language does not get trained.

GL 12. *Avoid giving answers to its own questions* - The CA should not give answers to its own questions. It can give hints or suggestions, but the user needs to speak.

GL 13. *Utterances and questions of sufficient length* - Short utterances and questions will not train the user on listening.

GL 14. *Set of questions with the same meaning/intentions at each conversation state* - This ties into the *rephrasing* strategy. By rephrasing, the users do not constantly hear the same utterances when a conversation breakdown occurs, and it, in turn, might help to overcome a conversation breakdown.

GL 15. *Not too sensitive dialogues* - Compared to the corresponding requirement (N8), the "built-in" part was removed. This was done because a CA does not necessarily need built-in dialogues. If it is created using an LLM, for example, it is still possible to prompt the generations such that not too personal or sensitive questions are asked.

GL 16. *Temporary user data memory* - To maintain privacy but yet keep a conversation personalised, user data can be stored during the conversation. This way, the CA can recall and address, during the conversation, some of the information the user has given. After the conversation ends, these data should be erased.

GL 17. *Reasonable amount of implicit and reinforced feedback* - As the teachers predicted in the interviews, participants would like to receive explicit feedback on their pronunciation, grammar, and/or word use. However, this is not in line with the Delft method. Instead, implicit feedback should suffice and otherwise reinforced feedback, such as compliments and encouragements, can help show a user that they are doing well. Some participants agreed that implicit feedback is already sufficiently useful. As the theory behind the Delft method stresses, vulnerable students should be handled carefully with corrections and be praised more for their efforts, which is in line with this guideline.

GL 18. *Tracking time* - To prevent loops or dead ends, a set time limit can help. At some point, a conversation should end.

ID	Guideline: the CA has ...	Rationale: so that users ...	Req. ID
GL19	multi-modal functionalities for navigating through the conversation	can control the conversation if it somehow stops working.	N6
GL20	a mask with human facial features.	see a human-like face to talk to.	R6
GL21	the possibility for students to indicate the desired amount of feedback.	have a personalised amount of feedback.	R10
GL22	possibility for students to indicate their current DM lesson.	can practice all lessons from the courses.	N10

Table 10.9: *The proposed design guidelines with lower priority. Each row shows, from left to right, the ID of a guideline, the actual guideline, the reasoning behind the guideline, and to which requirement ID from previous chapters they correspond.*

GL 19. *Multi-modal functionalities for navigating through the conversation* - The buttons in both prototypes were wizarding buttons, meaning that the user could not control the conversations through these. However, during the co-design and user tests, it appeared that users would like to have these buttons or something similar to control the conversation themselves through other modalities than speech if some errors occur. This should be a feature whose use should be minimised, as the goal is to solve errors through speech.

GL 20. *Human facial features* - The human facial features are important, but the face does not have to be fully human. Cartoon characters and similar were mentioned as better alternatives by the participants. The appearance of Furhat was not too successful, as many participants did not keep looking at it throughout the conversations.

GL 21. *Indicate desired amount of feedback* - Since people differ, users want to be able to choose how much feedback they will receive. Some might need more reinforced feedback than others.

GL 22. *Indicate current DM lesson* - This guideline comes from a won't-have-requirement.

Having all lessons available is mainly important for a full-fledged CA end product, which was out of the scope for the current work.

10.4 Limitations and Future Research

First of all, it should be noted that there is some bias in this current work. Both DM experts were female, and most of the participants in the user tests were female as well (seven out of 11). Since female NT2 learners profit more from higher educational L2 training than male learners in L2 acquisition (Van der Slik et al., 2015), feedback from women on language tools can differ as well. Secondly, as noted earlier, one of the experts worked on developing the Delft method and therefore might have a positively biased view on the method.

There were also some technical limitations: the speaker and microphone used during the testing of both Prototype I and Prototype II were built into the researcher's laptop. This means that since there was a distance between the monitor and the laptop, the location of the audio source and input point felt a bit off to the participants. Furthermore, the WTC questionnaire consisted only of statements within the DM classroom context. This, while the CA could also be useful for outside-classroom environments and situations. It cannot be stated whether this CA is useful for WTC in that context, since this has not been addressed in the user tests.

The researcher noted that in addition to verbal communication, non-verbal communication was used frequently during the conversations. Mostly facial expressions were used, for example, when a participant was confused. Situations happened where a user did not know how to answer, but only showed that through their facial expression. The CA could not notice this, which caused illogical answers or interruptions from the CA. A physical Furhat has a camera and functionalities to recognise facial gestures, which would enhance conversations in terms of overcoming conversation breakdowns. Participants could indicate their non-understanding using their face or hand gestures. This is not included in the design guidelines, since the current work focuses on conversational design. However, participants were interested in trying a physical version, so future research can compare such a robot to the virtual one. A second topic for future research can be the use of LLMs to generate dialogues. The main challenge would be to stick to the limited vocabulary for each DM lesson. If that is possible, conversations can vary more, and it is also possible to have the user ask questions back to the CA. In addition, it saves development time because there is no need to design a full dialogue for each lesson.

As can be seen in the proposed design guidelines, users would like to see customisation: ideas such as choosing the speed of talking, choosing the level of speech, or choosing the amount of feedback they will receive. These are some interesting points that could

improve the CA's functionality, but would require some research and testing in future work as well.

Conversational strategies can be explored further, too. For now, six strategies were implemented and tested. One that was not tested, although it was planned to do so, is giving hints. For this, the used voice model should allow for changing the intonation within utterances to create sentences that need to be finished by the user, i.e. a form of cloze test (fill in the blanks), but then with speech instead of writing. In this way, it might be possible to assist the user when they do not know what kind of answer to give to the robot. In essence, it is how teachers approach such situations, too, in DM classes. Adding more strategies might help in preventing and overcoming more conversational breakdowns. The current prototype was also taking initiative, meaning that users could not ask questions back to the CA. It appeared that higher NT2 level students tended to ask questions back to Furhat, such as "What was your favourite trip?". Therefore, it is recommended to explore initiative-switching for higher-level DM courses CAs.

11

Conclusion

This work set out to address the following problem:

"How to create a conversational agent, suiting the Delft method, that leads to more speaking opportunities for Dutch learners and motivates them to speak Dutch more?"

To explore this, a Furhat-based conversational agent was developed and evaluated following the interaction design lifecycle model. Delft method experts and their students interacted with the conversational agent, allowing for iterative refinement of design requirements and the formulation of guidelines for future development.

The evaluations revealed that the conversational agent holds promise for enhancing speaking opportunities and supporting learners' willingness to communicate in Dutch. Students responded positively to the concept, and they imagined it could be a useful tool for preparing classroom lessons. The idea of using such a conversational agent appeared to support or at least sustain students' willingness to communicate in class, suggesting that its integration could be compatible with classroom participation. In particular situations, such as when students sit near the teacher or are fully prepared for class, their willingness to communicate was even positively influenced. Future work should focus on outside-classroom speaking situations since participants were enthusiastic about using a conversational agent to practice those as well.

Automatic speech recognition was identified as the primary cause of conversation breakdowns, highlighting the importance of further developing GL5 - conversational strategies and fallbacks. Even though the strategies showed effectiveness in overcoming these issues, future work is needed to figure out how to manage these situations more effectively. If future work makes use of large language models, dynamic strategies

should be the focus. Static strategies and fallbacks are implemented at specific pre-programmed dialogue points, while dynamic strategies can be applied at almost any point in a dialogue generated on the spot. Another key issue was robot interruption (GL8), as the conversational agent was prone to interrupting users too often, which disrupted the flow of interaction. A method for defining waiting thresholds should be developed so that users will not get interrupted while thinking or speaking.

Furthermore, non-verbal and other aspects of the conversational agent, such as facial gestures, control buttons, or timing features, should be explored further since they can enhance the total user experience of interaction and also improve the overcoming of conversation breakdowns. These have lower priority, however, due to the conversational design focus of this work.

In sum, this project has opened a promising path toward developing a CA tailored for NT2 learners using the Delft method. However, significant work remains in refining the technology and testing its effectiveness in practice before it can become a reliable and widely usable educational tool.

References

Ayedoun, E., Hayashi, Y., and Seta, K. (2015). "A Conversational Agent to Encourage Willingness to Communicate in the Context of English as a Foreign Language". In: *Procedia Computer Science* 60, pp. 1433–1442. DOI: https://doi.org/10.1016/j.procs.2015.08.219.

Ayedoun, E., Hayashi, Y., and Seta, K. (2019). "Adding Communicative and Affective Strategies to an Embodied Conversational Agent to Enhance Second Language Learners' Willingness to Communicate". In: *International Journal of Artificial Intelligence in Education* 29, pp. 29–57. DOI: https://doi.org/10.1007/s40593-018-0171-6.

Ayedoun, E., Hayashi, Y., and Seta, K. (2020). "Toward Personalized Scaffolding and Fading of Motivational Support in L2 Learner–Dialogue Agent Interactions: An Exploratory Study". In: *IEEE Transactions on Learning Technologies* 13 (3), pp. 604–616. DOI: https://doi.org/10.1109/TLT.2020.2989776.

Bashori, M., Van Hout, R., Strik, H., and Cucchiarini, C. (2022). "Web-Based Language Learning and Speaking Anxiety". In: *Computer Assisted Language Learning* 35 (5-6), pp. 1058–1089. DOI: https://doi.org/10.1080/09588221.2020.1770293.

Benner, D., Elshan, E., Schöbel, S., and Janson, A. (Dec. 2021). "What do you mean? A Review on Recovery Strategies to Overcome Conversational Breakdowns of Conversational Agents". In: *International Conference on Information Systems*.

Blom, A. (2006). "NT2-les Zonder Grammatica". In: *Levende Talen Tijdschrift* 7 (4), pp. 20–27.

Chapelle, C. A. (2010). "The Spread of Computer-Assisted Language Learning". In: *Language Teaching* 43 (1), pp. 66–74. DOI: https://doi.org/10.1017/S0261444809005850.

Clegg, D. and Barker, R. (1994). *Case Method Fast-track: A RAD Approach*. Addison-Wesley Longman Publishing Co., Inc.

Derwing, T. M. and Munro, M. J. (2000). "Does Popular Speech Recognition Software Work with ESL Speech?" In: *TESOL Quarterly* 34 (3), pp. 592–603. DOI: https://doi.org/10.2307/3587748.

Derwing, T. M. and Munro, M. J. (Oct. 2009). "Putting Accent in its Place: Rethinking Obstacles to Communication". In: *Language Teaching* 42 (4), pp. 476–490. DOI: https://doi.org/10.1017/S026144480800551X.

Divekar, R. R., Lepp, H., Chopade, P., Albin, A., Brenner, D., and Ramanarayanan, V. (2021). "Conversational Agents in Language Education: Where They Fit and Their Research Challenges". In: *International Conference on Human-Computer Interaction*. Cham: Springer International Publishing, pp. 272–279. DOI: https://doi.org/10.1007/978-3-030-90179-0_35.

EF English Proficiency Index (2023). EF Education First. URL: https://www.ef.nl/epi/.

Furhat (2024). *Furhat Developer Docs*. URL: https://docs.furhat.io/.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., and Freynik, S. (2014). "Technologies for Foreign Language Learning: a Review of Technology Types and Their Efectiveness". In: *Computer Assisted Language Learning* 27 (1), pp. 70–105. DOI: https://doi.org/10.1080/09588221.2012.700315.

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., and Maisonnier, B. (2009). "Mechatronic Design of NAO Humanoid". In: *2009 IEEE International Conference on Robotics and Automation*, pp. 769–774. DOI: https://doi.org/10.1109/ROBOT.2009.5152516.

Horwitz, E. K. (Jan. 2001). "Language Anxiety and Achievement". In: *Annual Review of Applied Linguistics* 21, pp. 112–126. DOI: https://doi.org/10.1017/S0267190501000071.

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). "Foreign Classroom Anxiety". In: *The Modern Language Journal* 70 (2), pp. 125–132.

Huang, G. and Moore, R. K. (2023). "Using Social Robots for Language Learning: Are We There Yet?" In: *Journal of China Computer-Assisted Language Learning* 3 (1), pp. 208–230. DOI: https://doi.org/10.1515/jccall-2023-0013.

Hulstijn, J. H. (Nov. 2018). "Honderd Jaar Theorie en Onderzoek van Tweedetaalverwerving: Afgedankte Inzichten en Nog Onopgeloste Raadsels". In: *Internationale Neerlandistiek* 56 (3), pp. 189–202. DOI: https://doi.org/10.5117/IN2018.3.002.HULS.

Inceoglu, S., Chen, W.-H., and Lim, H. (2023). "Assessment of L2 Intelligibility: Comparing L1 Listeners and Automatic Speech Recognition". In: *ReCall* 35 (1), 89=104. DOI: https://doi.org/10.1017/S0958344022000192.

Jepson, K (Sept. 2005). "Conversations—and negotiated interaction—in text and voice chat rooms". In: *Language Learning & Technology* 9 (3), pp. 79–98.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Upper Saddle River, NH,USA: Prentice-Hall, Inc., pp. 841–846. ISBN: 0131873210.

Koster, D. E. S. (2015). "A Dynamic, Usage-Based Approach to Teaching L2 Dutch". In: *Dutch Journal of Applied Linguistics* 4 (2), pp. 257–264. DOI: https://doi.org/10.1075/dujal.4.2.08kos.

Lee, K.-A. and Lim, S.-B. (2023). "Designing a Leveled Conversational Teachable Agent for English Language Learners". In: *Applied Sciences* 13 (11). DOI: https://doi.org/10.3390/app13116541.

Lee, S. and Jeon, J. (2024). "Visualizing a Disembodied agent: Young EFL Learners' Perceptions of Voice-Controlled Conversational Agents as Language Partners". In: *Computer Assisted Language Learning* 37 (5-6), pp. 1048–1073. DOI: https://doi.org/10.1080/09588221.2022.2067182.

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). "The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains". In: *34th Annual Conference of the Cognitive Science Society*.

Li, J. (2015). "The Benefit of Being Physically Present: A Survey of Experimental Works Comparing Copresent Robots, Telepresent Robots and Virtual Agents". In: *International Journal of Human-Computer Studies* 77, pp. 23–37.

MacIntyre, P. D. (2017). "An Overview of Language Anxiety Research and Trends in its Development". In: *New Insights into Language Anxiety: Theory, Research and Educational Implications*, pp. 11–30. DOI: https://doi.org/10.21832/9781783097722-003.

MacIntyre, P. D., Clément, R., Dörnyei, Z., and Noels, K. A. (Dec. 1998). "Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation". In: *The*

Modern Language Journal 82 (4), pp. 545–562. DOI: http://dx.doi.org/10.1111/j.1540-4781.1998.tb05543.x.

McCrocklin, S. and Edalatishams, I. (Dec. 2020). "Revisiting Popular Speech Recognition Software for ESL Speech". In: *TESOL Quarterly* 54 (4), pp. 1086–1097. DOI: https://doi.org/10.1002/tesq.3006.

McCroskey, J. C. and Baer, J. E. (Nov. 1985). "Willingness to Communicate: The Construct and Its Measurement". In: *Speech Communication Association*. Denver.

Menzel, K. E. and Carrell, L. J (Jan. 1999). "The Impact of Gender and Immediacy on Willingness to Talk and Perceived Learning". In: *Communication Education* 48 (1), pp. 31–40.

Moussalli, S. and Cardoso, W. (2016). "Are Commercial 'Personal Robots' Ready for Language Learning? Focus on Second Language Speech'". In: *CALL Communities and Culture - Short Papers from EUROCALL*. DOI: https://research-publishing.net/book?10.14705/rpnet.2016.EUROCALL2016.9781908416445.

Moussalli, S. and Cardoso, W. (2019). "Intelligent Personal Assistants: Can They Understand and Be Understood by Accented L2 Learners?" In: *Computer Assisted Language Learning* 33 (8). DOI: https://doi.org/10.1080/09588221.2019.1595664.

NT2 Boom (2019). URL: https://nt2methodes.nl/demo/groeneboek/v1/les01/leren/1.

Palanica, A., Thommandram, A., Lee, A., Li, M., and Fossat, Y. (2019). "Do You Understand The Words That Are Comin Outta My Mouth? Voice Assistant Comprehension of Medication Names". In: *Digital Medicine* 2 (1). DOI: https://doi.org/10.1038/s41746-019-0133-x.

Shadiev, R. and Yang, M. (2020). "Review of Studies on Technology-Enhanced Language Learning and Teaching". In: *Sustainability* 12 (2). DOI: https://doi.org/10.3390/su12020524.

Sharp, Helen, Preece, Jennifer, and Rogers, Yvonne (2019). *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons.

Søgaard, A. (Dec. 2022). "Should We Ban English NLP for a Year?" In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5254–5260).

Tai, T. Y. (2024). "Effects of Intelligent Personal Assistants on EFL Learners' Oral Proficiency Outside the Classroom". In: *Computer Assisted Language Learning* 37 (5-6), pp. 1281–1310. DOI: https://doi.org/10.1080/09588221.2022.2075013.

Tai, T. Y. and Chen, H. H. J. (2023). "The Impact of Google Assistant on Adoloscent EFL Learners' Willingness to Communicate". In: *Interactive Learning Environments* 31 (3), pp. 1485–1502. DOI: https://doi.org/10.1080/10494820.2020.1841801.

Tai, T. Y. and Chen, H. H. J. (2024). "The Impact of Intelligent Personal Assistants on Adolescent EFL Learners' Speaking Proficiency". In: *Computer Assisted Language Learning* 37 (5-6), pp. 1224–1251. DOI: https://doi.org/10.1080/09588221.2022.2070219.

Teske, K. (2017). "Duolingo". In: *CALICO Journal* 34 (3), pp. 393–401. DOI: https://doi.org/10.1558/cj.32509.

Van Boxtel, S., Wesdijk, C., Kraal, P., Schenk, L., and Van Laar, A. (2021). *Docentenhandleiding deel C. Hoe werkt de Delftse methode? Inzichten uit wetenschap en lespraktijk*. Boom NT2.

Van der Slik, F., Hout, R., and Schepens, J. (Nov. 2015). "The Gender Gap in Second Language Acquisition: Gender Differences in the Acquisition of Dutch among Immigrants from 88 Countries with 49 Mother Tongues". In: *PloS one* 10. DOI: https://doi.org/10.1371/journal.pone.0142056.

Van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., and Strik, H. (2016). "Evaluating Automatic Speech Recognition-Based Language Learning Systems: a Case Study". In: *Computer Assisted Language Learning* 29 (4), pp. 833–851. DOI: https://doi.org/10.1080/09588221.2016.1167090.

Varonis, E. M. and Gass, S. (1985). "Non-Native/Non-Native Conversations: A Model for Negotiation of Meaning". In: *Applied Linguistics* 6 (1), pp. 71–90.
Wesdijk, C., Annema, J., and Van Boxtel, S. (2021a). *Docentenhandleiding deel A. Lesuitwerkingen les 1-42 van het Groene Boek*. Boom NT2.
Wesdijk, C., Van Boxtel, S., and Van Laar, A. (2021b). *Docentenhandleiding deel B. Handleiding Delftse methode: praktische didaktiek, gebruik en implementatie*. Boom NT2.
Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P. R., Clark, L., and Cowan, B. R. (2020). "See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. Oldenburg, Germany.
Xiao, F., Zhao, P., Sha, H., Yang, D., and Warschauer, M. (2024). "Conversational Agents in Language Learning". In: *Journal of China Computer-Assisted Language Learning* 4 (2). DOI: https://doi.org/10.1515/jccall-2022-0032.
Yu, M. (2015). "An Examination of the Dynamic Feature of WTC Through Dyadic Group Interaction". In: *System* 55, pp. 11–20.

Appendices

A

Semi-Structured Interview Questions

Here, one can read the questions planned to be asked during the semi-structured interviews with the Delft method teachers. The questions are in Dutch since this was the language used.

A.1 Introductie

Ik doe onderzoek naar hoe een conversational agent kan helpen met het leren van de Nederlandse taal aan anderstaligen. Nu zou ik graag het perspectief willen vanuit de docenten, om zo designrichtlijnen op te kunnen zetten en deze toe te kunnen passen op het ontwikkelen van een prototype. Daarnaast onderzoek ik hoe zo'n agent kan bijdragen aan het verbeteren van de bereidheid tot communiceren (willingness to communicate).

A.2 Introductievragen

1. Hoe lang werk je al met de Delftse methode?
2. Kun je heel kort uitleggen wat de Delftse methode is en hoe het anders is dan andere methoden?
3. Wat zijn de sterkste punten van de Delftse methode?
4. Zijn er ook nadelen? Zo ja, welke?

A.3 Directe vragen

1. Welke digitale/elektronische hulpmiddelen worden er gebruikt binnen de Delftse methode?
   - Zijn de ervaringen (vanuit docenten en wat je hoort vanuit leerlingen) hiervan positief?
2. Wat denk je dat een conversational agent is? En wat deze allemaal kan?
   - Uitleg over wat ik bedoel met een conversational agent: dit is een of digitale of fysieke agent waarmee je een gesprek kan voeren. Deze reageert op wat jij zegt en stelt vragen om het gesprek op gang te houden. ChatGPT kan bijvoorbeeld als conversational agent worden gezien, maar ook een robot waarmee je echt kan praten (en dus niet alleen bevelen geven).
3. Wat zou zo'n conversational agent moeten kunnen?
4. Hoe zou zo'n conversational agent precies van toegevoegde waarde kunnen zijn voor de Delftse methode?
   - Op welk moment binnen de methode?
   - Waar zou deze vooral (bij uitstek) goed in moeten zijn? Bijvoorbeeld het uitlokken van het spreken, geen fouten maken, herkennen van fouten?
5. Vanaf welk niveau NT2 zou een CA kunnen geïntroduceerd worden?
6. De leraar mag alleen woorden gebruiken die geleerd zijn uit de tekst. Gaat dit dan voornamelijk om inhoudelijke woorden? Of ook lidwoorden, voorzetsels, voornaamwoorden, etc.?
   - Ik zag dat er in les 1 naar landen wordt gevraagd. Deze staan natuurlijk niet allemaal in de tekst. Zijn er meer van zulke voorbeelden?
7. Hoe zijn jouw ervaringen met conversational agents?
   - Tekstuele chatbots
   - Sprekende assistenten (Google Assistant, Siri)
   - Large Language Models (ChatGPT)
8. Voorgaanden hebben allemaal geen personage (lichamelijke of digitale representatie) dat de agent representeert. In hoeverre denk/vind jij dat dit belangrijk zou zijn in de context van de Delftse methode?
9. Furhat is de conversatierobot die ik ga gebruiken. Hier heb ik een voorbeeldvideo van hoe deze eventueel gebruikt kan worden in het aanleren van een tweede taal: video. Deze kan zowel virtueel als fysiek gebruikt worden. Wat denk je van deze agent?
10. Zie je het voor je hoe een conversational agent kan bijdragen aan de bereidheid tot communiceren (willingness to communicate)? Zo ja, in welke aspecten precies?
11. Er zitten natuurlijk ook negatieve kanten aan het gebruik van AI in onderwijs. Zo kan het bijvoorbeeld zijn dat het niet uit te leggen is waarom een agent iets precies zegt. Wat zijn de discussiepunten hierover binnen jullie vakgroep?
12. Sta je open voor nog een sessie wanneer ik verder ben in het ontwikkelen van de conversational agent?

A.4 Opvolgvragen en diepgaande vragen (voor meer informatie indien nodig)

- Kan je me een voorbeeld geven van X?
- Kan je Y in meer details uitleggen?
- Waarom denk je dat?

A.5 Structurerende vragen

- Laten we verder gaan naar...
- Zullen we het nu over X hebben?
- Ik wil dan nu vragen of je al ervaring hebt met Y...

A.6 Afsluiting

1. Zijn er verder nog opmerkingen die je kwijt wilt?
2. Dan wil ik je bedanken voor het meedoen aan dit interview.

B

Mapping of Experts Interviews Results

Down below, parts of the map with the results of the interviews with the DM experts are presented. Their opinions, remarks, knowledge, etc., were organised per theme and similar or contradicting points got connected. These maps served partially as a foundation for the first design requirements. The full map can be viewed online.

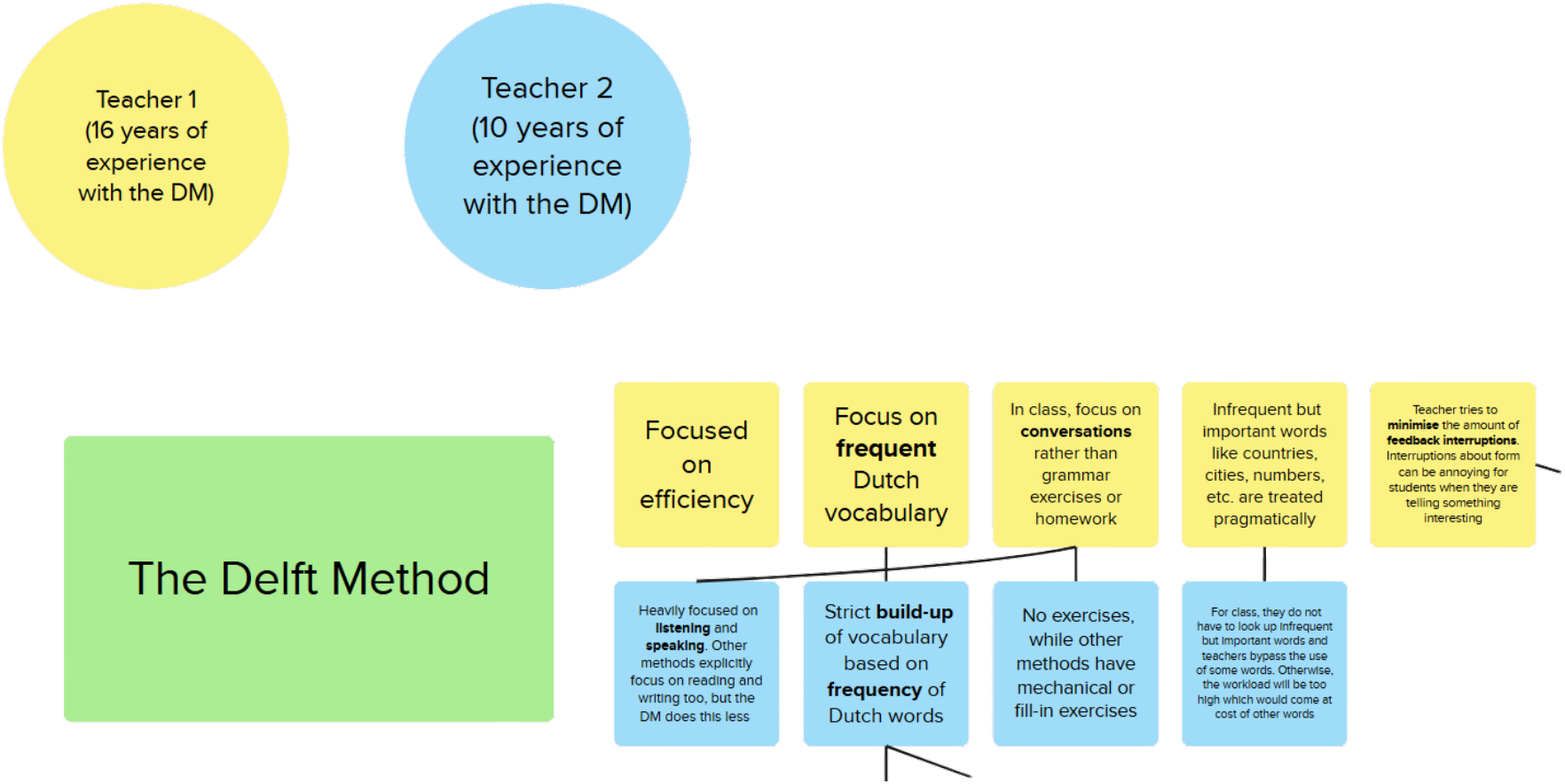


Figure B.1: *Part of the map highlighting the opinions, remarks, etc. from the DM teachers, organised by theme. The yellow post-its are from Teacher I (16 years of experience with the DM), while the blue post-its are from Teacher II (10 years of experience).*

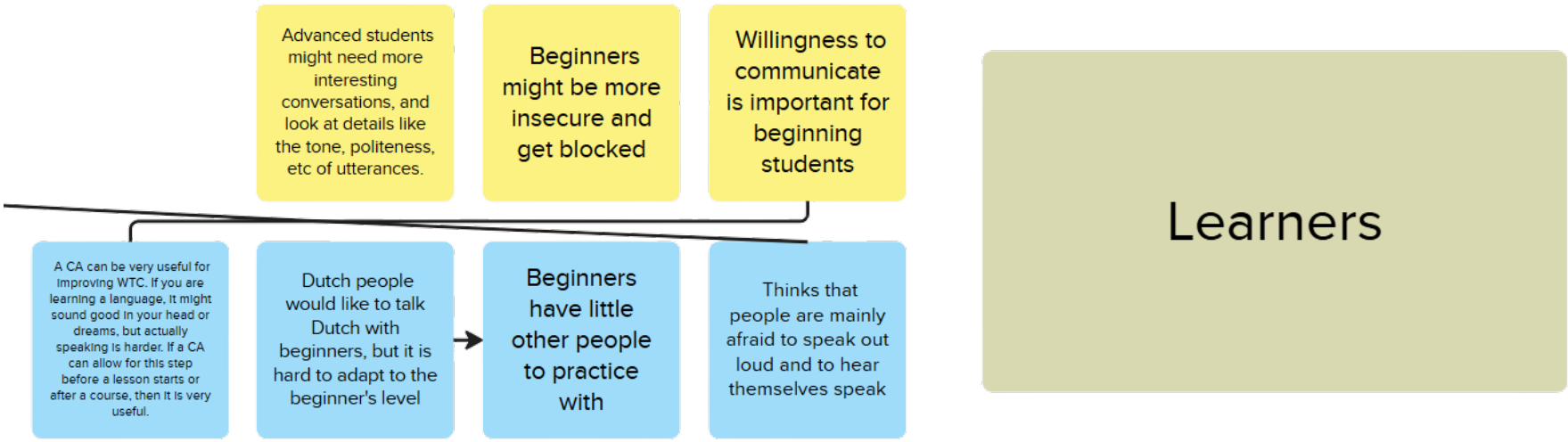


Figure B.2: *Part of the map with the results related to learners.*

C

Dialogue Designs Flowcharts - Prototype I

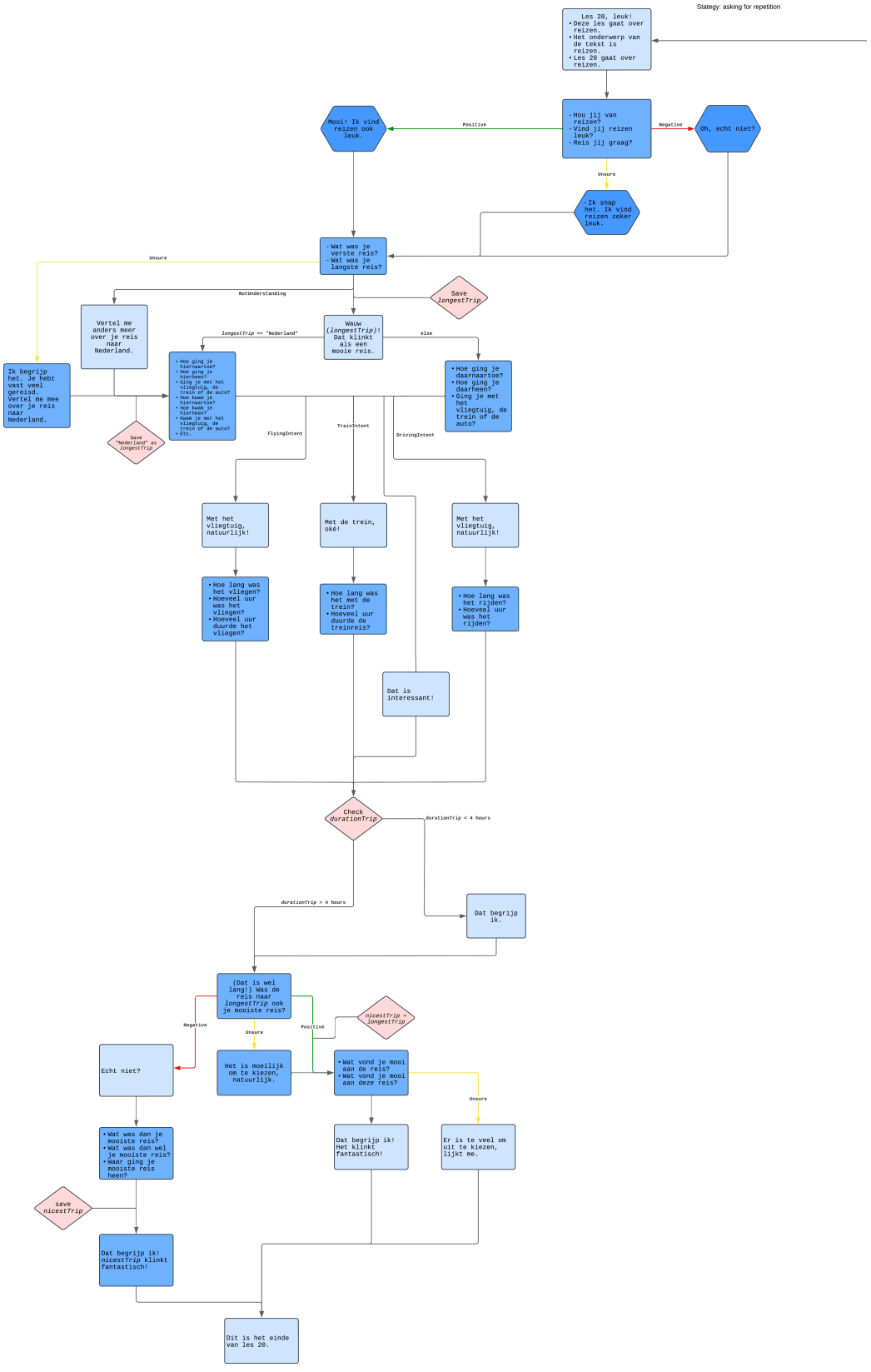


Figure C.1: *Les 20*

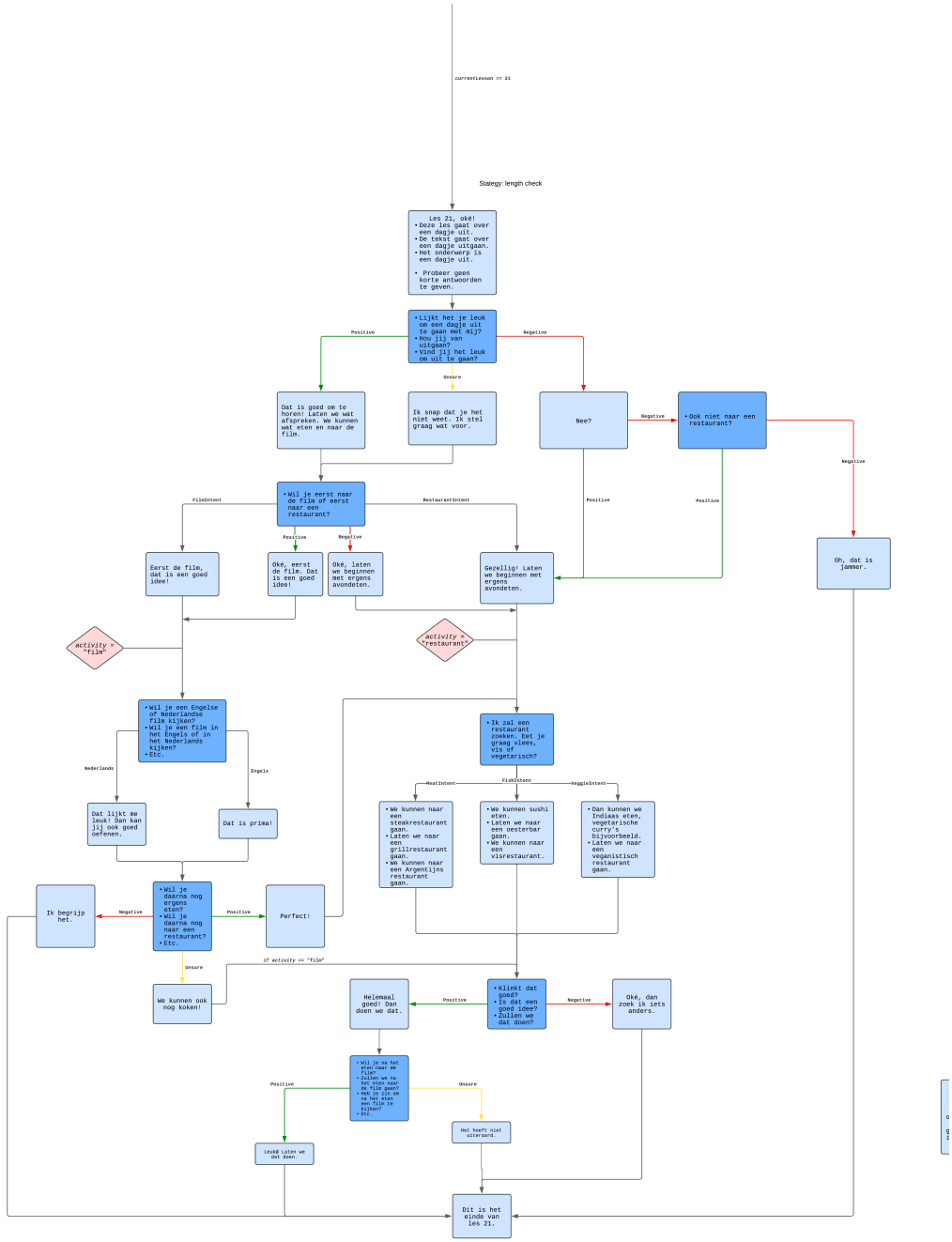


Figure C.2: *Les 21*

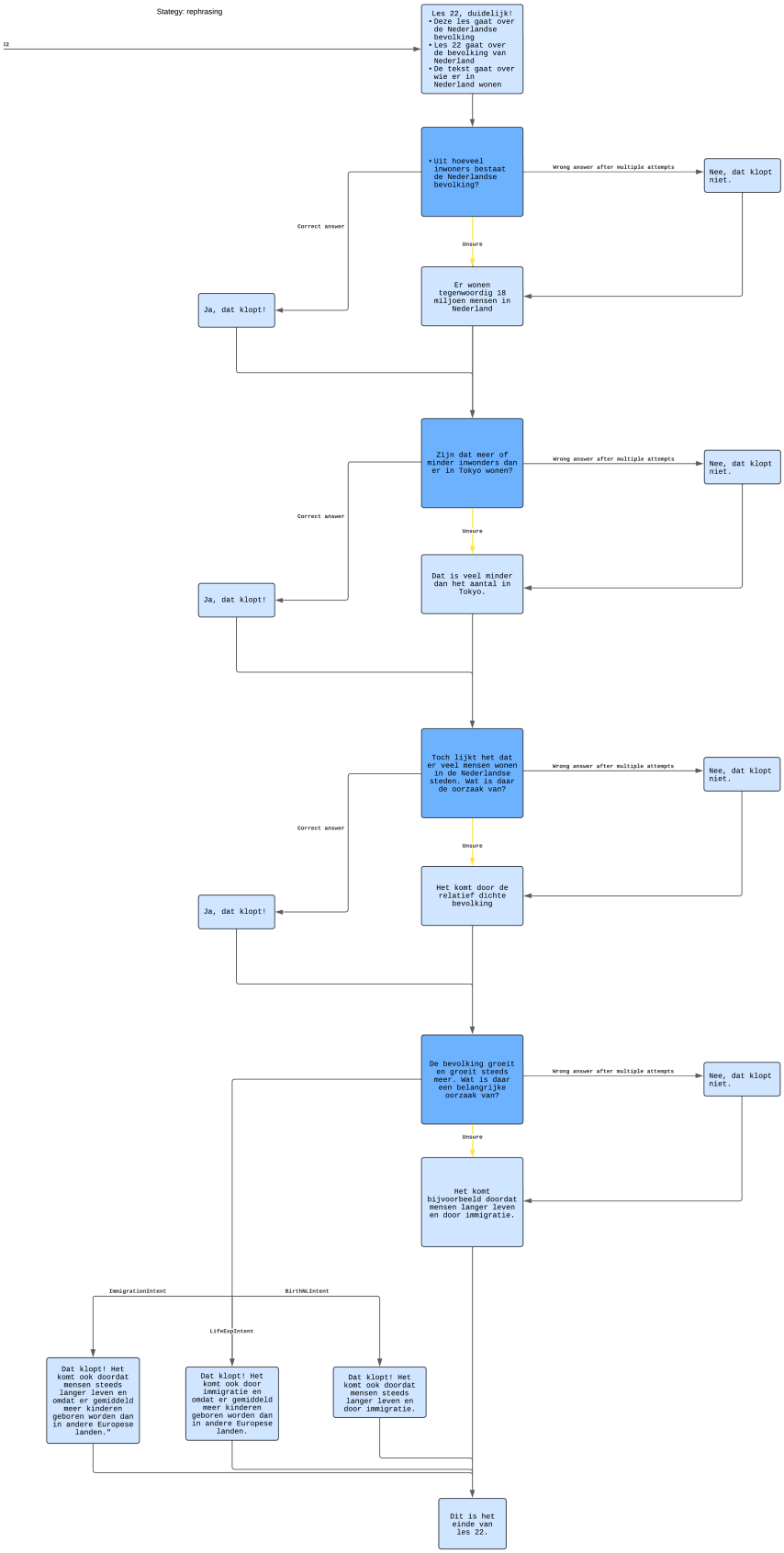


Figure C.3: *Les 22*

D

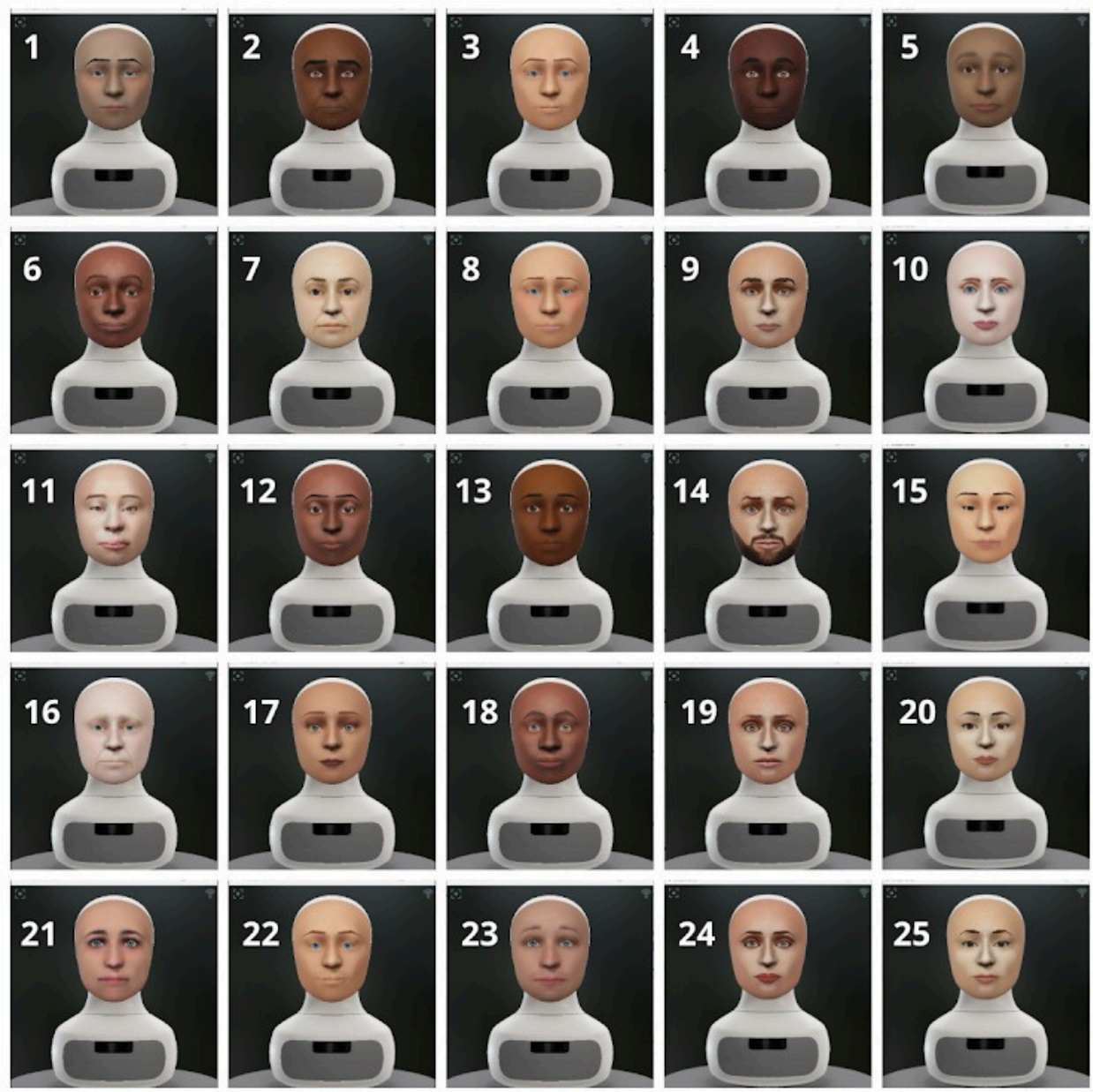
Co-Design Sessions

D.1 Interview

1. Ik hoor graag je eerste indrukken en meningen. Wat vond je van de gesprekjes?
2. Uiterlijk
   - Wat vind je van het uiterlijk van de robot? Let wel op, alleen het gezicht zou kunnen veranderen.
   - Zou de robot van gezicht/personage mogen veranderen? Bijvoorbeeld zoals in de Duolingo app, waar je meerdere personages hebt die elkaar afwisselen bij opdrachtjes.
3. Delftse methode
   - Wat vind je van de keuze voor deze lessen van de Delftse methode?
   - Sluit het gesprek van de robot goed aan bij de lessen?
4. Miscommunicatie; Er werden dus drie verschillende strategieën in totaal gebruikt door de robot om miscommunicatie te verhelpen: vragen om herhaling door de gebruiker, vragen naar bevestiging en herformuleren.
   - Wat vond je ervan hoe de robot miscommunicaties aanpakte?
   - Zijn er nog andere strategieën die zouden kunnen werken?
     - Zijn er voorbeelden die jullie in de les toepassen?
5. Multimodaliteit
   - Tijdens het gesprek zag je alleen de robot. Zou het wel of niet een goed idee zijn om het gesprek tekstueel te laten zien tijdens het gesprek aan studenten?
     - Zouden knoppen ook helpen?

D.2 Survey

1. Welke gezichten zijn het meest toepasselijk? (Meerdere antwoorden mogelijk)*



Vink alle toepasselijke opties aan.

- [ ] Optie 1
- [ ] Optie 2
- [ ] Optie 3
- [ ] Optie 4
- [ ] Optie 5
- [ ] Optie 6
- [ ] Optie 7
- [ ] Optie 8
- [ ] Optie 9
- [ ] Optie 10
- [ ] Optie 11
- [ ] Optie 12
- [ ] Optie 13

- [ ] Optie 14
- [ ] Optie 15
- [ ] Optie 16
- [ ] Optie 17
- [ ] Optie 18
- [ ] Optie 19
- [ ] Optie 20
- [ ] Optie 21
- [ ] Optie 22
- [ ] Optie 23
- [ ] Optie 24
- [ ] Optie 25

Stellingen

Onderstaande stellingen gaan over de verschillende aspecten van de robot en de gesprekjes. Elke stelling mag je beoordelen van 1 t/m 5. De getallen corresponderen met het volgende:

1. Helemaal mee oneens
2. Mee oneens
3. Neutraal
4. Mee eens
5. Helemaal mee eens

2. De robot verstaat mij goed*
Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

3. De robot begrijpt mij goed*
Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

4. De robot zou internationale studenten ook verstaan*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

5. De robot zou internationale studenten ook begrijpen*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

6. De gesprekken passen goed bij de bijbehorende lessen *

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

7. De woorden die de robot gebruikte kwamen overeen met de juiste woordenschat *

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

8. De gesprekjes verliepen logisch*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

9. De gesprekjes waren interessant *

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

10. De gesprekjes duurde lang genoeg*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

11. De gesprekjes stimuleren tot oefenen met spreken*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

12. De gesprekjes stimuleren tot oefenen met luisteren*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

13. Een student kan goed testen met de robot of hij/zij de teksten beheerst *

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

14. De robot zag er serieus uit *

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

15. De robot kwam serieus over*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

16. De robot past bij de doelgroep*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

17. Ik voelde mij comfortabel tijdens het gesprek*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

18. De robot gaf genoeg feedback*

Markeer slechts één ovaal.

	1	2	3	4	5	
Helemaal mee oneens	◯	◯	◯	◯	◯	Helemaal mee eens

E

Mapping of Co-Design Results

Down below, a part of the map with the results of the co-design sessions with the DM experts is presented. This map served partially as a foundation for the second set of design requirements. The complete map can be viewed online.

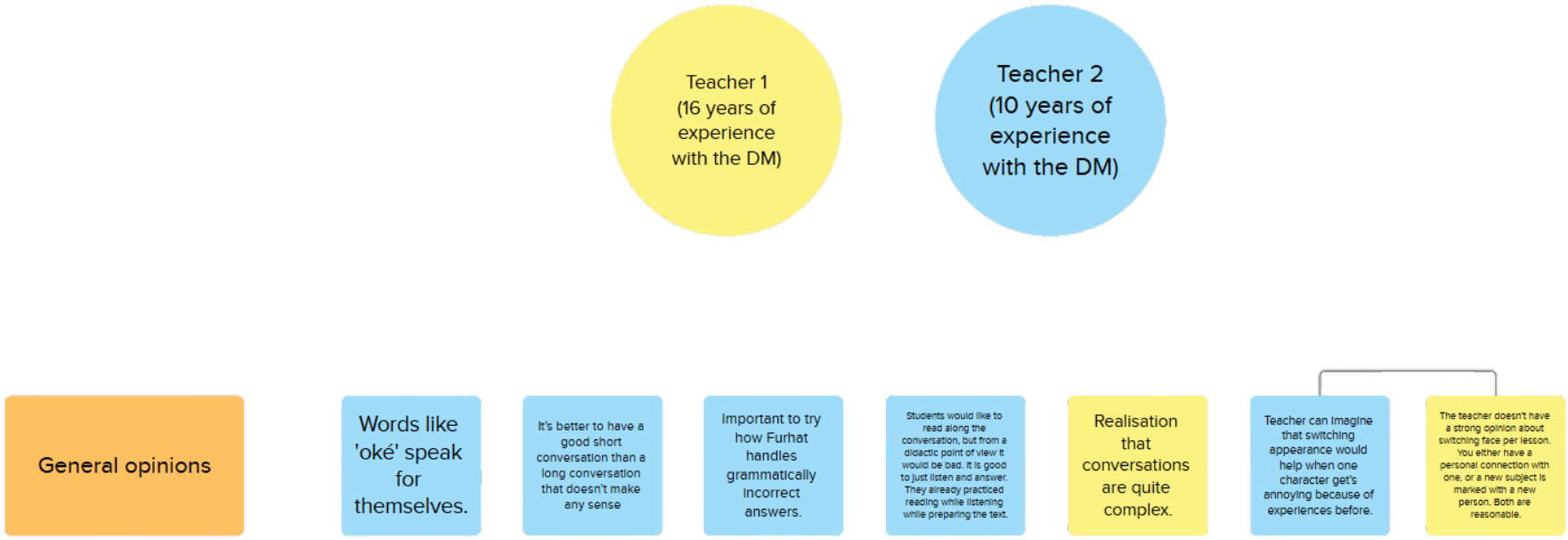


Figure E.1: *The map for thematic analysis of the co-design sessions.*

F

Questionnaires - User Tests

This appendix includes the pre- and post-questionnaires that the users had to complete on a laptop.

Pre-Questionnaire

* Required

Demographics

1.What is your age? *

- ◯ Under 18
- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45+

2.What is your gender? *

- ◯ Man
- ◯ Woman
- ◯ Other
- ◯ Prefer not to say

3.Which country are you from? *

4.What is your mother tongue? *

5.Which languages, besides English (and beginner's level Dutch), do you speak? *

6.For which of the following lessons of the Delftse Methode (DM) beginner's course (Green Book) were you present in class? *

- [ ] Les 20 (De mooiste reis)
- [ ] Les 21 (Een dagje uit)
- [ ] I did not do the DM beginners' course
- [ ] I do not remember
- [ ] Other

7.At which DM lesson of the Green Book are you now? *

- ◯ I do not follow the Green Book course currently
- ◯ Other

8.Which type(s) of DM course have you followed or are you following? *

- [ ] Beginners intensive (Green Book)
- [ ] Elementary 1 (Green Book)
- [ ] Elementary 2 (Green Book)
- [ ] Intermediate intensive (Second Round Book)
- [ ] Intermediate 1 (Second Round Book)
- [ ] Intermediate 2 (Second Round Book)
- [ ] Advanced intensive (Third Round Book)
- [ ] Advanced 1 (Third Round Book)
- [ ] Advanced 2 (Third Round Book)
- [ ] Advanced +
- [ ] Other

9.Do you study, a PhD, or work? *

- ◯ Bachelor University student
- ◯ Master University student
- ◯ PhD
- ◯ EngD
- ◯ Employed
- ◯ Unemployed
- ◯ Other

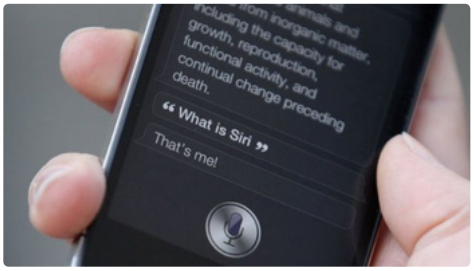
Experience with Conversational Agents

10. Which of the following chatbots/conversational agents/assistants do you have experience with? *

☐ ChatGPT

☐ Gemini (Bard)

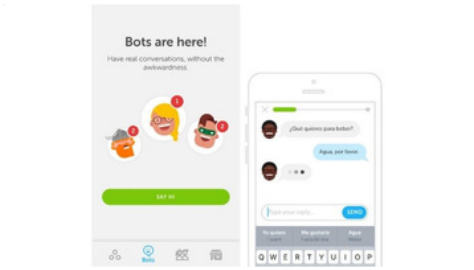
☐ Microsoft Copilot



☐ Siri

☐ Alexa

☐ Furhat



☐ Duolingo chatbot

☐ None of the above

☐ Other

11. How would you rate your experience with *chatbots*? **Note**: So, bots you can chat with through **typing**. *

☆ ☆ ☆ ☆ ☆

12. How would you describe your experience with *chatbots*? *

13. How would you rate your experience with *spoken digital assistants* and/or *conversational agents*? **Note**: So, bots you can interact with through **speech**. *

☆ ☆ ☆ ☆ ☆

14. How would you describe your experience with *spoken digital assistants* / *conversational agents*? *

15. Indicate for each item how often you would choose to talk in that situation or under that condition during a class of the Delft method. *

	Never	Almost never	Sometimes	Often	Very often
When the class is engaged in an open discussion.	○	○	○	○	○
When the class size is small.	○	○	○	○	○
When the teacher asks for a response from the class.	○	○	○	○	○
When the topic is interesting.	○	○	○	○	○
When my views differ from my classmates' views.	○	○	○	○	○
When I am sitting far away from the teacher.	○	○	○	○	○
When I followed all steps of the Delft method to study the text of the current class.	○	○	○	○	○
When almost the whole class is actively participating.	○	○	○	○	○
When the class is engaged in a heated debate.	○	○	○	○	○
When I am comfortable with the subject matter.	○	○	○	○	○
When an assignment is being discussed.	○	○	○	○	○
When no one else is talking.	○	○	○	○	○
When I am sitting close to the teacher.	○	○	○	○	○
When my views differ from the teacher's views.	○	○	○	○	○
When I am angry about a topic.	○	○	○	○	○
When I know the correct answer.	○	○	○	○	○
When I can really help clarify the discussion.	○	○	○	○	○
When I dislike my classmates.	○	○	○	○	○

Post-Questionnaire

* Required

Rating Situations and Conditions during Class

You will again get to read situations that could happen or conditions that could be met during a typical Delft method class.

Imagine that you could use the Furhat robot to practice a conversation for each lesson in the Delft method, before coming to class. You would thus have frequent contact with the robot and it could do a conversation about every lesson. How would that influence your willingness to talk?

1. Indicate for each item how often you would choose to talk in that situation or under that condition during a class of the Delft method, imagining that you can prepare for class with the robot. *

	Never	Almost never	Sometimes	Often	Very often
When the class is engaged in an open discussion.	◯	◯	◯	◯	◯
When the class size is small.	◯	◯	◯	◯	◯
When the teacher asks for a response from the class.	◯	◯	◯	◯	◯
When the topic is interesting.	◯	◯	◯	◯	◯
When my views differ from my classmates' views.	◯	◯	◯	◯	◯
When I am sitting far away from the teacher.	◯	◯	◯	◯	◯
When I followed all steps of the Delft method to study the text of the current class.	◯	◯	◯	◯	◯
When almost the whole class is actively participating.	◯	◯	◯	◯	◯
When the class is engaged in a heated debate.	◯	◯	◯	◯	◯
When I am comfortable with the subject matter.	◯	◯	◯	◯	◯
When an assignment is being discussed.	◯	◯	◯	◯	◯
When no one else is talking.	◯	◯	◯	◯	◯
When I am sitting close to the teacher.	◯	◯	◯	◯	◯
When my views differ from the teacher's views.	◯	◯	◯	◯	◯
When I am angry about a topic.	◯	◯	◯	◯	◯
When I know the correct answer.	◯	◯	◯	◯	◯
When I can really help clarify the discussion.	◯	◯	◯	◯	◯
When I dislike my classmates.	◯	◯	◯	◯	◯

Robot Statement Ratings

The statements below regard the different aspects of the robots and the conversations. You are supposed to rate each statement from 1 meaning totally disagree and 5 meaning totally agree.

2. Question

	1. Totally disagree	**2.** Somewhat disagree	**3.** Neutral	**4.** Somewhat agree	**5.** Totally agree
The robot hears me well	○	○	○	○	○
The robot understands me well	○	○	○	○	○
The conversations fit the lessons well	○	○	○	○	○
The words used by the robot correspond with the correct vocabulary of the lessons	○	○	○	○	○
The conversations went logical	○	○	○	○	○
The conversations were interesting	○	○	○	○	○
The conversations were of sufficient length (time-wise)	○	○	○	○	○
The conversations encourage me to practice speaking	○	○	○	○	○
The conversations encourage me to practice listening	○	○	○	○	○
I can test well if I know the texts from the book properly	○	○	○	○	○
The robot looked serious	○	○	○	○	○
The robot fits the target group	○	○	○	○	○
I felt comfortable during the conversation	○	○	○	○	○
The robot gave enough feedback	○	○	○	○	○

G

Code

```
22 val LikesTravelling = state(parent = RepetitionStrat) {
23     var notUnderstandingCounter = 0
24
25     onEntry {
26         val utterances = listOf("Hou jij van reizen?", "Vind jij reizen leuk?", "Reis
           ↪  jij graag?")
27
28         val selectedUtterance = findValidUtterance(utterances, "20")
29
30         if (selectedUtterance != null) {
31             furhat.ask(selectedUtterance)
32         }
33     }
34
35     onResponse<Negative> {
36         furhat.say("Oh, echt niet?")
37         goto(LongestTrip)
38     }
39
40     onResponse<Positive> {
41         furhat.say("Mooi! Ik vind reizen ook leuk.")
42         goto(LongestTrip)
43     }
44
45     onResponse<NotUnderstanding> {
46         notUnderstandingCounter++
47         if (notUnderstandingCounter < 2) {
48             reentry()
49         } else {
50             goto(LongestTrip)
51         }
52
53     }
54
55     onResponse<Unsure> {
56         furhat.say("Ik snap het. Ik vind reizen zeker leuk.")
57         goto(LongestTrip)
58     }
59
60 }
```

Listing G.1: *Example state Prototype I.*

```
class EndConversation: Intent() {
    override fun getExamples(lang: Language): List<String> {
        return listOf(
            "ik wil stoppen",
            "ik ben klaar",
            "zullen we stoppen",
            "eindigen",
            "stoppen",
            "klaar",
            "hou op",
            "het gesprek is voorbij",
            "ik wil niet meer praten",
            "het gesprek is over",
            "laten we het hierbij laten",
            "genoeg nu",
            "dat is genoeg",
            "stop maar",
            "ik ben klaar met dit gesprek",
            "ik wil afsluiten",
            "we stoppen hier",
            "ik wil het niet meer voortzetten",
            "tijd om te stoppen",
            "dit is het einde van het gesprek",
            "ik ben er klaar mee"
```

Listing G.2: *Example intent from Prototype I (EndConversation).*

```
val StateName: State = state(parent = LesXY) {
    var notUnderstandingCounter = 0
    val previousUtterances = mutableListOf<String>()

    onEntry{
        val utterances = listOf("Optie 1", "Optie 2", "etc.")
        val availableUtterances = utterances - previousUtterances
        val selectedUtterance = findValidUtterance(availableUtterances, "XY")

        if (selectedUtterance != null) {
            previousUtterances.add(selectedUtterance)
            furhat.ask(selectedUtterance)
        } else {
            if (previousUtterances.isNotEmpty()) {
                furhat.ask(previousUtterances.last()) // Repeat the last utterance
            } else {
                furhat.say("We gaan naar de volgende vraag.")
                goto(NextState)           }
        }
    }

    onResponse<EndConversation> {
        goto(Stopping)        }

    onResponse<RepetitionReq> {
         println("Previous utterances: $previousUtterances")
         if (previousUtterances.isNotEmpty()) {
            furhat.ask(previousUtterances.last()) // Repeat the last utterance
         } else {
             furhat.say("We gaan naar de volgende vraag.")
             goto(NextState) // Fallback in case no utterance is stored}
    }

    onResponse<NotUnderstanding> {
       notUnderstandingCounter++
       if (notUnderstandingCounter < 3) {
           reentry()
       } else {
           furhat.say("We gaan naar de volgende vraag.")
           goto(NextState)       }
    }

   onResponse<ClosedQuestion> {
       furhat.say("")
       goto(NextState)   }

   onResponse<OpenQuestion> {
       val inputSentence = suffLengthCheck(it.text, 3)
           if (!inputSentence.first) {
                furhat.say("Kan je je zin wat langer maken?")
                reentry()}
       furhat.say("")
       goto(NextState) }
```

Listing G.3: *State template.*

```
411             "kunnen we stoppen",
412             "kunt u stoppen",
413             "kun je stoppen",
414             "zullen we hiermee ophouden",
415             "het is genoeg geweest",
416             "ik heb geen zin meer om te praten",
417             "ik heb hier genoeg van",
418             "dit gesprek hoeft niet verder te gaan",
419             "ik vind het wel goed zo",
420             "het is tijd om af te sluiten",
421             "ik wil het gesprek beëindigen",
422             "we kunnen nu stoppen",
423             "ik ben klaar met dit onderwerp",
424             "ik wil hier niet meer over praten",
425             "zullen we het beëindigen",
426             "mag ik nu stoppen",
427             "kunt u dit gesprek afronden",
428             "laten we hier een punt achter zetten",
429             "het is klaar",
430             "voor mij hoeft het niet meer",
431             "we kunnen hiermee stoppen",
432             "ik vind het gesprek klaar",
433             "ik wil niet verder praten",
434             "u mag stoppen",
435             "stop alstublieft",
436             "kun je ophouden",
437             "kunt u ophouden",
438             "zullen we stoppen met praten",
439             "ik stel voor om nu te stoppen",
440             "ik zie geen reden om door te gaan",
441             "dit lijkt een goed moment om te stoppen",
442             "ik beëindig dit gesprek",
443             "het is tijd om te stoppen",
444             "ik zou graag willen afsluiten",
445             "dit gesprek kan stoppen",
446             "ik wil niets meer zeggen",
447             "er is niets meer te bespreken",
448             "ik wil dat we stoppen",
449             "ik heb hier genoeg van",
450             "ik wil afronden",
451             "u kunt stoppen",
452             "ik hoef niets meer te horen"
```

Listing G.4: *The added sentences for the EndConversation intent from Prototype II.*

```
// Adapted word list of the Delft method
// Each word is mapped to the first lesson they appear in
val wordMap = mapOf(
    // Les 1
    "les" to "1", "de" to "1", "hoe" to "1", "heet" to "1", "heten" to "1", "je" to "1", "hallo" to "1", "ik" to "1",
    "ben" to "1", "mijn" to "1", "naam" to "1", "is" to "1", "docent" to "1", "wie" to "1", "jij" to "1", "wat" to "1",
    "dag" to "1", "mevrouw" to "1", "uit" to "1", "welk" to "1", "land" to "1", "het" to "1", "kom" to "1",
    "komen" to "1", "zegt" to "1", "zeggen" to "1", "u" to "1", "waar" to "1", "vandaan" to "1", "Frankrijk" to "1",
    "en" to "1", "meneer" to "1", "uw" to "1", "komt" to "1", "China" to "1", "woont" to "1", "wonen" to "1",
    "in" to "1", "welke" to "1", "stad" to "1", "woon" to "1", "nu" to "1", "Den Haag" to "1", "straat" to "1",
    "centrum" to "1", "op" to "1", "nummer" to "1", "telefoonnummer" to "1", "ook" to "1", "nee" to "1", "hij" to "1",
    "dichtbij" to "1", "haar" to "1", "achternaam" to "1", "zij" to "1", "Nederland" to "1", "niet" to "1",
    // Les 2
    "voornaam" to "2", "adres" to "2", "postcode" to "2", "plaats" to "2", "telefoon" to "2", "mobiel" to "2",
    "werk" to "2", "Engeland" to "2", "heb" to "2", "hebben" to "2", "Engelse" to "2", "Nederlandse" to "2",
    "Nederlands" to "2", "nationaliteit" to "2", "lang" to "2", "hier" to "2", "zes" to "2", "jaar" to "2",
    "leeftijd" to "2", "wanneer" to "2", "geboren" to "2", "maart" to "2", "alleen" to "2", "bij" to "2", "geen" to "2",
    "vrienden" to "2", "vriend" to "2", "of" to "2", "familie" to "2", "nan" to "2", "getrouwd" to "2",
    "trouwen" to "2", "met" to "2", "een" to "2", "Nederlander" to "2", "zijn" to "2", "we" to "2", "jullie" to "2",
    "huis" to "2", "buiten" to "2", "kinderen" to "2", "kind" to "2", "ja" to "2", "twee" to "2", "jongen" to "2",
    "van" to "2", "drie" to "2", "meisje" to "2", "vijf" to "2", "maanden" to "2", "maand" to "2", "ze" to "2",
    "man" to "2", "zoon" to "2", "dochter" to "2", "hun" to "2", "moeder" to "2", "vader" to "2", "heeft" to "2",
```

Listing G.5: *Word map.*