Learning Icelandic in *Virtual Reykjavik*:

Simulating real-life conversations with embodied conversational agents using multimodal clarification requests

Branislav Bédi

Ritgerð lögð fram til doktorsprófs
2020

HÁSKÓLI ÍSLANDS
HUGVÍSINDASVIÐ
MÁLA- OG MENNINGARDEILD

Learning Icelandic in *Virtual Reykjavik*:
Simulating real-life conversartions with embodied conversational agents using multimodal clarification requests

Branislav Bédi

Ritgerð lögð fram til doktorsprófs

Háskóli Íslands
Hugvísindasvið
Mála- og Menningardeild
Nóvember 2020

Mála- og menningardeild Háskóla Íslands

hefur metið ritgerð þessa hæfa til varnar
við doktorspróf í annarsmálsfræðum

Reykjavík, 26. október 2020

Birna Arnbörnsdóttir

deildarforseti

Oddný Sverrisdóttir

varadeildarforseti

Doktorsnefnd:

Birna Arnbjörnsdóttir, leiðbeinandi

Ana Paiva

Hannes Högni Vilhjálmsson

Learning Icelandic In Virtual Reykjavik*: Simulating real-life conversations with embodied conversational agents using multimodal clarification requests*

Reykjavík 2020

ISBN 978-9935-9491-7-2
ORCID 0000-0001-7637-8737

Útdráttur

Þessi doktorsritgerð er hluti af verkefninu *Icelandic Language and Culture Training in Virtual Reykjavik*, þrívíddartölvuleik sem gerir þeim sem eru að læra íslensku sem annað mál kleift að æfa tal og hlustun. Markmið verkefnisins var að búa til tölvuleik með sýndarspjallverum (e. *embodied conversational agents*) sem byggju yfir raunsærri fjölþættri hegðun, með það langtímamarkmið að styðja við hagnýta kennslu á íslensku máli og menningu þar sem mál úr raunverulegum samskiptum er notað. Markmið doktorsverkefnisins beindist að því að rannsaka raunveruleg yrt og óyrt atriði í skýringarbeiðnum meðal Íslendinga (e. *clarification requests*, *CRs*). Lögð voru til sex fjölþætt líkön af skýringarbeiðnum sem áttu að stuðla að raunhæfari samspili manna og sýndarspjallvera í *Virtual Reykjavik*. Þróun doktorsritgerðar fór fram í þremur lotum. Fyrst var gerð stutt könnun til að komast að því hvaða væntingar notendur hefðu til *Virtual Reykjavik*-þrívíddarforritsins. Nemendurnir sögðust eiga í erfiðleikum með að æfa sig í að tala íslensku við þá sem hafa íslensku að móðurmáli og kynnu því að meta að fá sýndarnámsumhverfi til að æfa sig í tali. Kennslufræðilegur grunnur *Virtual Reykjavik* tekur mið af samskiptaaðferðum, námi á grundvelli verkefna og leikja og fjölþættum og einstaklingsmiðuðum aðferðum í tungumálanámi. Sýndarspjallverur *Virtual Reykjavik* búa yfir fjölþættri hegðun sem er í samræmi við íslenska menningu. Með því að taka þátt í leiknum komast notendur í tæri við íslenskt mál og menningu í sýndarnámsveruleika áður en þeir eiga samskipti við Íslendinga.

Meginviðfang rannsóknarinnar var samskiptaþátturinn skýringarbeiðni (CR) en nauðsynlegt var að afmarka rannsóknina við einn samskiptaþátt svo unnt væri að nota fjölþætta greiningu sem dugði til að forrita sýndarverurnar. Skýringarbeiðni er ein algengasta tegund segða í samtölum (Purver, 2004). Hún hjálpar til við að skýra það sem áður hefur verið sagt en sem viðmælandi hefur af einhverjum sökum ekki skilið og stuðlar þannig að góðu samtalsflæði. Af þessum sökum eru skýringarbeiðnir mjög mikilvægar til þess að ná fram raunsæjum samskiptum milli notanda og sýndarspjallveru í kerfum eins og okkar sem sameina sjálfvirka talgreiningu og samtöl sem skipulögð eru fyrir fram. Í næstu lotu rannsóknarinnar var málgögnum safnað til þess að greina yrta og óyrta þætti í mismunandi tegundum af skýringarbeiðnum. Vegna þess hversu flókið talmál er og fjölbreytileg samtöl geta verið var aðeins safnað samtölum þar sem ókunnugir spurðu til vegar í miðbæ Reykjavíkur. Þetta endurspeglaðist svo í þeim verkefnum sem nemendur

þyrftu að leysa í *Virtual Reykjavik*. Þar spyrja þeir sýndarspjallverur til vegar í miðbæ Reykjavíkur og verurnar nota skýringaraðferðir til að vísa til vegar á sem raunsæjastan hátt. Þó ber ekki að líta svo á að þetta sé tæmandi rannsókn á eðli skýringarbeiðna heldur fjölþætt lýsing á skýringarbeiðnum, notkun þeirra í sérstökum samræðuaðstæðum í leiknum og beitingu þeirra til að líkja eftir mannlegri hegðun.

Sex mismunandi fjölþættar skýringarbeiðnategundir voru búnar til á grundvelli gagnagrunns með myndbandsupptökum af raunverulegum samtölum milli fólks með íslensku að móðurmáli og fólks sem ekki hefur íslensku að móðurmáli. Þetta voru í heild 165 upptökur, 1.59.02 klst. á lengd, 108 pör fólks þar sem annar aðilinn hefur íslensku að móðurmáli en hinn ekki og 57 pör þar sem báðir aðilar hafa íslensku að móðurmáli, karlmenn og konur. Aldur þeirra sem höfðu íslensku að móðurmáli var á bilinu 18–70 ár og meðalaldurinn u.þ.b. 35 ár en þeir sem ekki höfðu íslensku að móðurmáli voru á aldrinum 20–40 ára og meðalaldur þar u.þ.b. 30 ár. Úr þessum gagnagrunni var búinn til fjölþættur stofn skýringarbeiðna sem samanstóð af yrtum og óyrtum gögnum fyrir hverja tegund af skýringarbeiðni. Myndbandsupptökur voru greindar með ELAN merkingar- og skýringapakkanum. Í hverri greiningu var fjölþættum gögnum lýst. Fjölþættri nálgun við tungumál og fjölþættri greiningu á samskiptum var beitt til að greina yrta og óyrta þætti skýringarbeiðna. Vegna takmarka á umfangi rannsóknarinnar voru aðeins tvær gerðir beiðna notaðar, úrfelling og innskotsaðferð.

Að lokum var framkvæmd notendakönnun til að komast að því hvernig nemendur skynjuðu fjölþætta hegðun sýndarspjallveranna í leiknum og hvort þeir tækju eftir þessum tveimur tegundum skýringarbeiðna í honum. Nemendum þótti innskotsaðferðin vera eðlilegust þótt þeim hefði fundist henni stundum vera beitt dálítið ruddalega eða hún verið notuð of mikið af sýndarspjallverunum. Það hversu spjallverurnar notuðu mikið skýringarbeiðnirnar var ekki mælt þar sem einblínt var á nemendur sem notendur í þessari frumútgáfu leiksins. Könnunin leiddi í ljós fjölda möguleika til að betrumbæta fjölþætta hegðun spjallveranna í framtíðarútgáfum leiksins. Sérstaklega bentu notendur á að ákveðin svipbrigði og að spjallverurnar gætu ekki brosað gerði það að verkum að þær virkuðu „óhugnanlegar“.

Í stuttu máli eru færð rök fyrir því í ritgerðinni að þrívíddartölvuleikir nýtist vel til að kenna íslenska tungu og menningu, með sérstakri áherslu á að æfa talmálsfærni. Fjallað er um og stutt með kennslufræðilegum kenningum hvernig bæta megi námsupplifun og kalla fram alvörusamskipti í sýndarveruleika með raunsærri og fjölþættri hegðun spjallvera. Skoðaðar voru sex skýringaraðferðir sem fólk með íslensku að móðurmáli notaði til að vísa til vegar, annars vegar af fólki með íslensku að móðurmáli og hins vegar þeim sem ekki hafa íslensku að móðurmáli. Í ritgerðinni er einnig bent á hugsanlegar nýjar rannsóknir í sambandi við skýringarbeiðnir og *Virtual Reykjavik*. Skoða mætti frekar fjölþættar skýringarbeiðnir í samtölum við aðrar aðstæður og í öðrum tungumálum. Slíkt myndi gagnast við að betrumbæta þær skýringarbeiðnir sem sýndarspjallverur í *Virtual Reykjavik* nota. Ágætis byrjun á áframhaldandi vinnu væri að framkvæma nýja könnun með fullkomnari leiðbeiningum, námsefni og stoðbúnaði, talgreinikerfi sem virkar á allan hátt í *Virtual Reykjavik* og með sýndarspjallverum sem byggju yfir fleiri eiginleikum, gætu t.d. brosað.

Þakkir

Leiðbeinandi minn, Birna Arnbjörnsdóttir, veitti mér ómetanlega aðstoð við rannsóknina og ritgerðarskrifin. Hannes Högni Vilhjálmsson leiðbeindi mér gegnum verkferlið í *Virtual Reykjavik*-verkefninu, rannsóknina á fjölþættri gagnagreiningu, auk þess að veita mér ótal ráðleggingar. Ana Paiva var óspör á athugasemdir um efnið og veitti aðstoð við framkvæmd rannsóknarinnar. Gott samstarf og aðstoð allra í *Virtual Reykjavik*-teyminu hélt mér gangandi allt ferlið. Teikningar af röð skýringarbeiðna eru eftir Loga Geir Þorláksson. Ég vil þakka Guðrúnu Nordal, forstöðumanni Stofnunar Árna Magnússonar í íslenskum fræðum, fyrir stuðninginn á lokametrum ritgerðarskrifanna. Ég stend í þakkarskuld við alla þátttakendur sem buðu sig fram, samstarfsfélaga og vini sem lögðu hönd á plóginn við gagnasöfnun í rannsókninni sem hér liggur til grundvallar. Að lokum ber mér að þakka Háskóla Íslands, Háskólanum í Reykjavík, samstarfsfólki við Center for Analysis and Design of Intelligent Agents (CADIA) og RANNÍS fyrir stuðning.

Abstract

This thesis forms part of the project Icelandic Language and Culture Training in Virtual Reykjavik, a 3D computer game that enables learners of Icelandic to practise oral language and listening. The aim of the project was to build a computer game populated with embodied conversational agents (ECAs) endowed with realistic multimodal behaviour, with a long-term goal of supporting authentic teaching of Icelandic language and culture. The part of the project reported in this thesis focused on examining human verbal and non-verbal features in clarification requests (CRs). Six multimodal CR models were suggested for implementation, with the intention of promoting a more realistic human-agent interaction in *Virtual Reykjavik*. The research took place in three phases. First, a small survey was carried out, eliciting learners' expectations from *Virtual Reykjavik.* It informed about learners' expectations of a 3D application. Learners reported difficulties in practising spoken Icelandic with native speakers in real life and for this reason said they would appreciate a virtual learning environment for practising oral language. The pedagogical foundation of *Virtual Reykjavik* considers the communicative approach in language instruction, task- and game-based learning, and multimodal and individual language learning approaches. *Virtual Reykjavik* was populated with ECAs endowed with multimodal behaviour that is authentic to Icelandic culture. Engaging in the game provided learners with an opportunity to experience Icelandic language as it is spoken in the target culture but in a virtual learning environment, and prior to engaging with speakers in the real world.

The communicative function CR was chosen as the main object of multimodal analysis, in order to narrow down the focus to a specific topic in natural language research. CR is one of the most commonly used utterance-types in spoken conversations (Purver, 2004); it helps to clarify what has previously been said but for whatever reason not understood by the recipient, and as such facilitates smooth conversational flow. For these reasons, CR is very important in achieving a realistic human-agent interaction in systems, like ours, which combine automatic speech recognition and pre-planned dialogues. In this second phase, natural language data was collected in order to analyse the verbal and non-verbal features in various types of CRs. Due to the complexity of spoken language and a wide range of possible conversational scenarios, data were collected only during first encounters asking for directions to a location in central Reykjavik. This in turn reflected

the same task learners would need to do in *Virtual Reykjavik* - they would ask agents for directions in central *Virtual Reykjavik* and the agents would use clarification strategies in an authentic way. It should, however, not be seen as an exhaustive treatise about the nature of CRs but rather as a multimodal description of CRs, their use in a particular conversational scenario in the game, and their application to the development of human-like behaviour. Based on a database of video recordings of real-life conversations between native and non-native speakers of Icelandic, six different multimodal CR types were characterised. (165 recordings with total recorded time 1 hour, 59 minutes and 2 seconds; 108 native-non-native speaker pairs and 57 native-native speaker pairs, men and women; ages of native speakers between 18-70 with average age approximately 35 years, and ages of non-native speakers between 20-40 with average age approximately 30 years). Out of this database, a multimodal corpus of CRs was created, consisting of verbal and non-verbal data for each type of CR. Video recordings were analysed using the ELAN tagging and annotation package. Each analysis consisted of a description of multimodal data. The multimodal approach to language and the multimodal interaction analysis were used to analyse the verbal and non-verbal features of CRs. Due to resource constraints, only two types, the Ellipsis and the Fragment (Interjection Strategy), were implemented.

Finally, a user response study was conducted in order to find out how learners perceived multimodal behaviour of ECAs in the game, and whether surveyed learners noticed the two implemented CRs. Learners perceived the CR Fragment (Interjection Strategy) as the most natural, despite its being perceived as slightly rude or used too frequently by the ECAs. The frequency of use of CRs by the ECAs was not measured, since the focus was on learners as users of this game prototype. The study revealed many possibilities for improving the multimodal behaviour of ECAs which could be implemented in future versions. In particular, certain facial expressions, and their lack of ability to smile, were commonly perceived by learners as “creepy”.

In summary, this thesis presents the rationale for building a 3D computer game for teaching Icelandic language and culture, with a focus on practising oral language skills. It presents pedagogical background for including authentic features into the multimodal behaviour of ECAs in a computer game to achieve a more realistic human-agent interaction, and thus to contribute to an improved learning experience in an online virtual learning environment. Six clarification strategies used by native speakers of Icelandic were observed when they were approached by other native and non-native speakers asking for directions. The thesis also outlines points for future work on CRs and *Virtual Reykjavik*.

Exploration of multimodal CRs in other conversational settings and languages would be useful for further improving ECA CRs used in *Virtual Reykjavik*. A good starting point for a continuation would be to conduct a new study with more complete instructions, learning materials and scaffolding, a fully functioning speech recognition system in *Virtual Reykjavik*, and ECAs endowed with additional features including smiling.

Acknowledgements

My supervisor, Prof. Birna Arnbjörnsdóttir, provided invaluable support and input on this thesis. Prof. Hannes Högni Vilhjálmsson guided me throughout the process of working on the *Virtual Reykjavik* project, the research work on multimodal data analysis, as well as providing much advice. Prof. Ana Paiva was generous in her comments on the content and the practical part of this thesis. I was sustained throughout by the collaborative spirit and succour of the whole *Virtual Reykjavik* team. The drawings of CR sequences are the work of Logi Geir Þorláksson. Gratitude goes to Guðrún Nordal, Director of the Árni Magnússon Institute for Icelandic Studies, for her support during the final stage of writing this thesis. I am indebted to all the volunteering participants, colleagues and friends who assisted with collecting data in the studies presented. Finally, I must thank the University of Iceland, Reykjavik University, my colleagues at the Center for Analysis and Design of Intelligent Agents (CADIA), and the Icelandic Research Fund RANNÍS for their support.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Introduction

This is an interdisciplinary thesis which has its home in the field of humanities, particularly in applied linguistics and second language learning, and also in the field of computer science. Other fields such as linguistics, applied linguistics, human-computer interaction, and natural language processing (NLP) are also explored herein. These domains assist to understand the design process aimed at realistic behaviour in ECAs in this game, towards enabling a more authentic language- and culture-learning experience to Icelandic learners. This thesis is part of a larger project, Icelandic Language and Culture Training in Virtual Reykjavik. The goal of the project is to develop a 3D computer game for learning Icelandic language and culture in *Virtual Reykjavik*. Specifically, the goal is to create a serious computer game in a 3D virtual learning environment populated with virtual characters who can speak and act like real people, and help learners of Icelandic practise spoken language. The purpose of this doctoral thesis was to support this development by collecting and analysing multimodal data from real-life interactions in first encounters between humans using clarification strategies and proposing theoretical models for multimodal clarification requests that can be implemented in *Virtual Reykjavik* into the multimodal behaviour of virtual characters. The team of programmers had already developed and implemented different interactional function categories in Functional Markup Language (FML). These interactional function categories, e.g. initiate and close a conversation, speech act, turn-taking, and grounding, and their different types are based on Cafaro's et al. (2014) work, which was also part of the larger project's work. The role of the main study presented in this thesis was to provide a description of behaviour that supports or carries one of the types of the communicative functions, the *clarification-request* that is part of the interactional function grounding. Data from real-life interactions informs the design of realistic agents in the game and thus contributes to a more realistic virtual learning experience for users practising oral language skills. This study also presents the results of two auxiliary studies which informed the *Virtual Reykjavik* project. The purpose of the first of the two auxiliary studies was to gauge learners' views about using 3D virtual learning environments for oral practice. The purpose of the second auxiliary study was to measure learners' views and experiences when playing the game with virtual characters endowed with multimodal features in two clarification strategies identified in the main study. The studies are described briefly below.

The first auxiliary study is a brief online survey of beginner and intermediate learners of Icelandic at the University of Iceland. The aim here was to gain an insight into the learners' needs for such a game and to a problem as to why it is important to practise spoken Icelandic in a VLE in Iceland. Although this survey was conducted approximately one year after the commencement of the larger project, it only provides, though in retrospect for the *Virtual Reykjavik* project team, motivation for creating a much needed virtual learning space populated with 3D embodied conversational agents (ECAs) helping learners to practise spoken language skills in a game-like environment, and information about learners' expectations from a 3D game for learning Icelandic that the project team can use in the future when designing learning materials and game scenarios for the computer game. This auxiliary study is conducted in a form of a survey and its purpose is to determine and compare the language skills learners want to practise in both the traditional language course and the 3D computer game *Virtual Reykjavik*, what elements such a game should contain to make it more enjoyable playing it, and what advantages and disadvantages of learning Icelandic language there are in such a game. In addition, a question about how learners feel about using Icelandic in a face-to-face interaction with local native people should give a hint that practicing spoken language skills in a safe interim space, i.e. where learners can make mistakes and use feedback to correct their own language output while practicing spoken interactions with virtual agents, would be practical even for learners living in Iceland. Although the questions in the survey do not include any direct questions about multimodal behaviour of agents using CRs, or questions about multimodality in computer games, they do, however, include open-ended questions about a 3D virtual environment that ask about learners' expectations which elements to include in the game and the design of virtual characters. The question about what kind of multimodal features should be included in CRs that ECAs would use in the game, is part of the main study. Peterson (2010a) suggests that computer games can provide an optimal language learning environment for negotiation of meaning, i.e. reaching a clear understanding of each other, with other peers which results in the production of the target language output, participation, enjoyment, and a reduction in anxiety or improved self-confidence (p. 74). In the view of the latter two points, the brief online survey reflected on learners' feelings when speaking Icelandic face-to-face to native speakers, i.e., they may feel nervous or insecure. Such applications could hasten the learning process and enable access to an online learning space. Practising spoken language skills in a VLE may also be very useful for those living in Iceland, due to the fact that local people are widely exposed to English and often switch to English when speaking to learners of Icelandic. This is similarly supported by the study of Arnbjörnsdóttir (2011), who reported that due to the wide exposure to English in Iceland, locals often use English in conversations with others in Iceland, including non-native Icelandic speakers. Furthermore, Theodórsdóttir and Eskildsen (2011) argued that due to the geographical location of Iceland between mainland Europe and Northern America, and having close relations with the United States because of a US Navy base operating until 2006[1], English may be even more salient in Iceland that in other countries. Nonetheless, Hult (2003, p. 44) reported that English also prevails in Sweden and was reported to be widely used by native Swedish speakers in educational, public, commercial and government settings, in both written and spoken interactions. For the above reasons, Icelandic spoken language practice may become difficult to conduct outside of traditional classes, as the tendency of locals is to switch into English when speaking to foreign learners struggling with spoken Icelandic (Bédi et al., 2017, p. 79). The survey helped to understand the situation learners of Icelandic experience in Iceland; they feel negative or even embarrassed when speaking Icelandic face-to-face with native speakers, and therefore might benefit from an interim learning space where they can practise oral language (Morton et al., 2012).

The second and main phase of this thesis consists of a study on the multimodal features of natural language that should help ECAs to speak and act in an authentic way. It investigates multimodal features in clarification requests (CRs). CRs are an important communicative function that occur frequently in real life conversations. It is the most commonly used (88%) open repair initiator in the Icelandic corpus of linguistic data collected from conversations with friends and family, analysed and presented in the article by Gísladóttir (2015), with the most frequently used type, an interjection *Ha? (Huh?)* (Gísladóttir, 2015, p. 314). In English corpora, it is also one of the most frequent utterances, the purpose of which is to maintain a smooth flow of a conversation (Purver, 2004). Here, the aim was to include the open repair initiator CR into the conversational architecture of ECAs for two main reasons: 1) to teach learners how native speakers ask for a clarification, which may be different to other languages and cultures, and 2) a practical reason to help maintain a smooth flow of a conversation between agents and learners, when technical problems or

[1] https://www.foreignaffairs.com/articles/united-states/2016-02-24/return-keflavik-station

the learner's unclear pronunciation, or incorrect use of words, prevent agents from understanding them. This communicative function may, however, be executed differently in different cultures and languages. When humans learn a different language, they often transfer their own native language (L1) features into that of an L2, which may or not be appropriate. Cultural behaviour can also be transferred in this way (Mapson, 2015, p. 164). Authentic L2 input, or exposure to language of native speakers is, thus, very important to the L2 language learners because it helps them to understand pragmatic behaviours that are problematic to learn in a formal classroom setting. Therefore, when modelling the conversational behaviour of ECAs in Icelandic, authentic features used by native speakers of Icelandic must be used. In the game, the CR function allows learners to repeat themselves without interrupting the conversation because the virtual characters will perform the function in a realistic manner based on observations of real-life interactions. In this way, the learners will become familiar with the behaviour of native speakers and learn to understand its various features that may be different to their own language and culture. For this, a multimodal approach to language is used that helps to understand how language is learned and used, what multimodal features include (particularly utterances), and therefore what features are necessary for learners to learn and to be able to practise authentic language use. Here, a study was conducted to gather data on authentic interactions involving CRs. The aim of the main study in this thesis was to establish the verbal and non-verbal features in such utterances that would help ECAs in the game act and speak more authentically. In order to propose models for multimodal CRs that ECAs would execute in the game, the study involved collecting data from two groups of speakers using CRs, native-native and native-non-native speaker pairs, both men and women. For this reason, the study has two variables: gender and native (language) origin. The gender variable represents how men or women might produce CR strategies when speaking to other men or women. The native origin helps to investigate whether native speakers produce different CR strategies when they speak to other native speakers or non-native speakers. Subject age, however, was not considered as a variable, because it would require a large amount of data to be collected and compared between numerous speaker pairs of different age groups. Compared to the study by Gísladóttir (2015) that included only native speakers but did not provide any multimodal analysis of CRs, this study has a larger impact on designing multimodal behaviour of ECAs, in that it contributes to the knowledge of spoken human language based on a particular culture including data from native and non-native speakers and provides a multimodal description of the CR utterance and its various types. It serves the purpose of creating a realistic human-agent interaction for enabling an authentic practice of communicative skills, not only in Icelandic. Moreover, it contributes to the design of serious/educational games, in that it supports the idea that didactical materials should be designed differently, i.e. using real-life conversational scenarios, and building lessons for practising spoken language based on them. This thesis contributes to the method and approach of communicative teaching for learning a language and culture in a virtual environment. Yet, this study is one of few that discusses the impact of natural language and its multimodal features in L2 education in connection with CALL, and the development of serious computer games for learning. With a detailed study of the repair mechanism, this study also contributes to the understanding of CRs in Icelandic and how they are multimodally produced. A multimodal analysis of CRs has not yet been conducted in Icelandic, neither were any found in any other research in different languages, at the time of writing. This study thus partly examines conversations between L1 and L2 speakers in Icelandic, which makes it one of the studies in Iceland that compare how native speakers speak and behave when they are interacting with non-native speakers, e.g., Theodórsdóttir and Eskildsen (2011), Theodórsdóttir (2011; 2018). This study also contributes to the pioneering work of the CADIA team at Reykjavik University, which among other things, focuses on designing new behaviours that will be modelled and implemented in CADIA Populus with a new method for bringing virtual people to life in an interactive graphical environment. This is achieved by making players aware of their social surroundings, linking the scene to social forces that act with their bodies to continuously produce appropriate body motions in a group conversation (Vilhjálmsson, 2011, p. 5). In Icelandic, there has been little research done on conversational interaction (Gísladóttir, 2015), except for studies on turn-final *eða* (or) (Blöndal, 2008) and the interactional function of *nú* (now) and *núna* (now), the comparison of *nú* (now) and *er það* (is it) (Hilmisdóttir 2007, 2010, 2011, 2016), and the direct speech (Bjarnadóttir, 2009) in Icelandic talk-in-interaction. All of these studies included only native speakers, whereas only one study investigated talk-in-interaction between L2 learners of Icelandic and native speakers of Icelandic (Theodorsdóttir, 2011), in the context of business and learning the L2 (conversion in a bakery), and compared a similar interaction to the one held between native speakers of Icelandic (Theodorsdóttir, 2011). That study showed that interactions between L2 and native speakers of Icelandic has a two-fold focus on both topic and linguistic form, whereas native speakers focus naturally only on the topic discussed. It did not however investigate the multimodal features used in that interaction. The main study herein on multimodal CRs was undertaken in order to endow ECAs with human-like behaviour in this communicative function.

The second auxiliary study was conducted in the so called third phase of the project to test its outcomes. This short study helped to investigate 1) how learners experienced playing *Virtual Reykjavik*; 2) how they perceived the interaction with ECAs in the game; 3) how they perceived the authenticity of their multimodal behaviour (i.e. speech, facial expressions, hand gestures and posture) and 3) to study whether the implemented multimodal behaviour helped them to understand how ECAs elicit CRs. During this study, the game[2] was tested for the first time on adult learners of Icelandic, who were first-year students of Icelandic Practical Diploma Course at the University of Iceland. Due to the general scope of this user-response study with only two out of six CR strategies implemented into the ECAs' multimodal behaviour, i.e. the CR Ellipsis (Hitt húsið) and the CR Interjection Strategy (ha?), the results relate to the general user perception of the application and the perception of the two CRs. Figure 1 below features a player (learner) approaching an agent with the purpose of asking for directions. The player uses a head set with earphones and a microphone, to be able to speak and listen to the other agent's responses.

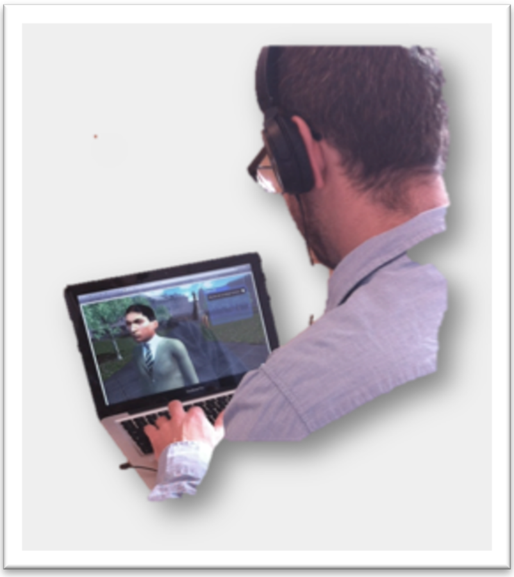


Figure 1: Learner interacting with an ECA in *Virtual Reykjavik*.

[2] The tested game was a prototype and included only one story *Týnda hljómsveitin* (The lost music band), which includes the chapter *Hvar er Hitt Húsið?* (Where is Hitt Húsið?).

3D computer games involve interaction between different kinds of agents. These agents may have a shape of a human body or of other animated characters. Agents used by human users as their animated representations in the game, are called avatars. Based on players' commands, the avatars act and speak accordingly. Each avatar is thus an agent whose actions are controlled by the human player. In such game environments, one can, however, find agents that are not represented by any human players, and therefore are not avatars. These are agents whose actions, responses and movements have been pre-programmed. They have a specific role within each game scenario and are part of the game. For instance, in *Virtual Reykjavik,* the human users (language learners) use agents to explore the environment in the game by walking, approaching and interacting with other agents, and to represent them in the game. In other words, learners use avatars to explore the game environment and meet other agents who may not be avatars. During interactions between the learner's avatar and another agent, various communicative functions take place. The non-avatar agent can only respond to the player's avatar agent in a specific way, whereas the player's avatar can ask, and respond to, any kind of question and move freely within the game. The number of responses from the non-avatar agent, however, is limited. For this reason, the communicative tasks in each game scenario are designed based on a specific conversational scenario. Based on a database of responses to possible questions and answers in each conversational scenario, the non-avatar agents can select the most appropriate one, which in turn limits their capability of interacting freely. This means that when the learner's avatar approaches an agent and speaks to them (*via* microphone), the speech recognition system recognises the learner's speech and transcribes it to text. The text represents an input for the non-avatar agent. If the input has been understood, the non-avatar agent can give an adequate response. The system checks the speech input produced by the learner's avatar. This process involves using NLP tools to validate certain features of the performed dialogue, e.g. grammar errors, semantic coherence, etc. If, however, the input has not been understood due to the learner's choice of words, or pronunciation issues, or technical problems, then the non-avatar agent needs to ask the learner (learner's agent) to repeat and clarify what was said. In *Virtual Reykjavik*, the learner has a first-person view, such that the graphical perspective is from the viewpoint of player's character (see below). Players typically do not see the body of their avatar, though they may see the avatar's legs and feet (see Figure 2).

Figure 2: Learner's view of their avatar in the game. The player (learner) enters the game and starts walking toward an agent. Screenshot from the first scene situating the learner's avatar in Austurvöllur Square upon game start in *Virtual Reykjavik*.

In order to support a natural flow of conversation between the player's avatar and other agents in the game, communicative functions can be used. One such function is the clarification request (CR). This function triggers a repair sequence to clarify what has been said in a dialogue (Thorne et al., 2009, p. 811). In order for the non-avatar agent to execute this communicative function in a more natural way, i.e. not robotic, the agent's embodied conversational behaviour should be endowed with authentic verbal and non-verbal features that would simulate behaviour in spoken natural language by native speakers. The learner's first-person view of agents standing in the Austurvöllur Square in *Virtual Reykjavik* presented upon game start is in Figure 3.

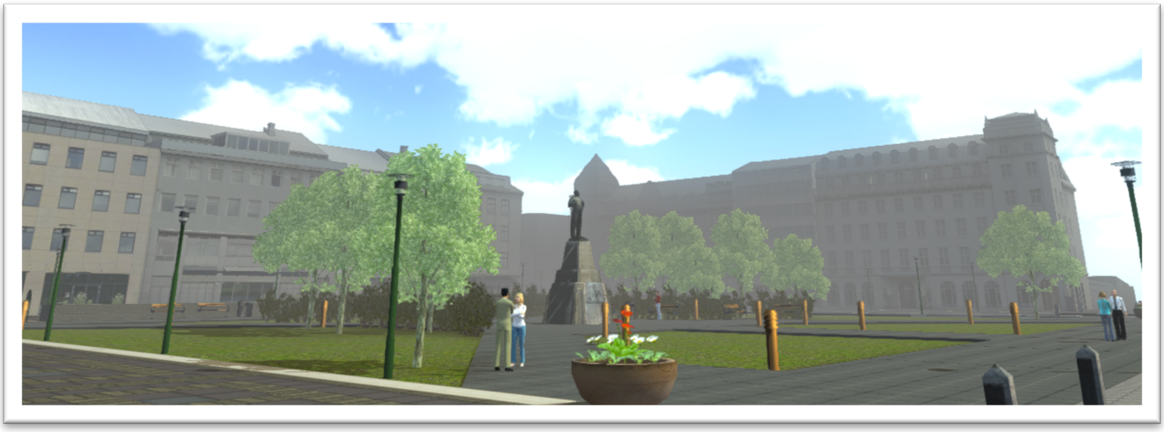


Figure 3: Learner's first-person view. Austurvöllur Square in *Virtual Reykjavik* is a place which learners see upon game start.

Spoken natural language is a very complex phenomenon (Mitchener, 2016). When conducting research in natural language with a focus on multimodal descriptions of spoken utterances, conversations between people need to be video recorded and analysed. This can be a very tedious research strategy, since multiple conversations include multiple turns in dialogues between participants. Each turn in a dialogue may consist of several phrases. Each of these phrases may consist of one or more utterances. A single utterance may include multiple verbal and non-verbal features that help the speaker (producer of the utterance) express the meaning of what they intend to say. These multimodal features need to be described in terms of how they are executed and in which context they occur. This in turn gives a clearer picture about how humans speak, and therefore what is needed to design a more realistic behaviour of animated agents in a game when speaking to language learners. Analysing behaviour in long conversations would require much effort, and for this reason, the present thesis focuses only on one utterance, the CR. Here, attention will be given to describe particular multimodal behaviours in detail. It aims to shed light on how native speakers speak, *via* analysis of verbal and non-verbal features used when asking for a clarification to other native speakers and non-native speakers. After gaining insight in how various types of CRs are executed, multimodal models for implementation into the multimodal behaviour of ECAs, that *Virtual Reykjavik* is populated with, will be presented. When implemented and consequently executed by the agents, a virtual simulation of a real-life conversation could be possible. In this way, learners can achieve a more authentic user experience of the game, which is designed for teaching the language and culture of a given community. In contrast to children acquiring language in their early age based on exposure to the natural language environment, adults often require much time and effort to consciously learn a new language. In today's world, a computer game could assist to bridge the gap between virtual and real worlds, to enable learners to practise spoken language. Learners can thus build confidence in the use of language with virtual characters, while carrying out a task in the game.

Despite Iceland being a small country with comparably a very small number of speakers (338,349 inhabitants on 1 January 2017, Statistics Iceland)[3], Icelandic as a foreign and second language (L2) is widely taught in Iceland, but less widely used in spoken form with foreign speakers because the general preference by native speakers would be to switch

[3] http://www.statice.is/

to English (Jonsson, 2019). In Iceland, especially due to immigration, which, according to Statistics Iceland, was 10.6% of Iceland's population in 2017[4] and 14.1% in 2019[5] - the highest in the previous ten years, and increases in student exchange programmes to Iceland, the teaching of Icelandic as L2 has become popular in the country. In 2017, at the University of Iceland, there were 350 students from 44 countries registered at Practical Diploma and BA courses for Icelandic as a second language[6]. The number of foreign students learning Icelandic currently, i.e., at the time of writing this thesis in 2017, represents about 1% of the country's population (338,349 inhabitants on 1 January 2017, Statistics Iceland)[7]. Based on the results from the first auxiliary study, this thesis also addresses and informs the discussion on the need for having a different learning tool for Icelandic, especially for enhancing oral communication in a 3D virtual learning environment. This kind of a tool is much needed also for learners living in Iceland to practise the language in a safe interim space where mistakes are allowed and where, unlike in real life, learners do not need to feel anxiety when speaking to virtual characters, who never get tired of repeating exercises.

Virtual learning environments (VLEs) have become increasingly common in language education (Dillenbourg, 2000; Sykes et al., 2008; Che, 2016) and in daily lives of people when practising language outside of the language classroom (Kessler, 2018). They often represent a practical way to quickly access and foster L2 skills, from anywhere where there is a computer and internet, towards supporting individual learning (Kessler, 2018). This thesis discusses about how a 3D computer games can help learners practise spoken Icelandic and learn about culture. In 2013, the Center for Analysis and Design of Intelligent Agents (CADIA) at Reykjavik University received a grant from the Icelandic Research Fund[8] to launch a new project in collaboration with the *Icelandic Online*[9] team. The project's name is Icelandic Language and Culture Training in Virtual Reykjavik (also known as *Virtual Reykjavik*), and its aim is to enable future learners of Icelandic to play a

[4] https://statice.is/publications/news-archive/population/immigrants-and-persons-with-foreign-background-2017/

[5] https://statice.is/publications/news-archive/inhabitants/immigrants-and-persons-with-foreign-background-8903/

[6] Statistical information about this has been given to me upon my request *via* email to the Students' Services Department at the University of Iceland.

[7] http://www.statice.is/

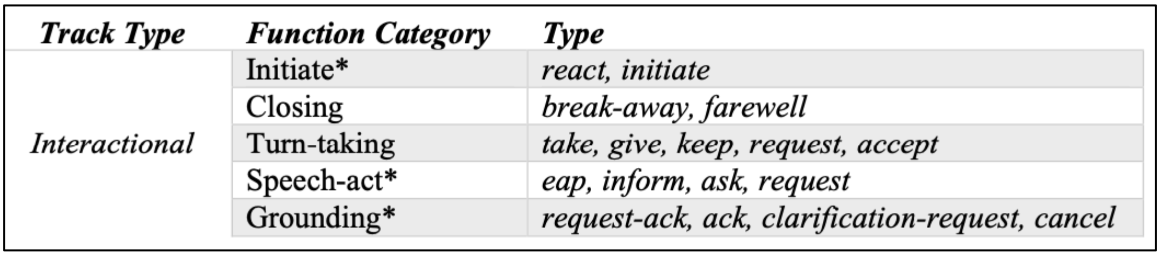
[8] https://www.rannis.is/frettir/rannsoknasjodur/thjalfun-islensks-mals-og-menningar-i-syndar-reykjavik-verkefni-lokid

[9] http://icelandiconline.is/index.html

serious game in order to learn Icelandic language and culture in a context-specific virtual learning environment (VLE). However, the *Virtual Reykjavik* application is intended to those who can already speak some oral language. For the purpose of preparing learners for practising orals skills in the application, *Icelandic Online* (launched in 2004 by the University of Iceland under the supervision of Birna Arnbjörnsdóttir), an open-source tool for teaching and learning Icelandic, represents a 2D web-based curated course (Arnbjörnsdóttir, 2008, p. 48) that can effectively serve as a preparatory stage to improve learners' vocabulary, grammar, pronunciation, and listening on ready-made exercises, e.g. first encounters asking for directions. *Virtual Reykjavik* is based on a previously developed application, the Tactical Language and Culture Training System (TLCTS) (Vilhjálmsson, 2011, p. 2), a programme that teaches a language and unfamiliar culture in a 3D game and lesson environment. The insights from Vilhjálmsson's previous work was used in order to develop a new language and culture learning tool for Icelandic. As an online application, it allows its users to connect *via* Internet at any time and to play the game by solving tasks, speaking to virtual characters, and gaining points for correct use of Icelandic language. The game uses a state-of-the-art Automatic Speech Recognition (ASR) for Icelandic, the Google Speech Recognition System, which is part of this application. It allows its users to experience a verbal human-agent interaction. Users can interact with ECAs that are endowed with multimodal features of natural language and behaviour of real native speakers. With the help of the ASR and the text-to-speech (TTS) system IVONA, the learners will be able to speak to virtual characters and receive a meaningful answer, which is crucial for a successful spoken interaction. This application can be considered as an 'intelligent programme'. It combines a corpus-driven approach, the natural language processing (NLP) with language learning techniques, the ASR and text-to-speech software enabling ECAs to interact with learners' avatar. The corpus-driven approach refers to the use of corpora of actually occurring language data, whether it is written or spoken, that is used for teaching foreign languages (Meunier, 2011, p. 460). According to Meunier (2011), different kinds of corpora, including multimodal corpora, with data from authentic interactions should be collected and analysed to help teach different languages and language modes (p. 467). This may lead to a more learner-centered, context-depended and culture-bound approach (ibid., p. 469). The approach may be more qualitative to corpus analysis, as it can be more appropriate and manageable for teachers and learners (ibid., p. 469). The frequency of occurrences should only inform about what language (vocabulary, grammar, etc.) to teach in which context and at what level (ibid. p. 471). Here, the corpus represents a collection of verbal and nonverbal data from real-life interactions, phrases and sentences used in CRs, which is then used for designing a dialogue in the game scenario. The NLP includes speech recognition tools processing natural language input from the user, using the ASR for Icelandic developed by Google. The text-to-speech gives voice to ECA's language input. The ECAs use text that are, e.g., sentences stored in a database to which voice is given. The programme is delivered through a three–dimensional (3D) graphic entity, which has the ability to interact with a human user by text or speech, either on the web or standalone computer (Morie et al., 2012, p. 2). *Virtual Reykjavik*, as well as other 3D games, such as the *DARWARS Tactical Language Training System* (Johnson et al., 2004) and the *Danish Simulator* (Hansen 2016), uses an ASR system to help learners get immersed into interactive dialogues with agents. An ASR system is one of the ways to enhance the development of spoken communicative skills and cultural understanding. For Icelandic, *Virtual Reykjavik* also offers a virtual learning experience where dialogues unfold naturally, enabling both participants in a conversation, the learner and the agent (see Figure 1), to speak to each other freely, while restricted to the tasks and the dialogue structure in each scene of the game. In its complete version, *Virtual Reykjavik* is also expected to include other intelligent features, such as feedback that inform individual learners about their progress, mistakes, and suggestions for improvement. As such, it will belong to the Intelligent Computer Assisted Language Learning (ICALL) effort that is currently becoming widespread in the field of technologies for education. *Virtual Reykjavik* targets two main groups of learners: young adults and adult learners of Icelandic as L2, to enable them to practise their speaking and listening skills and bridge the gap between the use of language in a classroom and real life. As any other serious game in this category, it should enable players to solve tasks and reach goals with pedagogical objectives, which would help to maximize their learning and have fun while learning. The methodological framework in *Virtual Reykjavik* supports the following main learning approaches: game-based learning, task-based learning, communicative approach, multimodal approach to language teaching, and individual/individualised learning. The combination of these should not only enable learners to be engaged in solving various tasks in the game, but also to collect points and receive feedback in the form of for example transcription of dialogues, suggested correct phrases, and grammar hints, while communicating with ECAs in spoken Icelandic. The ECAs would each use communicative functions in a natural way throughout the whole game when speaking to learners and replying to their questions.

However, very little data is available on non-verbal features of verbal interactions in Icelandic. For this reason, the main study in this thesis examines natural language use, particularly the use of CRs in first encounters when asking for directions, which is conducted for the purpose of *Virtual Reykjavik*. The verbal and non-verbal features in spoken language interactions represent a challenge because they need to be examined in natural settings in the same situational scenario as the game proposes. As the aim of the project is to create ECAs that are aware of the learner and are able to react believably to his/her social presence (Vilhjálmsson, 2011, p. 6), the ECAs then must also be able to speak and act believably in an authentic manner, and in this way expose learners to the real language used in a VLE. This is the first study of its kind in Iceland. No other studies with multimodal features of native speakers of Icelandic engaging in first encounters asking for directions are available for modelling conversational behaviour of ECAs used in *Virtual Reykjavik*. The main novelty in this thesis is using a multimodal approach to investigate CRs in real-life interactions in the same conversational scenario as would be used in the game for practising the language. Moreover, the novelty here also falls into including non-native participants interacting with native speakers of Icelandic and comparing results with other native-native speaker pairs. Consequently, six novel multimodal models for CRs are proposed. Two of the six multimodal CR models would be delivered to the programming team in a form of an annotated overview table for the FML and BML (see multimodal models for CRs in Table 22-27). Table 1 specifies the FML part and shows the Track Type for suggested interactional functions together with their types used in the project *Virtual Reykjavik*. The CR is one of the types in Grounding.

Table 1: Overview of interactional function categories in FML (Functional Markup Language) and their types based on Cafaro's et al. (2014) work that are used in the project *Virtual Reykjavik*. Used under author's permission (Ólafsson, 2015, p. 8).



Track Type	*Function Category*	*Type*
Interactional	Initiate*	*react, initiate*
	Closing	*break-away, farewell*
	Turn-taking	*take, give, keep, request, accept*
	Speech-act*	*eap, inform, ask, request*
	Grounding*	*request-ack, ack, clarification-request, cancel*

The team would implement two multimodal CR types proposed in this thesis into the conversational behaviour of ECAs. It would be for the first time that learners would practise language and culture skills with ECAs in the same game scenario as research was conducted in real life in Iceland. When planning a dialogue between virtual agents and human users, the agent's body needs to be built so as to be able to produce the right variety of communicative functions that the dialogue requires (Vilhjálmsson, 2009, p. 47). Figure 4 below shows a first-person view form the learner's avatar.

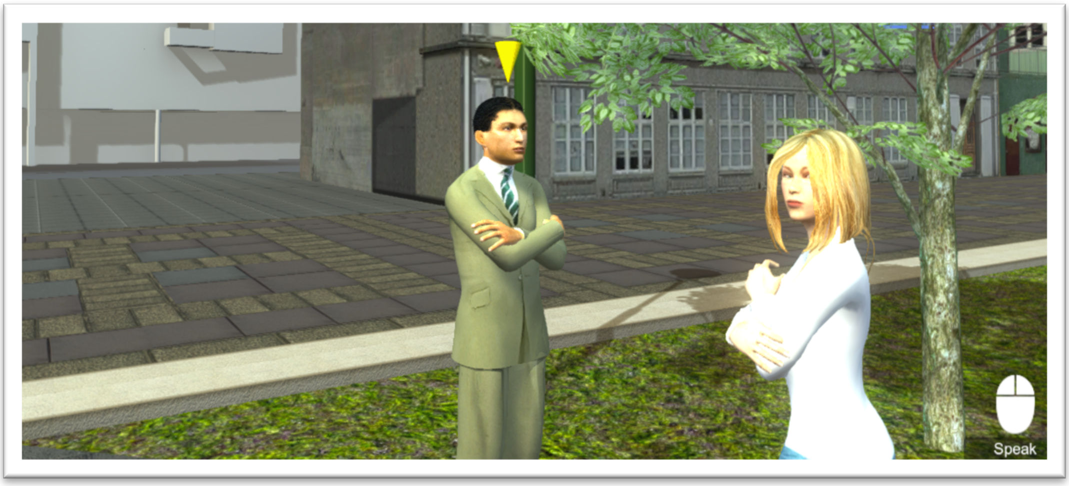


Figure 4: First scenario upon game start featuring agents in Austuvöllur Square in the game. The yellow arrow shows the agent at whom the learner is looking.

The learner sees this situation upon initiating the game. The location where the game starts is Austurvöllur Square in central Reykjavik. The task of the learner is to approach one of the agents and ask for directions.

The prospect for practising speaking skills and learning about the language and culture in an online setting is an ideal way for Icelandic learners living both in and outside of Iceland, because learners can access this interim learning space online. Learners should experience an authentic communication with virtual characters that is close to reality. Such an experience might help them learn Icelandic vocabulary, spoken language phrases that are used in reality (as in contrast to textbooks), and paralinguistic features, such as intonation, and the manner of speaking that is culturally bound. As opposed to traditional language learning textbooks that teach learners formal phrases, this application will teach them dialogues that simulate real-life conversations, and thus train them to hearing and using the language as it is spoken by native Icelandic speakers. Another very practical feature is that the characters are consistent in using Icelandic in communication and do not switch to English. The agents use language as observed from real people in similar real-life situations and interact with learners based on the particular conversational scenario and learners' tasks. The learner's avatar has a shape of a person and is automatically assigned to the learner. When looking down or to the side, the learner can only see feet and hands of their own avatar. When the learner approaches agents in the game, a yellow arrow appears above the agent's head. It is there to show which agent the player (learner) is looking at. Players use keyboard controls to move their avatar's head and body towards that agent. This is a way for the learner to express non-verbal behaviour. This application thus represents the first virtual interim space where Icelandic language technology is combined with artificial intelligence and gaming, to give learners of Icelandic an opportunity to practise speaking with virtual characters simulating real Icelanders. The following section introduces to the background and motivation to this thesis.

1.2 Background and Motivation

The most advanced online course for learning Icelandic is *Icelandic Online*, www.icelandiconline.com, which provides curated materials for practising various language skills, except for speaking. This web-based course is mainly built for practising vocabulary, grammar, pronunciation, reading, and listening, by providing elaborate exercises and feedback for each skill (Arnbjörnsdóttir, 2004). Today, the variety of online courses for Icelandic has grown. Besides *Icelandic Online*, learners can choose between web-based courses offering learning materials for vocabulary, grammar and listening activities, such as *Tungumálatorg*[10], listening activities on *YouTube*[11], or checking pronunciation of a restricted word database on *Forvo*[12]. For practising speaking in a VLE with virtual characters with authentic language, there is no application available on the market yet. The reason for that could be due to lack of funding, or simply, as Hampel and Hauck (2006) claim, it could be due to the dominance of writing as part of the skills trained in language education in Western society, which might have lessened the inclusion of other modes (p. 4) in language practice online. This, as well as lack of available technology, might have postponed the inclusion of speaking practice into new applications. This thesis suggests that speaking can be practised with ECAs in *Virtual Reykjavik* as part of a language and culture training online.

A body of research (Halliday, 1989; Crystal, 2005; Nelson et al., 2005; Zhang, 2013; Hulme and Snowling, 2013) highlights the two different communication modes - speech and writing, that are different from each other due to contexts in which they occur.

[10] http://tungumalatorg.is
[11] https://www.youtube.com/channel/UC3Y0LUUjCHoKTOGYjyYcbDQ
[12] https://forvo.com/languages/is/

This consequently influences the way they are learned. Writing is context-reduced and lexically dense, whereas speaking is part of the complex system of language that draws from different resources (modalities), each with their own materials and affordances for making meaning (Hampel and Hauck, 2006, p. 5). For instance, spontaneous speech is unlike written texts because it contains many mistakes, short sentences, occurs spontaneously, and the whole speech may contain hesitations and silences (Halliday, 1989, p. 76). The development of pragmatic speaking skills may positively affect and thus enhance reading comprehension (Hulme and Snowling, 2013, p. 1). For this reason, there are also different ways of approaching the teaching of these modes. Speaking needs to be practised through oral exercises, supported by pragmatics. When building a computer game for speaking practice and creating the conversational scenarios the learner will engage in with the agent, several contextual references need to be taken into consideration. Spontaneous speech contains many multimodal cues, e.g. paralinguistic features of intonation, silence and speed (Halliday, 1989, p. 77). When humans speak, many verbal and non-verbal features are used to express ideas and thoughts. Such multimodal features help listeners to better understand what is being said and, in this way, to reach mutual understanding (common ground). Even in this process of reaching common ground (grounding), mistakes and misunderstandings occur. For instance, inappropriate vocabulary might be used by non-native speakers, or various paralinguistic features (e.g. intonation, stress) may be used incorrectly. Non-verbal expressions, such as facial features and gestures can also contribute to misunderstandings when they are used inappropriately in a different culture. Because mistakes or misunderstandings are a natural part of spontaneous speech, there are various mechanisms that control a smooth flow of a conversation. The CR is one of the mechanisms.

For language learners in today’s globalised and technologically advanced world, it is very important to have easily accessible tools for practising oral language skills. However, such applications are lacking. Some of the most popular applications, e.g. *Second Life*[13]*,* supports “speaking” (deliberate quotation marks) through Cypris Chat English (chatting while typing in messages). However, others including *The Tactical Language and Culture Training System*[14] for Arabic (2007) and *The Danish Simulator*[15] (2012), both of which are still considered prototypes of serious games, and the fully

[13] http://secondlife.com/

[14] https://www.alelo.com/case-study/tactical-iraqi-language-culture-training-system/

[15] https://www.alelo.com/case-study/the-danish-simulator/

developed application *ELSA (English Language Speech Assistant)*[16], which was launched only in 2016 for the purpose of correcting pronunciation, are the most known to offer oral language practice with the inclusion of speech recognition. *Virtual Reykjavik* differs from those above in that it applies specific verbal and non-verbal features into the ECAs' multimodal behaviour based on context, and in this way contributes to the authenticity of spoken language by the agents. The main study in this thesis examines clarification strategies in Icelandic in a multimodal way and thus contributes to the authenticity of spoken language by ECAs in *Virtual Reykjavik*.

Several studies published between 2004 and 2013 supporting L2 education in VLEs with virtual-reality features were analysed (Lin and Lan, 2015). The results inform about the most popular tools that support interactive communication; behaviours, affections and beliefs; and task-based instruction. The results also shed light on the need for including teachers for giving additional instructions to learners about how to use particular VLEs for completing tasks. This means that online tools are considered as additional learning tools that are part of the teaching process, both within and outside of a classroom, because they have the ability to track the learner's progress and provide individual feedback. However, what is not known yet is to what extent does the natural language behaviour, i.e. multimodal behaviour, which makes artificial humans (agents) in 3D applications believable or realistic, help learners develop their language and cultural skills. *Virtual Reykjavik* is a pilot project in this very scope which implements multimodal behaviour from real-life conversations into the conversational behaviour of ECAs in the game. This Icelandic Language and Culture Training in Virtual Reykjavik project had started nine months before the work on this three-phased project presented in this thesis begun. Amongst those communicative functions, that the team wanted to implement, it became clear the CR would be one of the two very useful functions that needed to be investigated multimodally in real-life conversations. The first communicative function was the Explicit Announcement of Presence (EAP) as decribed in (Ólafsson et al., 2015) which was part of this larger project *Virtual Reykjavik*.

The EAP was examined by Ólafsson (2015) in order to find out how strangers meet in first encounters. The results showed that in 34 cases out of 43, a real-life interaction between strangers was initiated by the EAP function, i.e. the participant calls attention to oneself in order to initiate the approach. This prompted the inclusion of this function into

[16] https://www.elsaspeak.com/home

Virtual Reykjavik. Therefore, the EAP communicative function was implemented as part of the *Virtual Reykjavik* system. By using this function, the learner's avatar could initiate a conversation with the agent in the game by pressing a key that triggers the instantiation of an EAP object. For instance, this signals the animation module of the user's avatar to generate behaviour appropriate for an EAP, for example, a hand wave or a head toss. If the virtual character that is being approached accepts this invitation, the conversation can begin. In his thesis, Ólafsson (2015, p. 17) describes the processes involved in the production of different communicative functions, including the EAP and CR. The processes are included in a block consisting of five different states of the agent (see Figure 5). State (0) represents the agent's initial state. From here, the agent, who has the turn in a dialogue, can perform two actions, either producing the communicative function *Ask*, which would consequently lead to state (1), or not performing any action, which would consequently lead to state (3). Here for instance, the learner's avatar arrives to the agent in the game and performs the communicative action *Ask* (Where is Hitt húsið?). This action leads to state (1). The system checks the speech input produced by the learner's avatar. This process involves using NLP tools to validate certain features of the performed dialogue, e.g. grammar errors, semantic coherence, etc. If the speech input is not understood by the agent in the game to whom the learner's avatar is talking, then that agent in the game performs the communicative function *ClarificationRequest* (e.g. "Ha?"or "Hitt húsið?"). This leads to state (4). Here, the system checks the input from the learner's avatar. If it is understood, then it leads to state (3). If it is not understood, then it leads to state (1). The learner's avatar will take turn and perform Ask (Where is Hitt húsið?) and this will lead again to state (1). When the input is understood, then the agent in the game performs the communicative function Inform, in which the agent gives the learner's avatar information about where the place is, which leads to state (2). When the learner's avatar does not understand the information, the learner (learner's avatar) can take turn and perform ClarificationRequest, thus leading to state (5). Then the agent performs Inform again, thus leading again to state (2). Finally, when the information given (input) is understood by the learner's avatar, it will lead to state (3). In order to prevent from "looping", i.e. going repeatedly between state (1) and (4), the scenario in the game includes three conversational tasks for the learner to follow. These three tasks represent a) seek attention, b) ask directions to Hitt húsið, and c) say goodbye, to both guide the learner in a dialogue and to indicate when his/her turn is. The learner sees these three tasks in the top right corner of the screen. The green check button indicates which task the learner has completed. The team of programmers working at *Virtual Reykjavik* will implement the ClarificationRequest function and its multimodal realisation into the AskInformBlock.

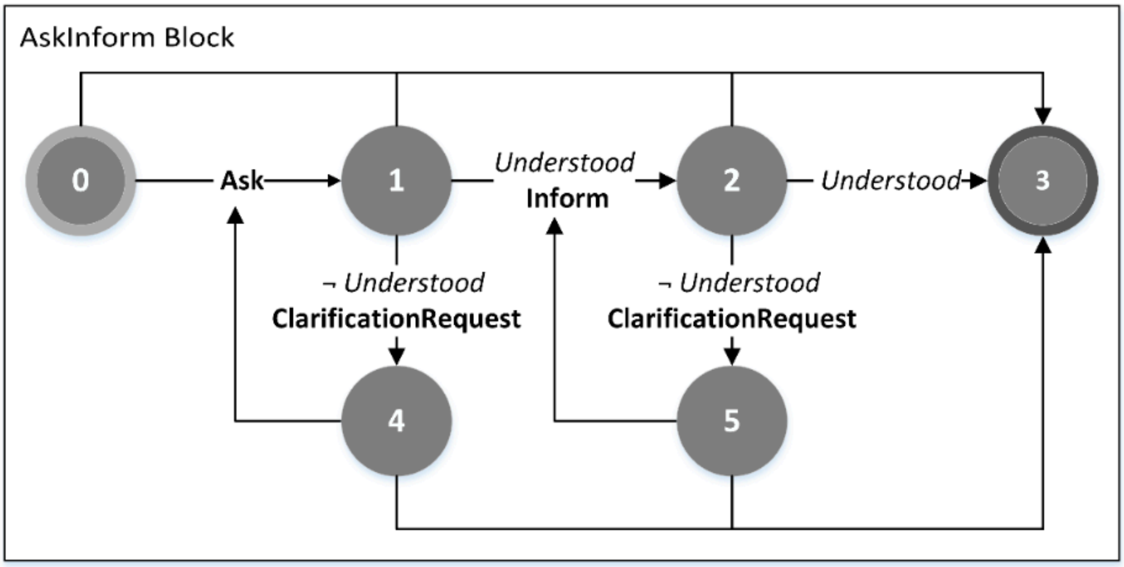


Figure 5: Visual reproduction of the AskInform Block in *Virtual Reykjavik*. This block consists of five states. Each number (0-5) represents the state of the conversation. The methods that produce agent‘s communicative functions are: *Ask, Inform, ClarificationRequest*. Depending on the input (intention, dominance, speech) coming from the learner’s avatar, the state machine will produce suitable communicative functions as the dialogue progresses. Reproduced under author’s permission (Ólafsson, 2015, p. 17).

This thesis provides empirical data helping to simulate real-life conversation with ECAs using multimodal CRs. The empirical data will be presented in the main study of this thesis. The results will be used to create multimodal CR models. These models will be delivered to the programming team in *Virtual Reykjavik* who will use it to implement two of the six suggested models. The team will use Unity’s Animator[17] to realize the multimodal behaviour in ECAs using the suggested models. In this way, the multimodal behaviour in CRs will give learners a more realistic experience of the game while keeping smooth flow in a conversation with virtual agents, thus contributing to a natural-language conversation in a safe interim space for learning Icelandic. The term *safe* indicates a virtual space where learners are allowed to make mistakes and where agents are not tired of practising spoken language skills with learners. As the first auxiliary study reveals, this is particularly important for L2 learners of Icelandic located in Iceland. The second auxiliary

[17] https://docs.unity3d.com/Manual/AnimationSection.html

study presents qualitative data from a pilot study of users playing *Virtual Reykjavik* and practising spoken language with ECAs. Due to English being the lingua franca used amongst and towards non-native Icelandic speakers, which is often very practical, the relatively widespread use of English in Iceland limits the target language exposure. The above tool should offer another possibility to 'meet' virtual native speakers of Icelandic (agents) and practise spoken language in an interim learning space, to prepare them for a more complex use of language in reality.

1.3 Research Questions and The Multimodal Approach

Ideally, realistic multimodal behaviour in virtual reality should be based on research in real-life interactions between humans. Such research provides authentic data to design a realistic behaviour in virtual agents, helping learners to develop language fluency in a simulation of reality. 3D virtual learning spaces are suitable for this purpose. Since language is a complex phenomenon and it is not possible to rigorously examine multimodal features in all communicative functions in a human conversation, the present thesis focuses only on one such function – the CR. The conversational situation is situated in first encounters asking for directions. This section introduces three overarching research questions guiding this three-phased thesis consisting of one main study and two auxiliary studies. After the introduction of the overarching research questions, particular research questions included in each of the three studies will be presented. The justification for using the Multimodal Approach in this thesis, especially in examining multimodal features in CRs in the main study, is provided in this section as well. The overarching, general research questions that helped to guide the main goal in this thesis, while contributing to answer partial goals in the larger project *Virtual Reykjavik*, are:

1. What are general expectations of L2 learners of Icelandic from a 3D game for learning Icelandic with virtual characters;
2. What multimodal features in CRs are needed in order to design a more realistic human-agent interaction in *Virtual Reykjavik*; and
3. What are learners' general experiences with playing *Virtual Reykjavik*, in which the ECAs use the suggested multimodal CR models in first encounters?

1.3.1 Questions guiding the auxiliary study on learners' expectations from **Virtual Reykjavik**

The first study is a preliminary needs study about *Virtual Reykjavik*. In this shorter study, however, the following questions helped to design the survey:

1. Learners' country of origin;
2. What language skills learners want to practise in the language course;
3. What language skills they expect to practise in a 3D computer game for Icelandic, which has virtual characters that are able to speak;
4. What elements such a 3D computer game should contain to make it enjoyable for them to play and learn;
5. What would be the advantages and disadvantages of such a game;
6. How do learners feel about using Icelandic in a face-to-face conversation with local native speakers?

The complete list of questions is presented in Appendix A. The survey was distributed online approximately one year after the project *Virtual Reykjavik* had started. As there had not been done any preliminary needs study before the start of the project, this short survey fills in the gap and represents the first step in designing a more rigorous user study about learners needs for *Virtual Reykjavik* in the future.

1.3.2 Research questions guiding the main study on multimodal CRs

In the second and main study of this thesis, the multimodal features in CR strategies between native and non-native speakers in real-life are examined. This study provides information about the verbal and non-verbal features used in CR responses by natives in a real-life conversation with first encounters asking for directions. The findings are then used to create models for multimodal CRs, in order to enhance the authenticity of the interaction between learners and agents in the game. In this context, the multimodal approach to language (Vigliocco et al., 2014) is used for exploring these features and for creating the multimodal model of CRs. For the former, Multimodal Interaction Analysis (Norris, 2004, 2013) has been chosen as a method for data analysis. This method is mainly concerned with the human being in interactions, i.e. how a person displays the structure of various higher-level actions, such as meeting and greeting someone, that consists of a chain of different utterances in a conversation between people. In addition to this, the Multimodal Corpus-Based Approach (Bateman, 2012), that talks about the body language and facial expressions that co-occur in speech, is used for creating layers of multimodal data, e.g.

transcription, intonation, position of speakers, use of non-verbal features in certain utterances in a multimodal language corpus. As for the latter, by examining the behaviour of people in natural language settings, one can gain information about the specific verbal and non-verbal features used in a human talk-in-interaction, and gain insight into the problem which helps to define the multimodal model for a CR that ECAs will use in *Virtual Reykjavik*. This thesis then proposes six novel theoretical models for six different types of multimodal CRs. Information in various types of CRs is described in more detail, which consequently helps to design realistic behaviour in the ECAs. Human-human interaction in face-to-face is often viewed as a fundamental or primary form of communication (Clark and Wilkes-Gibbs, 1986; Clark, 1996; Bavelas et al., 1997), from which other forms are derived. This thesis bases its findings in multimodal CRs on observing such interactions in real life.

The multimodality of language is a fundamental approach when studying natural language in interaction, because it helps to understand the behaviour that is carried out in each communicative function of a conversation. It ultimately helps to inform this thesis about theories and approaches used to support its main research. Multimodal Interaction Analysis (Norris, 2004, 2013) was chosen as a method for data analysis because this method is mainly concerned with the human being in interaction, i.e. how a person displays the structure of various higher-level actions, such as meeting and greeting someone that consists of a chain of different utterances in a conversation. The aim of this study is to find multimodal features that are present in CR utterances during first encounters. It aims to find what types of CR utterances are used, whether there are any differences in their use between men and women, or when native or non-native speakers are involved in the conversation. It sheds light on what types of utterances the ECAs need to be endowed with, and whether male or female characters need different features. Therefore, in addition to identifying the type of CRs used, the study has only two independent variables: gender and native origin. Player age is not considered as a variable, because it would require a large sample size to permit appropriate comparisons between numerous speaker pairs of different age groups and was beyond the resources available for this study. Nonetheless, the sample group includes speakers in the range of 18-70 years old, which will be sufficient to compare individual data among speakers of different ages. The research questions are composed based on the above and include two groups: one focus group, which had only native speaker pairs, and the other one as a control group, which had native (NS) and non-native (NNS) speaker pairs. The main aim is to guide the study to discover similarities or differences between the production of CRs between different speaker pairs. The complexity of the problem lead to the construction of four main research questions that are presented below. In a given scenario of asking for directions in while playing *Virtual Reykjavik*:

1. What are the common CRs that Icelandic NSs (random men and women, aged 18-70) make in a face-to-face interaction with other NSs and NNSs (actors, men and women, aged 18-70) in order to initiate speech repair?
2. Are there any differences between genders[18], NSs and NNSs, or in other words, can there be any difference found in making CRs when speaker pairs consist of different pairings of gender and native origin?

Once the data has been collected, the following decisions will be made:

3. Which of the verbal and non-verbal features most commonly used by NSs of Icelandic during CRs are critical for implementation into the multimodal behaviour of virtual characters, in order to simulate authentic interaction for better learning of language and culture? Or in other words, which of the multimodal features found in human CRs should be selected to build a model for multimodal CRs that can be implemented into the conversational behaviour of ECAs in order to simulate natural conversation in a virtual environment on computer, i.e. *in silico*, so that the users (learners) get a better language and culture learning experience?
4. To what extent do learners of Icelandic believe the natural behaviour incorporated in ECAs will help them improve learning of the language and culture?

After the analysis of results, six models for multimodal clarification strategies will be suggested for implementation, but due to time and resource constraints, only two will be implemented into the conversational architecture of ECAs in *Virtual Reykjavik*.

1.3.3 Questions guiding the auxiliary study about users' experiences with Virtual Reykjavik

The third study is an auxiliary study investigating how learners perceive (a) playing *Virtual Reykjavik*; (b) the interaction with ECAs in the game; (c) the authentic multimodal

[18] Carli (1989) suggests that a difference in gender affects partners' behaviour in interaction, e.g., men interacting with other men have less effective influence on each other than when interacting with women (p. 565). Similarly, Ridgeway and Smith-Lovin (1999) see a difference in interaction between genders because it perpetuates status beliefs, leading men and women to recreate the gender system even in everyday interaction (p. 191).

behaviour of ECAs; and (d) whether the game assists the learners with learning Icelandic. The following research questions have been stated to guide the pilot user study:

1. How do learners perceive playing *Virtual Reykjavik*?
2. How do learners perceive the interaction with ECAs in the game?
3. How do learners perceive the multimodal behaviour of the ECAs, i.e. speech, facial expressions, hand gestures and body posture while engaged in CRs?
4. Does the ECAs multimodal behaviour feel natural?
5. How effective is the game for learning Icelandic language and culture?

The next section informs about the contribution and organisation of this thesis.

1.4 Contribution and Organisation of the Thesis

The main contribution of this thesis is that informs the development of realistic virtual spaces or interim learning spaces for practising language skills. The study also calls attention to the need to base learners' interactions with virtual agents on authentic human interactions supported by empirical data. The study specifically contributes six theoretical models for multimodal CRs in Icelandic. The significant original knowledge presented here is based on the empirical investigation of multimodal (verbal and non-verbal) features in CRs found in real-life conversations between both native and non-native speaker pairs in first encounters. This resulted in defining five types of CRs that had also been found in a previous research and one novel type of a non-verbal CR, which had not been found in any research on spoken interactions before (Gísladóttir, 2015; Purver, 2004; Dingemanse and Enfield, 2015). However, the non-verbal CR type had indeed been found in research on Argentine Sign Language (Manrique, 2016), but described as a "freeze-look". This means that multimodal analysis of spoken interactions is necessary in future research to be able to find, compare, and describe in detail language phenomena across spoken and sign languages.

The practical contribution in this thesis results from the theoretical investigation of multimodal CRs. The suggested multimodal models for CRs would help keep a natural flow of a conversation between learners and the agents in the game, and help the learners learn ways how locals ask for clarifications. According to Shin (2018), learners, by using embodied condition, will be able to embody experiences by viewing, playing, and feeling perceptual cues linked to those experiences (p. 68). The suggested models will be delivered to a team of programmers working in the project *Virtual Reykjavik*, who will implement them into the conversational behaviour of ECAs when executing the CR function in the game. Consequently, it will contribute to the improvement of fidelity, i.e. how true to real life the multimodal behaviour of ECAs is, which in this context represents a more realistic graphics and sensory input for learners practising communicative skills with ECAs in the game. On a world-wide scale, the contribution in this thesis is very significant in that it delivers a detailed description of each of the six types of multimodal CR models which had not been previously done in this scale in any of the previously reviewed research. Researchers and designers of 3D interfaces concerning ECAs in spoken dialogues can use these models to help realise multimodal behaviour in CRs. Although the main study in this thesis is language specific, e.g. the intonation of some of the CR types may be different in other languages, the data description is very detailed and can provide a guidance how to conduct a similar study and compare the results.

The two auxiliary studies presented in this thesis will partially contribute to the general goals of the larger project *Virtual Reykjavik*, but will also inform about the need to have 3D virtual characters for practising spoken communicative skills with learners of Icelandic in Iceland, and how the learners' experience is during the first pilot study when learners play the prototype of the game with ECAs that are endowed with multimodal features in two CRs. This will help to inform the future design of *Virtual Reykjavik* and contribute to designing a more rigorous study in the future.

This thesis is organised in five chapters. Chapter 1 is part of the general introduction to the topic and organisation of this thesis. Chapter 2, The Theoretical Framework, includes the theoretical background for this thesis, including the importance of virtual environments for language learning in today's world, and the role of CALL and learning approaches used to enhance L2 learning. It includes a section on modelling interactions between human users and agents in virtual environments, while highlighting theories and approaches that should be considered when building a dialogue in a simulation of a real-life environment in such a context. Chapter 3, The Studies, includes two supportive auxiliary studies and one main study on multimodal CRs. It includes the methodology and results, as well as the description of multimodal CR models. Two of these models are partially tested in the user pilot study on perception of the general interaction and multimodal behaviour of ECAs, as well as the learning effect. The user response study, which is one of the auxiliary studies, describes the learners' general experience from the game and gives information about the use of the two implemented CR strategies by ECAs. The impact of the main CR study and the user response study are discussed in Chapter 4. This thesis is then concluded in Chapter 5, in which contributions, limitations and future work are discussed.

1.5 Summary

This chapter presents the motivation and rationale for conducting a three-stage project in this thesis. It introduces the research questions in all three studies and the Multimodal Approach used for exploring the verbal and non-verbal features in CRs and for creating the multimodal model of CRs. It concludes with the organisation of this thesis. The next chapter discusses the theoretical framework for modelling a realistic human-agent interaction in a 3D computer game, to help learners speak Icelandic in an authentic manner. Based on the selected pedagogical background, this might be achieved by solving tasks, while interacting with virtual characters and solving particular tasks in the game. By the end of the following chapter, the focus will be narrowed down to multimodality and CRs in order to demonstrate the importance of research in natural language and including authentic features into the multimodal behaviour of ECAs in *Virtual Reykjavik*.

2 Theoretical Framework

2.1 Introduction

This chapter introduces the theories and approaches that form the theoretical basis for this thesis. It introduces the pedagogical strategies to reach the pedagogical goal of *Virtual Reykjavik*, i.e. an authentic human-agent interaction. It furthermore focuses on spoken language and multimodal features in the utterance of investigation, the CR. As language is a complex phenomenon and includes many consecutive utterances in natural speech, this particular utterance forms an optimal example to demonstrate the multimodal features that ECAs are endowed with in the game. In the beginning of this chapter, the interdisciplinary approach of Computer Assisted Language Learning (CALL) is introduced that supports *Virtual Reykjavik* as a 3D computer game for language learning. CALL combines a corpus-driven approach and an NLP approach with language learning techniques, that are designed for individual learning. The game has a speech recognition and a text-to-speech software that enable ECAs to interact with learners. With these features included, it shifts towards intelligent CALL (ICALL), which forms the next stage of CALL development. Consequently, the theoretical background to computer games in language learning is introduced, because the multimodal features of CRs suggested for implementation will be part of agent conversational behaviour in *Virtual Reykjavik*.

In order to create a realistic human-agent interaction in a 3D VLE such as *Virtual Reykjavik*, the second part of theoretical framework covers relevant theories and approaches that range from technology, through cognitive studies to linguistics, discourse, and a simulation of real-world scenarios in training situations with virtual characters. To create a realistic human-agent interaction, *Virtual Reykjavik* uses natural language as a source for creating both the learning materials and the interaction for practising spoken Icelandic. Since data from natural language are used to endow the multimodal behaviour of ECAs when executing the CRs, the topic of language as a complex system is also discussed here. In this context, the complexity theory will be introduced because it is used not only to describe the complexity of natural language and its multiple cues that help express the speaker's idea, but also to understand the complex process behind teaching an L2. As multimodal features of CRs are used for endowing ECAs that represent speakers of a particular culture, the CR strategies used by them show learners how native speakers ask for clarification. By observing these strategies during the course of a spoken interaction and understanding how they are made, the learners may possibly learn to reproduce them when conversing with Icelandic speakers in real life. How people perceive and produce language (embodied cognition) is also briefly discussed within this section. The concept of embodied cognition supports the view that language is produced and perceived through various senses and helps to understand how spoken language is used in context. This is particularly important when creating a realistic human-agent interaction, where the agents represent native speakers. In this context, the agents should be able to simulate a real-life conversation by employing various modalities, such as facial expressions, hand gestures, or body movement, to help them express the meaning of their spoken utterance.

When humans speak, they are said to perform ‘speech acts’, that contain one or more multimodal interactive utterances, i.e. each speaker employs multiple modalities to produce meaning. Speech acts are typically spoken words accompanied by other non-verbal cues. The CR is thus considered as a multimodal utterance and part of a speech act. In this context, speech act theory is an important part in the present theoretical framework. It informs about the turn-taking mechanism between participants in a conversation that are trying to reach a mutual understanding of what is being said. As a result, CRs help to achieve a mutual understanding between participants in a conversation, as participants use clarification strategies to ask the other speaker to clarify what they have previously said. In *Virtual Reykjavik,* speech acts are part of a pre-set conversational scenario. Each task prompts the learner to ask a particular question, to which the ECA answers appropriately. As all conversations happen in a particular context, it is important to mention how context is used for creating an authentic human-agent interaction for the purposes of teaching L2 and keeping the learners immersed and motivated when playing the game. Before discussing this issue, the following section places *Virtual Reykjavik* within a pedagogical background and consequently proceeds with a subchapter on Computer Assisted Language Learning (CALL). Lastly, computer games in language learning and the main learning strategies used within are discussed.

2.2 Pedagogical Background for *Virtual Reykjavik*

Computer games for language learning belong to the category of VLEs that are types of social spaces, which can vary from spaces enabling work with text and texting to complex 3D immersive worlds. The process of immersion turns them into more effective educational VLEs that provide spaces where learners participate in solving tasks to enable learning. VLEs not only support individual learning outside of the classroom but can also be included in the teaching and learning process within a traditional classroom. By including computer games for language learning, the learners enrich their activities because they integrate heterogeneous technologies into language learning (Dillenbourg et al., 2002, pp. 3-4). Computer games can also provide learners with opportunities for reflection (Vallance and Martin, 2012, p. 2 & 6) on how good their spoken language skills are, which words they have difficulties pronouncing, etc., depending on what language skill a particular game offers. In this way, learners can practise language skills outside of the classroom in so-called 'interim learning spaces', or spaces 'in between' the artificial setting of the language classroom and real-world use of language for communication. Interim spaces enable learners to access the L2 environment that simulates the real one and practise the various language skills that facilitate learning (Forteza and Pastor, 2014, p. 135). Many of these interim spaces also use different methodologies and approaches to provide appropriate materials for language learning (Romero and Carrió, 2014, p. 150). For instance, 3D computer games (Shudayfat et al., 2012; Barkand and Kush, 2009) and other real-world simulated 3D environments, such as *Second Life,* belong to 3D VLEs that represent a shared 3D virtual world. This world provides context for language learning. For instance, in computer games learners can use other agents to move between spaces, e.g. a living room or a castle. This provides a context for the use of language (vocabulary) in a different scene. In such environments, learners not only perform tasks and learn about new things, but they also receive feedback from other players (peers) or agents in the game. Learners have the opportunity to re-do tasks as many times as the game allows, while the system collects data about their progress. These represent 'intelligent environments' where learners are allowed to make mistakes for the purpose of learning, have personalised progress and learning scores based on their achievements, and achieve goals while collecting rewards. In the context of this thesis, *Virtual Reykjavik* belongs to such environments. This section describes the pedagogical foundation for *Virtual Reykjavik.*

2.2.1 Computer Assisted Language Learning

From a historical point of view, Farrington noted in 1989 that in order to make language learning more attractive to learners, the use of the latest digital technologies and software for language learning should be supported. He moreover noted that various CALL programmes "will soon, if this is not already true, only motivate if they are challenging, perplexing and interesting in themselves, like any other language learning activity" (Farrington, 1989, p. 67). The CALL approach has since encouraged a proliferation of online tools for language learning, i.e. various applications and programmes have been used in various devices that have assisted with L2 learning. Advancement of technologies and increased learners' demands for programmes are enabling more and more attractive L2 learning. Such applications should support interactive L2 learning by interacting with tutors and other learners in real-time, give instant feedback, offer correction of errors, include features for personalisation, and be equipped with speech recognition to enable practising oral language. Chapelle (2008) summarises CALL as the advancement of technologies which affords opportunities for communication to be mediated through internet by using various tools. Learners of languages from different parts of the world can connect with their peers and practise language even beyond classroom settings. In this context, teachers are also participants; they can use various tools to design contents that are available later for learners to practise language skills, and thus create learning activities with controlled input. CALL thus expands possibilities for developing individual learning and oral language practice. For instance, speaking can be increased through goal-based communicative activities that are implemented in various applications, such as computer games for language learning where speech recognition is used.

When looking back at the development of CALL, the late 20th and the beginning of the 21st century was a time when new opportunities in language learning arose due to globalization, the rise of the Internet, and the consequent development of new tools and technologies supporting web 2.0. The CALL approach was being especially pronounced in L2 education at this time (Levy, 1997, p. 1). CALL as an approach to L2 learning supports the use of computers as "an aid to the presentation, reinforcement and assessment of material to be learned, usually including a substantial interactive element" (Davies, 2000, p. 90) at an educational establishment as well as at home (Kenning, 1990, p. 67). Since then, this approach has gone through several stages of development. Warschauer and Healey (1998) defined three stages: the behaviouristic CALL, the communicative CALL, and the integrative CALL. Each stage reflected technological advancements and certain popular pedagogical approaches of that era. These three stages are described below.

Behaviouristic CALL (1950-1970) featured repetitive language drills. It was informed by the behaviourist learning model of stimulus-response and was most popular in the United States. It was the era of the first personal computers, which allowed students to work at an individual pace. The computer system PLATO[19], a tutorial system developed by the University of Illinois, offered a range of courses and exercises, and was the most popular at that time (Warschauer and Healey, 1998, p. 57). Communicative CALL (1970-1980) emerged due to greater possibilities for individual work while using more advanced personal computers. Communicative CALL was informed by cognitive theories which supported the view that learning was a process of discovery, expression and development. In this period, text reconstruction programmes and simulations were the most popular among learners, because learners could work either in pairs or groups, which stimulated discussion and discovery (Warschauer and Healey, 1998, p. 57). The third stage was the integrative CALL (1980-1998), which was followed previous CALL periods. It included task-based, content-based and project-based approaches that would seek to integrate various language practising skills, such as listening, speaking, writing and reading. Integrative CALL had new technological tools that were more integrated into the language learning process, both inside and outside of traditional classrooms. During this time, "the multimedia networked computer [was] the technology of integrative CALL" (Warschauer and Healey, 1998, p. 58). *Icelandic Online* represents one such tool. It is a web-based course and mainly built for practising vocabulary, grammar, pronunciation, reading, and listening, using exercises and feedback for each skill (Arnbjörnsdóttir, 2004). Speaking, however, could not be practised here as the technology has not been available.

By looking at the previous three stages, this thesis situates *Virtual Reykjavik* into a fourth stage of CALL development, the intelligent CALL, or ICALL (2000-present). At the end of the 20th and the beginning of the 21st century, an interest remains in including artificial intelligence (AI) technologies into language learning (Gamper and Knapp, 2002, p. 329) and development of language learning materials and applications, e.g., creating intelligent tutoring systems "which are capable of processing and giving feedback on free language input" (Finkbeiner and Knierim, 2008, p. 402). According to Schwienhorst (2008, p. 140), ICALL includes intelligent tutoring systems that not only analyse utterances and give feedback to learners in an online mode, but they can also do it offline. The system uses automated feedback and stores data to form a learner corpus, which is further used to create and revise learner modes. By definition, such systems need to have three types of intelligence: (1) the subject matter or domain must be known to the computer system well enough to be able to solve problems in the domain; (2) the system must be

[19] https://en.wikipedia.org/wiki/PLATO_(computer_system)

able to find out and detect the learner's approximation to that knowledge; and (3) the system must be 'intelligent' enough to implement strategies and pedagogies to reduce the difference between expert and student performance (Burns and Capps, 1988, p.1). Other ICALL features include the so-called inspectable or viewable user model (UM) which records the user's steps and mistakes, natural language processing (NLP) systems that check grammar and spelling, automatic speech recognition (ASR) systems that allow user's speech to be recognised and thus enable a 'face-to-face' dialogue with virtual agents, and also machine translation systems for comparing texts translated by the users and correct them according to model sentences (Gamper and Knapp, 2002, pp. 332-335). This intelligent CALL can thus be summarised as an interdisciplinary approach to CALL using systems that enable storing, evaluating, and structuring data used for feedback purposes in the design of individual learning. This new development in ICALL supports various forms of interaction with the system - both learning about the user's progress and acting as their tutor. It moreover supports realistic interaction that can be, for instance, achieved in serious games for language learning populated with ECAs that represent real speakers. The ECAs are also considered as intelligent programmes that are delivered through a two- or three–dimensional graphic entity, which has the ability to interact with a human user by text or speech, either on a web or standalone computer (Morie et al., 2012, p. 2). Numerous recent applications use a wide range of teaching and learning approaches, including an individual approach, and address a variety of language skills that learners can practise (Schulze and Heift, 2013, p. 258).

Today, more than ever before, teachers and learners can use various technologies to connect to various different VLEs to teach and learn different languages (Mancuso et al., 2010, p. 683). Various online learning websites, platforms and communities operate to create and use new learning scenarios for L2 (Escudero et al., 2013, p. 367), supporting individual learning, which is seen as a complement to traditional classes. Today, the development of high-quality applications that can be used in smart mobile devices, opening VLEs that are far wider in content and design than previously thought. These can offer users "digitally delivered immersive experiences, game-like virtual language learning environments, and real time, interactive and project-based language use collaborating with native speakers internationally" (Lindaman and Nolan, 2015, p. 17). In the following section, the theoretical background of computer games for L2 education will be introduced, towards placing *Virtual Reykjavik* into the fourth stage of ICALL.

2.2.2 Computer Games for Language Learning

Computer games for language learning provide learners with opportunities to practise conversations, because they support language output. Learners interact with virtual agents or other learners' avatars in the game through text chat or voice. They communicate with one another to reach a common ground in a conversation, i.e. to understand what actions and tasks they have to do, simply by having a dialogue. This process often involves various communication strategies, e.g. asking questions, providing answers, comprehension checks and clarification strategies (Peterson, 2010a, p. 73). Computers recently were considered as suboptimal for teaching the truly communicative aspects of language. Tschichold (2006) suggested that this should probably be better left to human teachers for the time being (p. 812-813). Since then, however, the technology has advanced greatly, while this chapter aims to argue the contrary. Today, computers can be seen as complementary tools to traditional language classrooms, giving learners another opportunity to use the target language in a written or spoken form in a different context and space when for instance playing computer game (Wattana, 2013). Even though computers cannot fully substitute a teacher in a classroom, they nonetheless have potential to assist in moving language education forward. This can be done by providing learners with opportunities for a real-time authentic interaction in the target language with either their peers or native speakers (Peterson, 2013, p. 128; Zhang et al., 2017). Learners can practise the target language while writing to chatbots or their peers, or speak to agents in the game *via* ASR. Recent intelligent applications provide various kinds of feedback either in a written form or recorded voice, or both. Learners can listen anytime to this information, which can gradually lead to improving some of their specific language skills (Zhonggen, 2019, p. 6).

With advances in technology, computer games have improved graphics, more sophisticated verbal and non-verbal behaviour of virtual agents, and the inclusion of AI to the system, which in turn helps to track the player's progress. Even though not all computer games serve the purpose of learning, some of them, e.g. *World of Warcraft*, support interaction in L2 with other learners involved in the game. This leads to practising the target language output amongst peers. Thorne et al. (2009, p. 811) conducted research between one Ukrainian and one American *World of Warcraft* player having a dialogue with 140 turns in English upon initial contact. Here, the Ukrainian player reportedly improved their English by corresponding (chatting) with a native English speaker. When learners experience a challenge, fantasy and curiosity while playing a computer game, they are more likely to be motivated to study (Malone, 1980, pp. 81-82). Przybylski et al. (2010) similarly observed that when learners achieve certain outcomes when playing, e.g. they socialise with other players, become immersed in playing, enhance their competence in certain skills, and are generally positively satisfied with the game (p. 163-164), they are also more likely to be motivated to use computer games for further study. Computer games that include a fun element can promote intrinsic motivation and engagement in learning activities, to help develop knowledge and language skills relevant to the game (Johnson et al., 2005, p. 311). Computer games also provide interim spaces in which language learners can take more risks and practise communicating without causing communication to break down. Selected examples of computer games below represent a category of serious computer games that were developed either for entertainment purposes but seemed to be also suitable for L2 practice, and for learning purposes of L2. These games shed light on the technological advancement of today, featuring a transition from traditional computer games to computer games with intelligent features.

One of the most popular computer games today is *World of Warcraft*[20]. It is a massive multiplayer online role-playing game (MMORPG) in a 3D virtual environment, that originated in 2004. However, this game was not originally designed for L2 education. Studies were conducted on the effects of using English as L2 (Sylvén and Sundqvist, 2012; Heathcote, 2012; Kallunki, 2016; Newgarden, 2015; Newgarden and Zheng, 2016) or Spanish as L2 (Rama et al., 2012) in conversations *via* text chat and voice (e.g. through Skype) between users who were native and non-native speakers of these languages. Similarly, the following examples, *Ever Quest 2*[21], *Ragnarok*[22], and the former project *Quest Atlantis*[23], belonged to a generation of MMORPG that represented a large online following, where thousands of players from different parts of the world engaged in game activities and tasks. For many players, this also had a positive side-effect of practising L2 skills. The common features in these online games are a 3D environment populated with animated agents of various kinds. Furthermore, there is active communication between players' avatars and other agents in the game. Here, players from diverse linguistic backgrounds meet in this domain for the purpose of entertainment. Non-native speakers of

[20] https://worldofwarcraft.com

[21] https://www.everquest2.com/home

[22] https://eu.4game.com/ro/#ro-classes-box-20-hash

[23] https://sashabarab.org/projects/quest-atlantis/

certain widespread languages (e.g. English or Spanish) especially benefit from this because they can practise them with other users. In games such as these, players collaborate, communicate with one another and negotiate meaning, i.e. discuss their actions, *via* text chat or speaking on Skype. Many native speakers of English participate in MMORPGs, often leading to creating an English-speaking playing environment. Other non-native speakers of English may also join, which often leads to the improvement of their L2 vocabulary (Bytheway, 2011). These computer games therefore attract L2 teachers and researchers to study progress in L2 learning. For this reason, they appear promising VLEs for L2 learning (Peterson, 2010b, p. 431).

I-FLEG is a 3D computer game, which is designed for learners of French as a second language. It is a research prototype, but currently integrated on an island in *Second Life*, where learners are engaged into playing. This part (I-FLEG) belongs to the category of serious games that "combines a situated, language learning environment with advanced artificial intelligence and natural language generation techniques which support user adaptivity and the automatic, context-aware generation of learning material" (Amoia et al., 2012, p. 24). It can be played by connecting to Allegro Island in *Second Life* and contains a game scenario in which the learner enters one of the selected houses. Each house represents one game unit, and explores different rooms containing different vocabulary according to the particular lexical level the learner seeks to learn (Amoia et al. 2012, p. 25). The learner has the form of an avatar with a first-person perspective, he/she can navigate freely in the game and interact with the virtual world by touching, moving or taking objects, or by sitting on them, and writing in a chat window that opens after clicking on objects. The learner gets a response in a form of displayed messages. In the current version of the game, the system only monitors player's interactions with objects. In the future, the goal is to integrate a dialogue system into the game in order to support situated conversational interaction. Feedback and other communication are provided through a head-up display (HUD) in a form of text messages. The HUD display presents various kinds of data without requiring the players to look away from their usual viewpoint. Here, the learners can see information about the current state of the game, the score and the elapsed time. The HUD is also used for navigating the learner through the game, communicating learning exercises, conveying feedback information, and for displaying the menu. The game provides learners with lexical and grammatical drills (Amoia, 2011, p. 3; Amoia et al., 2012, p. 25). The only agent in the game is the learner's avatar, which has a body of a person, enabling the learner to explore the game environment by walking around, sitting down and virtually touching objects. The avatar executes various restricted moves according to the learner's intentions and commands, allowing the user to act as if in real life. The limitations of the game are the lack of possibilities to practise communicative speaking skills and listening activities, because the feedback is usually received in a form of a written text on display and the prototype of the game is accessible only through *Second Life*. The main goal of the game is to teach vocabulary and help learners visualise words by finding the correct objects.

The Danish computer game *Mingoville*[24] has been designed for teaching English to two groups of learners, i.e. pre-school children (Preschool Program) and children aged 6-12 (Primary Program). It is marketed as a global educational resource with attractive design, that is easily accessible and has simple tasks (Meyer, 2013, p. 42). The game is populated with 2D animated characters that have the shape of animals in human clothes, who are part of a flamingo (bird) family of Pinkelton. These animated characters execute simple movements, such as opening beaks in a regular manner while speaking, closing and opening eyes and turning their heads left or right. This game does not have any speech recognition software implemented and for this reason is not possible to practise speaking. The learners navigate through the Mingoville country and build their English skills mainly on vocabulary while having fun playing. In the course of the game, the learners go through various kinds of activities, e.g. recognising words, building simple sentences with chunks of stones that they need to put in the correct order, and reading and listening to simple texts. Improving vocabulary and building simple sentences is the main focus of *Mingoville*.

The *Adventure German - The Mystery of the Nebra*[25] is a serious computer game for teaching German as L2 to adult learners. The game is set in a fictional environment and has a storyline. The players are on a quest solving a mystery; they interact in dialogues with other animated characters in the game while trying to gather knowledge (clues) to solve the mystery. The animated characters are represented by 2D drawings of people with simple body movements and facial features that move while speaking, walking or sitting down. Throughout the game, players need to solve various tasks on grammar and vocabulary. The main focus here is on grammar, vocabulary, practising written output and listening by solving various tasks. The speech recognition software is not part of this game and for this reason it is not possible to practise spoken language skills. However, the

[24] http://www.mingoville.com/

[25] https://www.goethe.de/en/spr/ueb/him.html

European project *Eveil-3D*[26] is trying to use the idea of this game and reconstruct a similar game for teaching German and French in a 3D virtual reality environment populated with 3D agents. It is a pilot project which seems to have ended in 2014 without release of a 3D game.

The following two computer games are examples of a successful fusion between gaming technology and language education. They use the tactical language and culture training approach, i.e. they are designed so as to teach learners various phrases and behaviour that is in the target culture. This also provides the foundation for *Virtual Reykjavik*. *The Tactical (Iraqi) Language and Culture Training System*, which is "a serious game platform that helps learners quickly acquire knowledge of foreign language and culture through a combination of narrative lessons that focus on particular skills, and interactive games to practise and apply these skills" (Johnson and Wu, 2008, p. 520). At that time, two courses were developed: *The Tactical Levantine Arabic* for a dialect of Arabic spoken in the Levant region, and *The Tactical Iraqi* for the Iraqi dialect (Johnson et al., 2005, p. 306). In the game, learners could practise spoken language skills, non-verbal communication and cultural knowledge relevant to face-to-face communication. Here, simulated conversations with non-player characters enables continual provision of feedback on learners' performance within each game-scenario context (Johnson, 2010, p. 175). The learners had their own avatar and communicated with other non-player characters *via* a combination of speech and gestures by selecting possible gestures with a mouse from a menu. One example of a gesture is the so-called palm-over-heart gesture, this means placing the palm of the right hand on the chest, which is common in the Arab world as a gesture of expressing thanks after shaking hands. The game used a game-based and task-based learning approach. It was developed in cooperation with Alelo[27] for the US Marine Corps Training and Education Command to prepare units for deployment overseas (Johnson and Wu, 2008, p. 522). For practising spoken communication skills, the game includes a speech recognition technology to support simulated dialogues, together with learning materials, such as paper-based study supplements, which include dialogues that learners can practise with other learners in order to improve their face-to-face conversation practice (Johnson, 2010, p. 180). It also supports written language skills, so that learners,

[26] https://www.eveil-3d.eu

[27] https://www.alelo.com/case-study/tactical-iraqi-language-culture-training-system/

who have got the time and inclination, can learn to read and write in Arabic (Johnson et al., 2007, p. 2).

Another example is the *Danish Simulator* (also known as *The Hunt for Harald*), which is an "interactive language and culture training system for Danish in the serious games category, combining game play and speech recognition for learning purposes" (Hansen, 2012, p. 1) also developed in cooperation with Alelo[28]. The aim of this game is to create an "immersive learning environment where the learner can learn and practise spoken language through the use of speech recognition aided by the added motivational factor of playing an actual computer game at the same time" (Hansen, 2012, p. 5). The capability of the ECA called Harald Bluetooth is to communicate multimodally. For instance, in order to give the learner positive feedback on pronunciation, the agent could produce gestures such as thumbs up (Hansen and Petersen, 2012, p. 187); or by clicking the button reading "Gestures" on a menu, the learner could shake hands or hug virtual people in the game, depending on which was more appropriate for the situation (Hansen and Petersen, 2012, p. 188). The role of non-verbal behaviour used in the game was two-fold. Firstly, to receive feedback from the agent, and secondly to allow Danish learners to express emotions by hugging other virtual agents in the game. The results from this study (Hansen and Petersen, 2012, p. 190-192) mainly discuss the pedagogical aspect of this game, for instance the importance of the presence of a teacher in a Danish language class where this game can be played, and the importance of a dialogue among students (users) in the Danish language classroom, where this game was tested. Hautopp (2014, p. 6) reported that while the students were playing the game, they also had the need to discuss language phenomena and sociocultural aspects of Danish language with other students, as well as with the teacher in the class. The above studies presented immersion of students in *Danish Simulator* as a way of communicating among themselves and with the teacher about linguistic sociocultural phenomena of Danish. The studies did not describe immersion of students from the point of view how they were playing and what features (e.g. game design, narrative plot) kept them busy exploring the game. The main approach used in *Danish Simulator* was communicative language teaching, which with the help of speech recognition, allowed learners to have a conversation with the virtual character Harald. The results from the user testing revealed that the game had enabled learners to become more familiar with spoken Danish, which helped them to participate more freely

[28] https://www.alelo.com/case-study/the-danish-simulator/

in conversations. Moreover, an external evaluation of this game showed that learners appreciate the anytime/anywhere availability of this platform, the pronunciation and conversation activities, allowing them to work in an individual and focused manner (Jensen, 2014, p. 15).

The examples of computer games here above represent relatively successful computer games where L2 could be practised. In these games, learners typically engage in individual and collaborative tasks while practising different language skills. The first four examples, *World of Warcraft*, *Ragnarok*, *Ever Quest*, and *Atlantis*, are MMOPRGs in which learners use natural language in text chat. Although these are not designed for L2 learning, they nonetheless practise the target language. Here, the learners use agents in the form of avatars (learners are impersonated into a selected agent) that have the shape of a human body and allows them to explore the game environment. However, the *I-FLEG* is only a module but similar to the MMORPG. This module is integrated into *Second Life*, which does not belong to the category of MMORPGs because it lacks a conflict and does not have set objective. In this module, natural language can be used in the form of text chat. Similarly, the learners use agents in the form of avatars that have the shape of a human body. Two further examples, *Mingoville* and *Adventure German*, however, belong to another category. These games were created for the purpose of L2 education and therefore are also deemed serious games. These do not support the use of natural language *via* an ASR because exercises are preprogrammed and learners need to accomplish tasks in order to proceed. The agents are 2D animated characters with simple body movements. On the other hand, *The Tactical (Iraqi) Language and Culture Training System* and *Danish Simulator* are serious computer games with more elaborate 3D agents. In these computer games, ECAs conduct more sophisticated movements, enabling the learners to interact face-to-face with them so that the player (learner) can see different facial and body movements on the computer screen that correspond with the context of what has been said. An advanced use of multimodal behaviour was presented in *The Tactical (Iraqi) Language and Culture Training System*, in which the verbal, non-verbal and cultural features implemented in the multimodal behaviour of ECAs enabled them to interact with other characters according to the tradition of the target language and culture. Most of the computer games above had limitations in the lack of agents imitating a sophisticated multimodal behaviour with authentic cultural features. However, *The Tactical Iraqi* had agents with some features of culturally specific multimodal behaviour; these were examples given by experts and taken from existing research. Such research has not yet been done for Icelandic, and therefore this thesis aims to describe the importance of observing native Icelandic speakers in a natural setting.

Despite the fact that computer games have increasingly replaced more traditional games as a leisure activity, they are now being explored as an attractive and effective method of teaching and learning language (Connolly et al., 2012, p. 661; Papadakis, 2018, p. 1). They represent a complementary tool to the traditional classes and fill in the gap between real and virtual world. Their popularity has increased amongst young teenagers and adult learners, who enjoy getting immersed into the ever-improving virtual learning environment of computer games and other technologies for language learning (Blyth, 2018). Computer games appeal to learners because they engage, entertain, and motivate them, while also promoting learning (Johnson et al., 2005, p. 306; Royle and Colfer, 2010, p. 13) as they increase students' motivation for learning. As Royle and Colfer (2010) argue, learners can acquire particular skill habits and affordances through gaming (p. 17). Games are responsive to the players' choices and actions, such that they "can inspire the loftiest form of cerebral cognition and engage the mot primal physical response, often simultaneously" (Salen and Zimmerman, 2004, p. 1); they represent a dynamic system – a system of possibilities, which players inhabit and explore (Salen and Zimmerman, 2004, p. 2). Even though traditional games have been part of language education for decades, Meyer (2009) suggests that computer games have never had a central position in foreign or second language education because gaming as such is, from the viewpoint of some teachers, still being claimed as disruptive and antithetical to schooling (p. 716). Despite this, Meyer suggests that serious games have great potential in foreign and second language education. Serious games have the ability to provide learners with specific learning environments that contextualise knowledge and immersive experiences, both outside and inside of formal education (Meyer, 2009, p. 715). In order to combine computer games with language learning, one needs to understand first how language learning proceeds in traditional classes. Meyer offers an insight into this process, which is skill based rather than content based, as in learning subjects such as history. Learning a language is different from learning any other subject in the curriculum, because, as Meyer (2009, p. 715) says:

> *it combines explicit learning of vocabulary and language rules with unconscious skill development in the fluent application of both these things… as it implies that (the learners) should be able to master both grammatical knowledge and fluency, the latter being often difficult to provide in classrooms where a couple of lessons a week may fail to provide the meaningful exposure to the foreign language required for learning.*

Computers are therefore ideal for enabling users to both learn vocabulary and grammatical rules, with the help of specially designed virtual learning environments. In this sense, computer games may well help learners to increase the exposure to the target language. With the new generation of computer games that include better visual graphics and technologies, such as speech recognition, there is a greater potential for language instruction and acquisition, because they allow the practice of speaking skills in the target language. This is something that CALL has been aiming for since its beginning, but which is now becoming reality in ICALL. The real-time exchange of meaning is especially important in facilitating communicative competence in the target language. Participation, enjoyment, and reduction of anxiety or minimising low self-confidence (Peterson, 2010a, p. 74) are amongst those things that many learners can benefit from playing computer games for learning language. Conversing with virtual characters in any computer at any time is no longer a remote possibility and using computer games as virtual learning spaces to enhance communicative competence in the target language and culture is on the near horizon, if not already here. The following section discusses important language learning approaches in connection with computer games.

2.2.3 Language Learning Approaches and Computer Games

When designing computer games for language learning, one needs to consider the relevant instructional approaches that support the development of learners' language and culture skills. The communicative approach together with task-based learning are frequently used methods to support practising oral language skills in computer games. The instructional approaches also provide the foundation for game-based learning. They moreover inform about the roles each player should have within the game scenario, i.e. which tasks should be completed in order to practise context-related language with other players in the game. In addition to these, the multimodal approach in language learning (Dale, 1969) plays a very important role in computer games, because it informs about all the important sensory channels, such as vision and audio, that are involved in the learning process of L2 in computer games. All of these approaches are discussed in a more detail below.

2.2.3.1 The Communicative Approach. In the traditional view, the Communicative Approach is about teaching learners' communicative competence (Hymes, 1972, 1992) in order to cope with everyday situations language-wise (Littlewood, 1981, p. ix), and to be exposed to meaningful interactions in the target language. The emphasis is on target language-like communication, where the message is the main concern rather than the grammatical form (Krashen, 1982, p. 17). This approach was popular in 1970s-1990s and was originally used for teaching languages in classrooms to enhance the oral language competence of learners (Richards, 2006, p. 9) who wanted "to be able to communicate socially on straightforward everyday matters with people from other countries who may come their way, and to be able to get around and lead a reasonably normal life when they visit another country" (Littlewood, 1981, p. ix). This approach was seen as an alternative to the previously form-based instructional approach, such as the Grammar Translation method, which mainly focused on learning grammar rules and translating texts from books. While previous instructional methods were teacher based, this new approach focused on learners and their need to practise oral language. This view opened up new perspectives on language: language should not only be considered in terms of structure (grammar and vocabulary) but also in terms of its communicative functions and what people do when they speak (Littlewood, 1981, p. x). This included developing syllabi built on communicative functions, such as greetings, requests, apologising, surprise, and notional categories such as time, sequence, quantity, and location (Toth, 2013, p. 3). Nowadays, it is considered an output-based approach in language education, enabling learners to practise language skills via certain communicative tasks, thus making the practice more authentic.

The Communicative Approach plays an important role in ICALL because it informs about the need to practise the target language via speaking. In practice, learners adopt conceptual language chunks consisting of a group of concepts. When learners routinely express certain communicative intentions, they activate these chunks and consequently use them in speaking (Kormos, 2013, p. 3). On the other hand, learners are also exposed to certain phrases that virtual agents repeatedly use in particular conversational scenarios. These phrases become chunks that learners associate with a particular context in which the agents use them. In this way, the learners can more likely use these phrases, or chunks, in a similar communicative task in real life. For this reason, computer games with implemented spoken and textual dialogue systems are suitable for practising the oral language skills of the learners. Learners become engaged in solving tasks to use the previously heard chunks when speaking to virtual agents. These tasks can provide realistic and meaningful opportunities for language production, and in combination with game-based learning, they may support fluency and communicative competence by permitting learners simulate or act-out real-life situations (Meyer, 2009, pp. 715-716). In this way, computer games can provide a useful VLE that gives the learners the opportunity to produce the target language output, participate in spoken and written conversations, enjoy playing while learning, and reduce anxiety and low self-confidence due to unproficiency in the target L2 (Peterson, 2010a, p. 74). The Communicative Approach has been the main mode of foreign language instruction for the past decades and remains one of the crucial approaches even in today's ICALL, which is output driven.

The general aim of the communicative approach is to assist learners with using the target language in real communications for establishing and maintaining contact, and for exchanging general information. This approach furthermore supports the theory of communicative competence (Hymes, 1972; Hymes 1992; Ellis 2011), which will not be discussed here. *Virtual Reykjavik* supports the communicative approach to language learning. The following section informs about the importance of task-based learning in computer games.

2.2.3.2 Task-Based Learning. Task-based language learning is an offshoot of the communicative approach. In an effort to increase communication in the classroom, the nature of tasks was re-examined with a view to create tasks that required the use of enhanced natural language in the classroom (Skehan, 1996). In traditional instruction, Skehan (1996) proposed a framework for implementation of task-based learning. He built on the developments in cognitive psychology, which supported a dual mode perspective for language processing that uses a collective action of verbal and non-verbal mental systems to process imagery and linguistic information (Clark and Paivio, 1991, p. 150). On the basis of task characteristics, Skehan (1996) characterises a strong and a weak form of task-based instruction (p. 39). The former uses tasks as the main unit of language teaching, and that everything else is subsidiary. The latter embeds tasks in a more complex pedagogical context, in which focused instruction precedes the use of tasks, and another type of instruction, which is based on the results from tasks, follows them. This weak form of task-based instruction is very close to general communicative language teaching (Littlewood, 1981; Skehan, 1996). The task-based approach is also implemented in educational games, because it helps to stimulate learners to proceed and solve problems in the game, while the communicative teaching enhances the development of speaking skills. Communicative tasks engage learners in comprehending, manipulating, producing, or interacting in the target language while attention is principally focused on meaning rather than form (Nunan, 2001, p. 10). However, a task does not only require a learner to act primarily as a language user and give focal attention to message conveyance, but also it allows the learner to decide what forms of language to use according to the given context when completing the task (Ellis, 2011, p. 5). By letting learners solve tasks that require choosing the appropriate verbal interaction, language learning can be stimulated (Robinson, 2011, p. 1).

Modern computer games involve learners into performing various tasks while practising various different language skills, e.g. listening, speaking, reading, and writing. Warschauer and Healey (1998) considered task-based learning as an important approach in providing the learners with authentic VLEs (p. 58). Today, computer games implement task-based learning strategies to enhance learning of real language use in simulations. One example is Bado and Franklin's (2014) *TraceEffects* game that engages learners into cooperative learning and the exchange of information, to promote a face-to-face interaction and the use of the target language. Another example is a more advanced computer game which uses communicative tasks to engage learners with animated characters representing local people, *The Tactical (Iraqi) Language and Culture Training System* (T(I)LCTS) (Johnson and Wu, 2008, p. 520). In this simulation of real environment, American soldiers practised communicative and cultural skills. Learners who had training in this simulation of a real environment were perceived differently in reality by local people, which shows that this system helped them transfer culture specific communicative skills from virtuality to reality. Apart from language and cultural skills, this game also enabled practising words and phrases, and to complete exercises and quiz items that require speaking and understanding of spoken language (Johnson, 2010, p. 178). The difference here was that the virtual characters in the T(I)LCTS simulation were endowed with realistic non-verbal behaviour, such as gestures (both "do's" and "don't's") that represented local cultural norms and etiquette of politeness, to help learners accomplish the social interaction tasks successfully (Johnson et al., 2004, p. 2). Evidence suggests that realistic communicative tasks in computer games may relatively easily help learners transfer their communicative skills into real-life situations, even though they have experienced them only in a computer simulation (Peterson, 2010a; Barrett and Johnson, 2010). One of the aims of *Virtual Reykjavik* is to enhance transfer of spoken communicative skills in Icelandic. The following section introduces game-based learning.

2.2.3.3 Game-Based Learning. There are various types of digital and computer games, each developed with a specific purpose or function depending on the target audience. According to Connolly et al. (2012), certain digital commercial games, such as *Mario Brothers* and *Grand Theft Auto*, serve as entertainment, but other computer games and serious games are designed for learning and changing behaviour not only in business, industry, marketing, healthcare, and government, but also in education (p. 662). *Virtual Reykjavik* belongs to this group of computer games. When computer games are used for learning a language and culture, this approach can be characterised as (digital) game-based learning (Prensky, 2001, p. 4).

Game-based learning is according to Prensky (2001) another way of teaching and learning, a radical idea, aimed primarily at individuals (p. 4) which should be about fun and engagement (p. 16). This approach should include educational content that is useful in real life (Prensky, 2001, p. 124) in order for learners to get familiar with the real environment. Such an approach should inform about how to keep the players (learners) engaged in both playing the game and the learning state simultaneously (Prensky, 2001, p. 125), while maintaining a story line to motivate them to finish the game (Prensky, 2001, pp. 132-133). In addition, game-based learning supports a learner's intrinsic motivation in language learning rather than non-gaming material (Francoisi, 2011, p. 11). It supports individual motivation, which is affected by the level of challenge, curiosity, control and fantasy in any learning situation (Pourabdollahian et al., 2012, p. 260). Furthermore, Prensky (2001) suggests that computer and video games, like nothing else, are a conjugated form of fun and play, and rules and goals. Computer and video games are interactive and adaptive, produce outcomes and feedback, have ego gratification through win states, give adrenaline through conflict/competition/challenge/opposition, spark our creativity by solving problems, create social groups through interaction and have representation and story that gives us emotion, to name only a few (p. 106). Even though there may be limitations in the level, quality, technical and logistic issues of the game, or in the learner's skill and desire to play, game-based learning seems to affect many areas, enhancing learners' performance while engaging in tasks.

In view of the above, game-based learning (not only of digital games) is a form of interactive content that often allows the player to learn in the first-person rather than from a third person perspective, using role play in the process (Chee, 2016, p. 52). According to Tobias et al. (2014), it can moreover allow for near and far transfer of tasks performed in computer games to external tasks performed in real life. They also argue that these tasks must overlap, otherwise the transfer may be very weak (p. 485). Virtual agents can help with enriching the interactive content of computer games, but it is technically very challenging to endow them with natural language and behaviour associated with particular communicative tasks. The learners can repeatedly observe how virtual agents speak and listen to their speech, which may help them produce similar utterances in a similar conversational context. The following section discusses the use of multimodal approaches in language learning. It informs about which sensory channels are involved in the learning process.

2.2.3.4 Multimodal Approach in Language Learning. The multimodal approach in language learning builds indirectly on Dale's (1969) model of *Cone of Experience*. This model supports the claim that the more sensory channels learners involve in a learning process, the more information and knowledge they retain. In this view, sensory teaching, or sensory education, is often mentioned when different modes (senses) are employed in language learning (Lightbown and Spada, 1990; Minogue and Jones, 2006; Felicia and Pitt, 2009; Pekarova, 2010; Abraham and Leiss, 2012; Ramirez, 2011; Covaci et al., 2018). This approach can also be considered from two perspectives. According to Jewitt (2013a), the first perspective refers to the language in a face-to-face interaction, which is communicated through gesture, gaze, facial expression, shifting of the body posture, which may be specific to a particular social or cultural context and the resources available to people at the time of making meaning, i.e. producing and perceiving language (p. 2). The second perspective refers to various educational tools, such as textbooks, maps, digital objects, models that include a mix of images, colours, texture, writing, music, and a spoken word through which different senses are employed in the process of learning, allowing learners to practise different language skills (ibid., p. 2). Combined, these two perspectives create a multimodal language learning environment which involves multiple senses, or modes, that learners use for making meaning of the information about the L2 (Farías et al., 2007; Early et al., 2015; Godhe and Magnusson, 2017). Each learner differs in their learning abilities and preferences; the more senses available in the instructional and task design of a particular learning environment, the greater improvement that can be achieved on an individual basis (Arbuthnott and Krätzig, 2015, p. 2).

In the context of this thesis, the multimodal approach in language learning is used in connection with embodied digital interfaces, namely the ECAs, in a 3D VLE. *Virtual Reykjavik* enables access to such an environment that is multimodal, because it uses spoken and written texts, images, sounds, and animated virtual characters with multimodal behaviour. Even though each virtual environment has its own principal mode of communication, ranging from text, audio, video, and graphics (Palomeque and Pujolà, 2018, p. 177), the visual and auditory senses are one of the main modes involved in language learning in computer games (da Silva, 2014, p. 158). There are also other senses that are involved in the learners' action of doing something, for instance receptive modes (reading and listening) and expressive modes (writing and speaking). The learners follow instructions and receive feedback while speaking and writing to other players or non-player characters in the game (da Silva, 2014; Rudis and Postic, 2017). The learners may also receive an additional sensory input by listening to conversations and reproducing the spoken language input by speaking. In this way, they can practise pronunciation in the target language (Rankin et al., 2006, p. 4).

Embodied conversational agents that 'know' how to use their bodies and language in a conversation play an important role in achieving the learning goals of *Virtual Reykjavik*, i.e. to practise authentic Icelandic conversation. As Jewitt (2013a) puts it: "The characters that have the most modes of communication are the key to game success – especially those with the potential to speak when approached by the player/avatar" (p. 24). In *Virtual Reykjavik*, the virtual agents are part of a multimodal interface and use multimodal behaviour (eye gaze, facial expressions, gesture, body posture) when speaking to Icelandic learners. Multimodal features of a spoken communication system are utilised frequently in real life between L2 learners and native speakers, to achieve a mutual understanding and linguistic development (Wild, 2015, p. 50). Similar features should be used by ECAs when speaking to L2 learners to prepare for real language use. Through various activities in such games, learners communicate with other agents by which they learn both to use the language adaptively in a game-solving situation (Zheng et al., 2012, p. 358) and to facilitate a face-to-face interaction with speakers of a particular culture (Bédi et al., 2016, p. 42). Since a multimodal approach is used in *Virtual Reykjavik* to support individual and individualised learning, the following section will be dedicated to this topic.

2.2.3.5 Individual Learning Approach. According to Prensky (2001), computer games for language learning are primarily aimed at individuals. They support individual and individualised learning, in that they allow each learner to proceed with playing by their own pace, and thus make the game more individualised to the needs of users. The benefits of individual and individualised learning were discussed as early as 1988 by Weinstein et al. This study claimed that collecting data from learners in classrooms and creating individual profiles may help create a more individualised learning experience, by suggesting learning issues that need more attention (Weinstein et al., 1988, p. 35). A similar approach can also be used in computer games for learning purposes, because several games allow tracking and storage of data from users, that can be used for future improvement and learning assessment (Medler, 2009, p. 181). Today, computer games have become part of contemporary culture, and with respect to enhancing L2 skills, they have also become educational tools based on an interdisciplinary approach to ICALL. Intelligent systems enabling storing, evaluating and structuring of data for feedback purposes are part of the design for individual learning and creating of individual lessons, with individual measurement of progress (Schulze and Heift, 2013, p. 258). When user profiles in computer games are personalised according to learners' achievements, these achievements often contribute to promoting learning and motivation (Almeida, 2012, p. 4). The individual approach to learning helps to assess goals of each player, and to track achievements and failures that are part of their learning process. For instance, when building communicative skills in the *Tactical Iraqi* application, learners can switch from one task to another, or from one game scenario to another. Learners can thus determine which skills they lack and focus their attention on improving these in the course of game-play that includes such tasks (Johnson, 2010, p. 178). Flexibility in time is another advantage of individual learning. For instance, players in the *Danish Simulator* appreciate the 'anytime, anywhere' availability of this platform, enabling flexible access to exercises for pronunciation and conversation activities, and thus working in an individual and focused manner (Jensen, 2014, p. 15).

Individual learning can also be characterised from many other perspectives. In the context of computer games for educational purposes, Kao and Windeatt (2014) suggest that it is motivation and categorisation of players that are the driving force behind learning. Intrinsic motivation affects learner's inner motivation, the willingness of oneself to achieve or learn something. Extrinsic motivation comes from outside and is a stimulating factor that motivates the learner to play the game by, e.g., collecting points and reaching a higher level than their peers, or receive some kind of acknowledgement that shows success in learning. This type of motivation stimulates the attitude, willingness, anxiety and the actions of a learner towards learning (Kao and Windeatt, 2014, p. 3). Kao and Windeatt (2014) furthermore argue that although the inner perspective is not widely accepted as a measure of academic performance, it is the personal sense of competence or success, which serves as the evidence of self-determination and self-regulation on the part of learners who are actively assessing their individual achievement or progress (ibid., p. 2). In this sense, computer games for language learning can stimulate learners' inner motivation, for instance, by providing feedback on progress and achievement (Pourabdollahian et al., 2012, p. 260), since even failure may motivate learners to do better and achieve goals. Learners execute various tasks in language-learning games, which makes them engaged in task-solving process, and thus learning. According to Dörnyei's (2003) model of motivational task processing, learners follow a certain 'action plan' that had been set up in the task; they continuously process "a multitude of stimuli coming from the environment and of the progress made toward the action outcome" (p. 15). Language-learning computer games furthermore provide alternative educational exercises outside of traditional classes, are flexible in time, and stand as a choice of one's individual learning styles and preferences (Kao and Windeatt, 2014, p. 4). They provide a complete playing interactive environment by engaging learners into a visually appealing environment, i.e. simulation of reality and embodiment of fantasy, have ultimate goals to motivate learners *via* fun elements and visual feedback, all of which can provide learners with an immersive experience and a sustained interest in the game (Morton et al., 2012, p. 12).

There are also many internal and external factors that affect the learning process of an individual. These factors interact with each other and affect each learner differently. As Tennyson and Breuer (2002) and Tennyson and Jorczak (2008) suggest, motivation, attitude, willingness, anxiety and action to learn are only a few components of a more complex system that represents the learning process of an individual. Based on their research, Tennyson and Breuer (2002) propose a comprehensible model of integrative information processing (Figure 6).

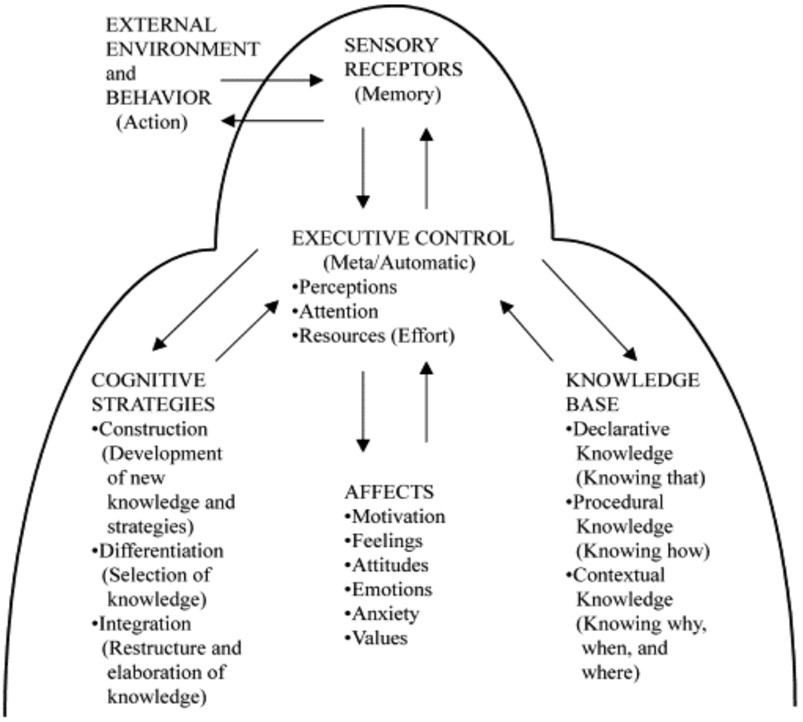

EXTERNAL ENVIRONMENT and BEHAVIOR (Action)
SENSORY RECEPTORS (Memory)
EXECUTIVE CONTROL (Meta/Automatic)
•Perceptions
•Attention
•Resources (Effort)
COGNITIVE STRATEGIES
•Construction (Development of new knowledge and strategies)
•Differentiation (Selection of knowledge)
•Integration (Restructure and elaboration of knowledge)
AFFECTS
•Motivation
•Feelings
•Attitudes
•Emotions
•Anxiety
•Values
KNOWLEDGE BASE
•Declarative Knowledge (Knowing that)
•Procedural Knowledge (Knowing how)
•Contextual Knowledge (Knowing why, when, and where)

Figure 6: Interactive cognitive learning and thinking model (Tennyson and Breuer, 2002). Reproduced with permission from Elsevier with license number 4780961380863.

This model uses complexity theory to describe one perspective on language learning. It views learning "as the result of complex and non-linear interactions of variables internal and external to the cognitive system of a learner" (Tennyson and Jorczak, 2008, p. 5). The model also states that its various components constantly interact with each other in all directions. In combination with sensory information from an external source, they can help improve the individual's knowledge base. Graphic design in computer games supports visual senses, sound and other audio affect hearing senses, and moving objects around in the game environment simulates tactile senses. The more objects with physical/living characteristics are implemented into the game environment and interact with the player, the more realistic it becomes for the learner to play, which eventually leads to a more natural experience in virtuality (Bossomaier, 2012, p. 32). The above interactive cognitive learning and thinking model informs about the complexity of this issue and suggests that learning *via* computer games requires the inclusion of many senses to successfully promote learning of L2. There are also various categories of players (gamers) that the individual learning approach helps to inform about. For instance, Edwards (2004) suggests that gamers have distinct goals in playing games, and for this reason, there are three types of players: the 'gamist', who seeks competition and challenge; the 'narrativist', who is satisfied if the playing session results in a good story; and the 'simulationist', who likes exploring and experiencing his/her own virtual "pocket universe" (p. 1). Player types are thus motivated by different aspects of the game, for instance gamers like to solve problems and achieve goals, narrativists are motivated by the game context, and simulationists by game interactions and representations (Tennyson and Jorczak, 2008, p. 18). Although there is no empirical proof offered for these three categories of game players, the player traits nonetheless play an important role in determining the effectiveness of games (Perez, 2007, p. 288). For instance, games offering explicit goals may motivate gamists more than other types of players (Perez, 2007, p. 288).

This thesis is a part of an effort to developing a computer game to support the learning of spoken language and culture in a computer game. The game is populated with ECAs that use Icelandic language and change their body behaviour when interacting with learners. The learners create individual profiles that can store data about their own progress. They play a game by following instructions and solving tasks. While playing, they employ various senses, such as visual, audible and partly also tactile senses (using the keyboard for moving around with one's avatar). This game supports individual learning, by giving learners the opportunity to practise, review and re-perform tasks that may help them improve their spoken communicative skills. This game is not designed as a multi-player game. The current version allows learners to play the game individually, and learners can experience a realistic interaction with ECAs (Ólafsson et al., 2015; Bédi et al., 2016). The theoretical foundation for designing a realistic interaction in computer games for language learning is presented next.

2.3 Designing a Realistic Human-Agent Interaction

This section presents a theoretical background to designing a realistic human-agent interaction that underpin the effort for designing a realistic human-interaction in *Virtual Reykjavik*. In face-to-face interactions, people use oral language to exchange information. They use mutual knowledge, beliefs, and assumptions to reach understanding. They moreover update one another on new information when necessary. This process is called negotiation of meaning, or grounding (Nakahama et al., 2001; Vandergriff, 2006). Agreeing, answering, assessing, responding, requesting, are all utterances used to achieve a mutual understanding, or a common ground, between participants in a conversation (Schegloff, 1992, p. 1300). Participants can use these utterances to reveal what they find problematic in the previous talk of the other speakers and signal that they have misunderstood something. In this case, CRs are often initiated. Toward the end of this section, grounding in a face-to-face interaction will be discussed from a multimodal point of view. The second last section will discuss the use of multimodal CRs by ECAs. The last section will focus on discussing the terms fidelity and feedback that are part of a simulation of real-world scenarios.

2.3.1 Technology Enabling Spoken Interaction in Computer Games

One of the necessary components for a realistic human-agent interaction in computer games is technology which enables spoken interaction between human users and virtual agents. Integration of automatic speech recognition (ASR) systems into computer games can enable players to hold a spoken conversation with virtual agents by speaking to a microphone and listening to what the agents say. The players' speech input is broken down into individual sounds that are recognised with the help of an algorithm into the most probable words, that are then transcribed into text. This speech dialogue between players and ECAs is operated in real time (Zouari and Chollet, 2008, p. 10522). The area of including ASR has prompted another area of research that deals with natural language processing (NLP) and interactive language learning through speech-enabled virtual scenarios. By holding an oral conversation with virtual agents, the learners can contextualise specific communicative tasks and eventually practise the target language in a computer game. This can be done in tasks that include a series of questions. Each task may represent a specific session where learners can practise different language skills in a given situation (Morton et al., 2012, p. 2). For instance, in *Virtual Reykjavik*, learners can practise oral language in different tasks, each task with a specific conversational scenario. The virtual agents will use a variety of clarification strategies that are specific to the conversational context. Currently, two implemented CR strategies were implemented, the Interjection Strategy *Ha?* (*Huh?)* and the Ellipsis *Hitt húsið (Hitt húsið)*. The ASR system enables the learner to practise the target language output by speaking. For comparison purposes, certain examples of computer games that have an integrated ASR system in order to enable learners to practise spoken language, are discussed.

One of the first examples of interactive systems is *Subarashii*, which used a prototypical ASR to offer spoken-language exercises in Japanese to beginning learners, who were native speakers of English (Bernstein et al., 1999). The learners play the game in the form of missions; they solve tasks in the context of an adventure game. Depending on the encounter with a person, the learner may initiate turn-taking in the course of a dialogue. This system was an answer to the early efforts of integrating technology into the interactive pedagogical design of CALL, which prompted the development of further applications and systems, such as the interactive multimedia computer system known as *Conversim* (Harless et al., 1999). *Conversim* incorporated speech recognition and digital video technologies to practise spoken Iraqi with virtual characters in real time, on a CD-ROM. This was a precursor to the creation of *The Tactical Iraqi* (TILCTS) serious computer game (Johnson et al., 2005, p. 306), which provides the basis for *Virtual Reykjavik*.

Interactions contribute to the development of learners' language and culture skills since it enables them to be active in this learning environment, solve tasks and contribute with their own language input into the interactions in the game. In order to enable a more realistic interaction between learners and virtual agents in the game, it is important that an ASR is integrated. *Virtual Reykjavik* similarly focuses on including spoken interactions in Icelandic between learners and virtual agents and uses the Google automatic speech recognition system for Icelandic. In this way, natural language can be used in a human-agent interaction. The following section focuses on natural language and discusses it from the point of view of a complex phenomenon. In order to better understand this phenomenon, complexity theory will be used. It will shed light on how multimodal features of natural language can be used for modelling a realistic human-agent interaction in computer games for language learning.

2.3.2 Language as a Complex Phenomenon

This section describes natural language from the point of view of language as a complex phenomenon. Here, the complexity theory will be used to help understand how multiple elements interact with each other, how they organise themselves into structures, adapt and undergo some kind of a spontaneous self-organization (Waldrop, 1992), which can be used for modelling a realistic human-agent interaction. Linguists studying the structure of language and AI researchers trying to model processes of thinking in computers use this theory to understand the complexity of a problem (Waldrop, 1992, p. 71). In the context of this thesis, it will be used to discuss the processes involved in producing and perceiving spoken CR utterances. Jörg (2011) suggests three complexities in this process: (1) the process of conveying the information, (2) the process of understanding the information, and (3) the process of those two to find a common ground (p. 208). For instance, when using spoken natural language in a face-to-face interaction, each participant has their own perception of meaning of an utterance/message conveyed by the other speaker. Based on the shared situational contextual and information around them, and the way in which the message has been conveyed, the spoken utterance should be perceived in a similar way by each participant in the conversation. If this has been achieved, then a common ground has been reached. Producing and receiving (or perceiving) messages is a process which happens in human communication. This process is part of human social skills (Lane et al., 2013, p. 2). By trying to understand the complex process involved in this situation, and by finding the order in the 'chaos', the information for creating a realistic human-agent interaction becomes much clearer. The main study in this thesis focuses on finding a common pattern in the production of multimodal cues in CRs, that will be used for creating multimodal models and implementing them into the conversational behaviour of ECAs.

Apart from analysing spoken language utterances, complexity theory is also used in the area of L2 education. For instance, Filipović (2015) suggested that every research phenomenon, including natural language, "should be analysed in all its complexity, made out of background information, agents and their interactions" (ibid., p. 31). Two approaches have been developed to support a complexity-driven language research: the micro-complexity driven approach and the macro-complexity driven approach (Filipović, 2015, p. 35). The former analyses language structure and forms, and the latter focuses on the dynamics of interaction between language and society. In the view of the macro-complexity driven approach, Filipović (2015, p. 37) suggests that use of language is culturally bound to language communities:

> *Regardless of the fact whether these communities are formal or informal, they are always dependent on cultural knowledge, cognitive cultural models and ideology which define their intra-group and inter-group relationships, hierarchies or lack of them, with all the socio-cultural baggage implied by it (politeness principles, conversation patterns, styles and registers, as well as implied or explicit balance of power and/or inequality with overarching consequences in the community members' lives both within their group and when facing other groups they interact* with.

All of this has serious impact on both how we act when we speak, i.e. how we employ our spoken and body language to convey a message, and how we perceive and understand communicative situations, verbal repertoires and conversational strategies used in a specific situation and by a specific community or culture (Filipović, 2015, p. 44). This approach points at different ways how humans communicate in different cultures, and that a face-to-face interaction between humans coming from different cultures may include elements that may not be the same for all speakers. This is important to be aware of in the context of L2 education. Learners should get familiar with ways how native speakers of the target language speak, what verbal and non-verbal features they use in certain phrases and conversational scenarios. This is furthermore supported by Larsen-Freeman (2013) who uses complexity theory to demonstrate that learners develop rather than learn another language. They actively transform their linguistic world and create their own patterns with meanings and uses. In this way, they gradually extend their system with alternative forms to their native language, to which learners have already made a neural commitment (Larsen-Freeman, 2013, pp. 2-3). Learners dynamically adapt their language resources to the context. In this process of language development, by letting learners revisit the same territory again and again, and by increasing the exposure to language at certain levels and scales, Larsen-Freeman (2013) suggests that learners can most likely then register particular language patterns in memory. This view is developed in this way: their first language (L1) acts as a filter on perception, and while certain forms are uniquely associated with a meaning or with a use, it is those forms with a unique meaning that will be easier for the learner to notice and to acquire, as opposed to other structures that have more meanings and learners are unfamiliar with (Larsen-Freeman, 2013, p. 4). This means that the frequency of exposure to patterns in language promotes learning (Larsen-Freeman, 2013, p. 4). This view on language development is relevant in the area of ICALL, where learners are given the possibility to actively interact by, e.g., spoken language and learn about the target language and culture by revisiting exercises in the computer game and train themselves in particular skills that the game offers while getting feedback. As Lane et al. (2013) suggest, “repeated practice opportunities with feedback are an essential component in the development of expertise” (p. 2). The same can be applied in the development of novel communicative skills.

To summarise, Filipović’s (2015) macro-complexity driven approach and the use of Larsen-Freeman’s (2013) complexity theory in L2 education suggest that language is both a complex phenomenon consisting of multiple verbal and non-verbal cues, and a function of communicative actions that are bound to certain cultures and communities. Body language, physical distance and cultural proxemics that the speakers take up to one another are part of spoken language interactions and need also to be investigated in this context (Norris, 2004, pp. 18-19). In order to help learners of Icelandic to develop their language resources, *Virtual Reykjavik* needs to present language in its complex way and how it is produced, including all multimodal cues. In order to create realistic human-agent interaction in a virtual setting, it is important to study not only the forms of talk but also the different multimodal strategies that accompany verbal interaction that help convey its message. The following section discusses embodied cognition which sheds light on how humans produce and perceive signals in a conversation *via* different senses (modes).

2.3.3 Embodied Cognition in Face-to-Face Interaction

The theory of embodied cognition plays an important role in this theoretical framework because it informs about the multimodal processes that take place in a face-to-face interaction between human speakers. The source of inspiration is Lakoff and Johnson's (1999) work. They argue that reason, and therefore mind, is fundamentally embodied (p. 3 and p. 17). Wilson and Foglia (2011) similarly advocate that the mind is embodied because we perceive meaning in spoken communication through all of our senses. Human speakers employ their face, eyes, hands, and the rest of the body both to convey and perceive meaning. These processes are unconscious and are based in our thought, operating beneath the level of our cognitive awareness (Lakoff and Johnson, 1999, p. 10). It means that humans are unaware of employing all of their other modalities when talking and listening to speakers. This suggests that all the following processes are employed in face-to-face conversation without humans being directly aware of them:

- Assessing memories relevant to what is being said;
- Comprehending a stream of sound as being language, dividing it into distinctive phonetic features and segments, identifying phonemes and grouping the into morphemes;
- Assigning a structure to the sentence in accord with the vast number of grammatical constructions in the native language;
- Picking out words and giving them meanings appropriate to context;
- Making semantic and pragmatic sense of the sentences as a whole;
- Framing what is said in terms relevant to the discussion;
- Constructing mental images where relevant and inspecting them;
- Filling in gaps in the discourse;
- Noticing and interpreting your interlocutor's body language;

- Anticipating where the conversation is going; and
- Planning what to say in response.

(Lakoff and Johnson, 1999, pp. 10-11).

O'Connell et al. (1990) have also observed that "speaking is embedded as an occasional event in a continuous stream of non-verbal behavior" (p. 365). They understood communication as continuous and argued that even though verbal switching (turn-taking) among participants occurs, the communication is held by an invisible bonding which consists of eye movement (e.g. looking), gestures and touch, with which the participants uphold conversation (pp. 365-366). All utterances are then supposed to be produced in a similar way, i.e. in the process of uttering words or sentences structured in a turn-taking pattern. Once again, human speakers usually do not notice such non-verbal processes in a face-to-face interaction, because they are inaccessible to their conscious awareness and control (Lakoff and Johnson, 1999, p. 38). They can, however, be only partly observable when such interactions are recorded on a video, studied and thoroughly analysed. The theory of embodied cognition then helps us to understand how the body and the mind is involved in the production of these utterances. Different aspects of the body will be examined when speakers cognitively process a CR. This means that the speakers are aware of verbally producing a request to clarify what has not been understood but are unaware of the non-verbal cues they have produced, along with the verbal ones.

On the other hand, humans may also use body movements consciously, for instance when they intend to bring something into visual consciousness of others for the purpose of seeking attention. In such situations, they can use their eyes, head and the rest of the body in various ways (Wilson and Foglia, 2011, p. 1). Specific facial expressions or hand gestures are very common here. In this view, Jacobsen (2015) follows Goffman's analysis on social interaction and suggests that "when an individual is in the immediate physical presence of other people, he or she will unavoidably seek to control the impression that others form of him, or her, in order to achieve individual or social goals. The individual will engage in (what is called) impression management" (p. 68) and will do so in two ways. By conscious information, i.e. production of verbal and non-verbal symbols; and by unconsciously emitting information that consist of signs and expressions (Jacobsen, 2015, p. 68-69). From this point of view, body movements can be used both conscious and unconscious ways. In this thesis, the latter realisation is the focus in CRs.

In the field of computer science, embodied cognition also plays a very important role in that it helps designers of virtual agents and synthetic agents (robots) to understand the role of the body in spoken interaction. To understand how body and speech are interconnected, researchers conduct studies analysing multimodal behaviour in natural language conversations. This helps them to build systems that reflect such behaviour in various virtual interfaces. ECAs also belong to these systems. To put it differently, specific cognitive processes found in human behaviour can be observed, described and implemented into a program. This creates functions with specific behaviours to make virtual agents act and behave in a more realistic manner when speaking to human users (language learners in this case). In this way, the agents can possess a system with its own dynamic characteristics, which can use speech and movements inspired by the human body (Clark, 1999, p. 345). If this is applied to the production of CRs discussed in this thesis, learners can perceive the interaction more naturally, leading to reduced disturbance and a better flow in communication between the learners and the agents. In addition, they can be exposed to the way how language is used by real people, and in this way, learn and practise it with ECAs in a VLE. From the perspective of embodied cognition, Goldman (2006) suggests that language production and language perception involve simultaneous processes (p. 110). This means that the same neural machinery is involved in the production of language as in the perception of it. This example supports the idea that "[hu]man acquisition of semantic representations does not occur based on pure language input" (Beinborn et al., 2018, p. 2325), but on the activation of various areas in the brain that correspond to the sensory modality associated with phrases and expressions. In the context of this thesis, learners can observe how ECAs produce certain types of CRs and in turn they can simulate this behaviour in real life. This means that the learners will most probably use such sensory modalities that they could observe on ECAs in the game.

One of the most important modes of perception of the world around us is vision. In spoken language interaction, vision is used in constructing meaning from the body. When different body movements are produced with or without the co-occurrence of sound, these body movements have a similar semantic structure and the same neural mechanisms as language production (Lakoff and Johnson, 1999, p. 41). Vision (of the listener) often helps to detect the body movements, or the body language, that are usually unconscious to the speaker. These body movements are an inseparable part of human spoken interaction. They have been the most important mode of expression in human-agent interaction that has been present since the creation of the first computer (Maybury and Wahlster, 1998, p. 8). Vision is one of the first and crucial senses to perceive the virtual environment. On this account, Wilson and Foglia (2011) quote James Gibson (1979, in Wilson and Foglia, 2011, p. 1):

> *Vision is not a mere brain process devoted to constructing mental models, but rather a skill of the whole situated, embodied agent, one whose movements are crucial to visual agency.*

Vision helps people to construct a 3D world from the information specified in the 2D image of virtual agents as found, for instance, in many computer interfaces (Wilson and Foglia, 2011). Even though the virtual agents with embodied behaviour appear in an animated form, it is through vision that the human users can perceive them as realistic. The reason for this is that human users employ cognitive embodiment; their brain deciphers and interprets the information coming in through their visual and audial sense. On the basis of neural processes, human users see virtual agents employing various multimodal behaviour that is specific to various communicative functions. The human users compare these to what they know from similar contexts in real-world situations. In *Virtual Reykjavik*, learners can interact with virtual agents by speaking to the microphone or typing text to a message window. They have an avatar with a first-person view which allows them to fully observe the game environment. They use two of their main modes, vision and audio. Vision is used not only to navigate in the game environment, but also to look at the virtual agents in the game and observe how they use their facial expressions and body language in conversations. Learners can also read the other agent's responses on the screen. In addition to vision, learners employ the audio mode when listening to other agent's oral responses. Through these senses of seeing and hearing, learners can decode and interpret multimodal information that the virtual agents convey in their responses. Maybury and Wahlster (1998) suggest that when designing intelligent interfaces with agents possessing multimodal behaviour, it is indeed important to consider different modalities that represent different human senses, such as vision, sound, smell, touch, and even taste (p. 4). In this view, embodied cognition should be considered when modelling a realistic human-agent interaction, especially teaching a foreign language with the appropriate behaviour bound to the community and culture. It is eventually the embodied condition that allows users to feel a sense of embodiment in virtual reality and thus make it to a more realistic experience (Shin, 2018, p. 68). The next section focuses on multimodal grounding, which considers multimodal features that participants in a face-to-face conversation use to achieve a mutual understanding.

2.3.4 Multimodal Grounding in Face-to-Face Interaction

Multimodal grounding describes how a common ground, i.e. mutual understanding, can be achieved between participants in a face-to-face conversation. To find common ground in a conversation, i.e., to find mutual knowledge, mutual beliefs, and mutual assumptions, participants in a conversation need to share information and update one another on new information, moment by moment, in order to coordinate on the progress of their discussion, a process known as grounding (Clark and Brennan, 1991, p. 127). Spoken language is a very complex set of verbal and non-verbal variables (Skinner, 1948, p. 131). As early as 1948, Skinner was one of the first advocates of non-verbal behaviour in spoken language studies, which at that time mainly focused on verbal signs. He argued that an utterance can be expressed in more than one way, and when considering human biology, muscles are involved in the production of verbal sound. According to this view, people use various facial expressions and body language when they produce sound in a face-to-face interaction. It is therefore not only a mere sound (verbal cue) that speakers produce but also non-verbal signals (non-verbal cues) that accompany this sound. The non-verbal behaviour is for this reason of equal importance to verbal behaviour in spoken language interaction (Skinner, 1948, p. 16). Human speakers connote non-verbal scenes and images as well as verbal signs in face-to-face interaction, which is seen by other scholars as the prime form, or the nucleus of, communication (Honeycutt, 2009, p. 198) , from which other forms of communication, such as writing or talking on the phone, are ultimately derived (Fillmore, 1981; Clark 1996; Bavelas et al., 1997; Müller, 2013). Participants in a conversation use it to draw on their common ground to maximise the likelihood of mutual understanding (Bavelas et al., 1997, p. 1).

In the view of the above, multimodal grounding considers spoken interaction as the production and perception of multimodal features. This approach sheds light on what kind of multimodal cues are used in different kinds of communicative functions, e.g. feedback or asking a for clarification. Multimodal grounding thus helps to look at how different patterns of multimodal behaviour are used by human speakers in real conversations on the one hand, and which multimodal cues can be used by embodied or disembodied (robots) agents when interacting with humans (Hee et al., 2017) in order to achieve a common ground (Nakano et al., 2003) in a more realistic human-agent interaction (Xu et al., 2016) on the other hand. The virtual agents thus represent embodied versions of natural language dialogue systems (Nijholt and Heylen, 2002, p. 333) that interact with human users. But since conversation in natural language between humans in reality is "a rich interaction among multiple verbal and non-verbal channels" (Quek et al., 2002, p. 172), it may be difficult to perfectly simulate it in a virtual environment. Regarding the complexity of work and further research needed in this area, Thórisson and Jonsdóttir (2008, p. 131) state:

> *Much of the work in the field of dialogue over the last 2-3 [now 4] decades has enforced strict turn taking between the system and the user, resulting in fairly unnatural, stilted dialogue. The challenge in building such systems lies, among other things, in the complexity of integration that needs to be done: Several complex systems, each composed of several complex subsystems – and those possibly going another level down – need to be combined in such a way as to produce coordinated action in light of complex multimodal* input.

Humans possess communication systems with multiple modalities that allow them to use multiple signals at the same time (Pelachaud and Poggi, 2002, p. 184), but when it comes to realisation of context-appropriate multimodal behaviour of ECAs, it may be difficult to synchronise all of these systems at once. Similarly, Vilhjálmsson (2009) mentions in the context of representing communicative functions and behaviour in multimodal communication that "[building] a fully functional and beautifully realized embodied conversational agent that is completely autonomous […] may take individual research groups more than a couple of years to put together all the components of a basic system" (p. 48).

Numerous contextual factors may have an effect on how participants produce and comprehend language. With face-to-face interaction in a natural setting, van Dijk (2008) suggests that it is the proposition of time, place, shared knowledge (common ground) of the participants that should be observed and analysed (p. 11). For instance, the social aspect of communication between human users and virtual agents (Pereira et al., 2014, p. 1449) or the gestures, postures, faces, bodies, movements, physical arrangements of other speakers in a particular environment (Blommaert and Rampton, 2011, p. 6), all shape the way people talk to each other and behave. Participants in a face-to-face interaction simultaneously use multiple semiotic resources, e.g. the speech and the body, graphically and socially present structure in the surroundings, sequential organisation, encompassing activity systems, etc., whereas certain sign phenomena are visible more and some less, depending on the social context (Goodwin, 2000, pp. 1489-1490). Social context may affect the combination of various multimodal communicative signals, such as words, prosody, gesture, face, body posture and movements that are displayed by ECAs. These are also determined by different aspects, such as (a) contents to communicate, (b) emotions, (c) personality, (d) culture, (e) style, (f) context, and these determine what the virtual character will say and how (Poggi et al., 2005, p. 3).

In the field of computer science, AI and NLP, the work on spoken interactions in a specific context has advanced much more than, for instance, in psychology (van Dijk, 2008, p. 10). The reason may lie in the different properties of discourse, i.e., pronouns, deictic expressions, etc., are studied in order to design a realistic conversational behaviour of ECAs. It is therefore important to study the properties of discourse in specific conversational contexts in order achieve a realistic human-agent interaction. In order to model ECAs that would understand a contextual situation, Prada and Paiva (2014) address various challenges that need to be solved. These include methodological challenges (develop models for interaction dynamics, take a user-centred approach, use data but also theories, develop methods for assessment); situation awareness challenges (develop computational mechanisms to understand others, care for user understanding, develop computational mechanisms to understand the context); interaction dynamics challenges (engage in long-term interactions, consider group dynamics, include the five senses); and societal challenges (account for responsibility) (Prada and Paiva, 2014, p. 7). This list of challenges indicate that it is a complex process to develop agents with high contextual awareness. Much research and technological improvement is required in order to fully achieve this. Prada and Paiva (2014) advise that one has to start investigating natural language step by step and use findings from each individual study as a partial contribution to the whole process of challenges.

Today, it is important to include speech recognition and speech related tasks into the system development of modern tools (Caglayan et al., 2019, p.1). These systems can help to achieve multimodal grounding in human-agent interaction by analysing phonemes in real time (Benoit, 1999; Caglayan et al., 2019). In the context of VLEs, an effective way to bring learners into a specific context for practising spoken language skills and learning about the culture and behaviour of people of the target language, is to populate VLEs with ECAs that use realistic multimodal behaviour and authentic phrases in a specific context. Learning tasks in this environment should be designed to support practising oral language skills by simulating a real-life scenario. These tasks could also be set into a specific conversational context in order to simulate situations that learners may recognise from real life, e.g. asking for directions to a particular place in central Reykjavik. The following section discusses the theoretical foundation for designing a simulation of real world in training situations with virtual humans for practising an authentic spoken interaction.

2.3.5 Simulating Real World Scenarios in Training Situations with Virtual Humans

In order to ensure that the simulation of real world in VLEs for practicing spoken interaction with virtual humans is as close as possible to the real world, research in natural language should be conducted. This research can provide information about what verbal and nonverbal cues virtual humans, i.e. agents, should have in order to conduct realistic interactions with human users (Filipović, 2015, p. 31). Models, predicting grounding structure and spoken turns in conversations, can be developed based on such research. Further findings can moreover shed light on how body and speech are interconnected, and which multimodal behaviour should be included in ECAs when interacting with human users. This would help to build systems with realistic virtual interfaces that simulate real life scenarios. Even though the list of challenges indicate that it is a complex process to accomplish, the technological work should not only focus on the exact reproduction of a real-world scenario but on designing effective training simulations that reflect real-life settings. This view is not only supported by some of the early pioneers of simulation fidelity in training system design (Ellis et al., 1968; Hays and Singer, 1989) but also by Shin (2018), who, after conducting a rigorous study on embodied experiences in a virtual environment, suggests that “no matter how functional and advanced the technology, the key is to focus on the story, not the technology itself or any special 3D effects. The real challenge is not so much that things can look too real or not real enough; instead, it involves the feel of the piece, as perceived by the users of VR stories” (p. 72). This view leads to the introduction of the notion fidelity and feedback in this thesis. The former deals with the question of “how similar to the actual task situation must a training situation be to provide effective training” (Hays and Singer, 1989, p. vi) and the latter with whether feedback can reduce the realism of the training situation but enhance learning (Hays and Singer, 1989, p. 15). Even though it can be difficult to obtain a precise answer to both questions, the following part will try to explain both terms in the context of this thesis and provide a theoretical definition of the notion fidelity, which sheds light on the complexity of work connected with designing a realistic conversation between ECAs and human learners in the game.

In the context of this thesis, the notion of fidelity is concerned with learning spoken language skills in a simulation of real-life conversations with agents in *Virtual Reykjavik*. More specifically, the notion of fidelity is concerned with how research on utilising technology for simulating real-world tasks can guide further development of tools and devices for a more successful training in the future (Jones et al., 1985, p. 97). Fidelity is thus used as a “conceptual bridge between the operational requirements and the training situation” (Hays and Singer, 1989, p. 1) to provide hands-on learning and training of cognitive or functional aspects of tasks. In the context of this thesis, fidelity thus provides understanding of how a realistic human-agent interaction should help L2 learners of Icelandic associate situations that they practise in the game based on solving tasks in a simulation of reality. The similarity of situation, frequency of practise and contiguity, i.e. a state of stimulus and response (ibid., p. 24)., can help prepare them for using the language in real life. It is, however, very important to remember that the notion of fidelity should be used only as a tool for understanding how developers can advance their training systems to reach effectiveness in learning and transfer of skills by a simulation of reality (Hays and Singer, 1989, p. 45). It should not dictate what high or low level of fidelity is, but to accumulate information based on surveys and research. This information can then be used for advising on the configuration of the training programs and the design of systems as a whole to achieve a more effective training in a simulation of reality (Hays and Singer, 1989, p. 46). For instance, if trainees or learners are provided with too much information during training, such as there are too many controls to operate in order to simulate a real-life scenario, it can lead to high fidelity but decrease the learning effect (ibid., p. 51). The learners may be too busy focusing on fixed procedures in tasks and therefore not achieve the learning effect, which they aimed at upon the training start. Even though high fidelity simulation models can provide learners to such exposures that would otherwise not be possible, or difficult to experience in real life on frequent basis, it should not be treated superior to low fidelity situations because both may lead to a similar performance. However, the high-fidelity situation may cause an undesirable effect of overconfidence (Massoth et al., 2019). Both high and low fidelity situations can simulate intensive real-life scenarios that students can enter and practise their skills in a save interim space in virtual reality (Grant et al., 2010, p. 178). But there might be a difference between developing L2 language and other skills dealing with, e.g., treating seriously ill patients. For instance, the development of L2 language skills requires the practice in a low anxiety simulation because cognitive learning of new information about grammar, vocabulary, pronunciation and multimodal behaviour in intercultural communication is used. Overall, it is practical to use the notion of fidelity to gather information to improve systems aimed at practising and learning different skills, but one has to differentiate the purpose. Whether it is in medical assessment of manikins and using other props representing the practice environment (Seropian, 2003), or in a pilot flight simulation (Perfect et al., 2013), or even in assessing the performance of robotic vision (Skinner et al., 2016).

In current research, fidelity is used to understand various kinds of performance in a simulation of a real-life conversation. The notion of fidelity is used here as a theoretical basis for experiencing social, visual and auditory sensors in the VR system to help with improving effective training, feeling of presence, and engagement of learners (Lane et al., 2013, p. 4). The research on multimodal features in CRs in real life conversations and the consequent creation of multimodal CR models should help the agents hold a more natural conversation and therefore better simulate real-life conversational scenarios. Based on an enhanced communicative behaviour of ECAs, the *Virtual Reykjavik* game could more effectively help L2 learners of Icelandic develop their communication skills with using both verbal and non-verbal features in that the learners experience how natives speak. Research in fidelity can moreover lead to a better understanding of immersion of learners into the games by studying the motivation, retention, feedback and immersion behind learners' playing the game.

Feedback is also very important in connection with fidelity. Intelligent tutoring systems can provide instant feedback to learners as they interact with virtual characters. The systems can also collect feedback about learners from pervious actions and provide further support to learners in a form hints, correct answers, thus leading to scaffolding knowledge and further enhancement of skills, which can furthermore result in a better task performance (Lane et al., 2013, p. 8). In computer games for language learning, feedback can be synchronous or asynchronous, delivered in the form of comments in messages or a recast of errors, provided either by the computer or another person such as the teacher or another learner-player in the game (Cornillie et al., 2012). Instant feedback can also be very useful in that it provides immediate response upon learners' completion of tasks, but it can also become overwhelming, resulting in a lower performance of learners (Burgos et al., 2007). In a simulation of reality, different feedback types can help learners to be more aware of the situation they are in the VLE. For instance, learners can be advised on their appearance (of avatars) when wearing a special gear or how well they handled a particular situation in a given task with virtual characters (Lane et al., 2013, p. 6). But when learning a new language, corrective feedback is very useful. Prompting-answer strategies, such as CRs, belong to an effective corrective feedback in ICALL for L2 learning (Ferreira et al., 2007, p. 392). These strategies can push learners both to notice language errors, particularly grammar and pronunciation, especially of beginner learners, and to repair their errors for themselves (Ferreira et al., 2007, p. 415). This may lead to a more interactive teaching model in ICALL systems with effective feedback (Ferreira et al, 2007, p. 415). In *Virtual Reykjavik*, the intention is also to include CR strategies but with multimodal features that ECAs will use in conversational tasks. This will not only help learners to notice and correct their language errors but also keep a more natural flow of conversation in case technical errors occur. Multimodal CR strategies can simulate a response of a person in real-life, thus leading to a smoother conversation between the agent and the learner. The following section discusses the theoretical foundation for designing realistic multimodal behaviour of ECAs when using clarification strategies in a spoken interaction with human users (language learners).

2.4 The Use of Clarification Requests by Embodied Conversational Agents

Lane's at al. (2013) understanding of the grounding process as explained above can be used to understand how messages are produced and received: "How one forms a message (consciously or not) depends again on context, beliefs, biases, and so on. Automated communicative skills are deeply rooted and, thus, difficult to modify in ways that enhance the odds of producing more effective outgoing messages" (p. 2). But, with the use of corrective feedback learners can learn novel communicative skills bound to culture. The learning of communicative skills can be compared to learning of other cognitive skills in that learners establish new knowledge structures, refine or tune these structures, and strengthen of such memory structures through use (Greene, 2003, p. 57). It can be said that repeated practice opportunities with feedback are an essential component in the development of learners' communicative skills (Lane et al., 2013, p. 2). In *Virtual Reykjavik*, using multimodal CRs does not only contribute to a more realistic interaction between human users (learners), because they will not only help with maintaining a smooth flow of a conversation, but also demonstrate to learners how native speakers ask for clarifications in real life and help them develop novel communication skills based on how ECAs use CRs in communication with them. The analysis of learners' needs and experience playing the game is important to study. In this context, the Concept of Flow and the model for training evaluation and effectiveness supporting the two auxiliary studies will be discussed.

Clarification requests will be considered as part of a turn-taking sequence because they appear in so-called ‘adjacency pairs’ in a dialogue between ECAs and learners. This means one turn represents a request and another turn a clarification of what has been said. The ECAs use the first part of the adjacency pair, the request for clarification. In the following sections, the definition and various kinds of CRs will be presented.

2.4.1 Clarification Requests

Clarification requests are often known under different terms, such as requests for clarification (Duncan and Niederehe, 1974, p. 236), repair initiators, or Next Turn Repair Initiators (NTRIs) (Schegloff, 1992, p. 1318), requests for repair (ReqRepair) (Traum and Allen, 1992, p. 4), CRs (Purver, 2004, p. 15), and clarification questions (Saxton et al., 2005, p. 393). They belong to the most commonly used communicative functions in a conversation, with occurrence at around 4% of all dialogue turns (Purver et al., 2003, p. 241). In this thesis, Purver’s (2004) term, CR, has been adopted. Even though this utterance is sometimes referred to a breakdown in communication (Saxton et al., 2005, p. 393; Purver, 2004, p. 15), or more specifically a breakdown in intersubjectivity between turns (Schegloff, 1992, p. 1299), other research, however, considers the CR as an initiator of repair that can help the speaker to return to the previous utterance and clarify it. When CRs are used, the conversation continues despite the failure in the turn-taking procedure. According to O’Connell et al. (1990), “(a) true breakdown would have to be a conversation stopper” (p. 346), which a CR is not. For this reason, the CR represents a function that helps to keep a smooth flow of a conversation. Similarly, in L2 education, CRs are also considered as markers of disfluency in speech and often referred to as self-repair or self-correction in speaking (Ellis and Yuan, 2005; Witton-Davies, 2010; Lowder and Ferreira, 2016). However, Mihas (2017) considers them as resources that learners can use for maintaining meaningful verbal interactions with other learners or language instructors (p. 221). This view has also been adopted in this thesis: the CRs have a useful property of keeping a conversation ongoing. For this reason, they have been incorporated into the design of realistic human-agent interaction in *Virtual Reykjavik,* that aims at both teaching Icelandic language and culture to L2 learners and having high visual and auditory fidelity, i.e. realistic graphics of virtual graphical interfaces represented by ECAs.

Various studies have identified different types of CRs. These will be discussed and presented in Table 2 below. For instance, Schegloff et al. (1977) investigated the organisation of repair in English conversation. They distinguished between two types: self-correction and other-correction (ibid., p. 361). In the other-correction type, they found five basic categories of the next turn repair initiation (NTRI) (Table 2). Cho (2007) conducted research in an English conversation between Korean speakers. When categorising CRs, Cho was inspired by Schegloff's et al. (1977) research and expanded the categories for five additional CRs (see Table 2 below). Purver (2004) identifies four major categories (Table 2) that are different to the ones from Schegloff et al. (1977) and suggests that CRs can vary widely not only in their form, but also in the information they carry (p. 15). Dingemanse and Enfield (2015) conducted research on other-initiation of repair across ten languages. They claim, however, that the five formats as identified by Schegloff et al. (1977) in an English conversation were not offered as cross-linguistic categories, but certain properties of those can be traced across formats and across languages (ibid., p. 102). Gísladóttir (2015) contributed to this research on other-initiation repair across languages by focusing on Icelandic. Gísladóttir found six major categories, which were divided into two main types - open and restricted CRs (p. 313). These categories are interjection, question-word strategy, formulaic: expression incorporating question words or interjections, request, offer, alternative question (Table 2). The category of explicit type "on-the-record", i.e. when the intention of the communicative act is clear and non-deniable, and implicit "off-the-record", i.e. when it is not possible to attribute one clear communicative intention, was used by Brown and Levinson (1987) and Manrique and Enfield (2015). In this regard, Manrique (2016) described an implicit type of CR, the "freeze look", which is a non-verbal type used in Argentine Sign Language (Table 2). A non-exhaustive summary of different types of known CRs based on the research listed above is presented in Table 2.

Table 2: Overview of different CR categories.

Study	CR Type	Example (CR)	Expressive Type
Schegloff et al. (1977)	1. Huh? What?	*Huh? What?*	Explicit
	2. Wh-words	*Who? Where? When?*	Explicit
	3. Partial repeat + wh-word	*The who? Met who? To a where?*	Explicit
	4. Partial repeat	*The…? Met …? To a …?*	Explicit
	5. You mean + possible understanding of prior turn	*You mean + … ?*	*Explicit*

Purver (2004)	1. Full Explicit Query	*A: Did Bo leave?* *B: I am sorry, what did you say? (CR)*	Explicit
	2. Echoes	*A: Did Bo leave?* *B: Did Bo leave? (CR)*	Explicit
	3. Ellipsis	*A: Did Bo leave?* *B: Bo? (CR)*	Explicit
	4. Fragments	*A: Did Bo leave?* *B: Eh? (CR)*	Explicit
Cho (2007)	1. Huh? What?	*Huh? What?*	Explicit
	2. Wh-words	*What's that?*	Explicit
	3. Partial or full repeat + candidate info	*A: I was in information.* *B: Information room? (CR)*	Explicit
	4. Partial repeat	*A: Just trip.* *B: Trip?(CR)*	Explicit
	5. Full repeat	*A: How about you?* *B: How about me? (CR)*	Explicit
	6. You mean + candidate understanding	*A: Ahh subway or bus?* *B: Ahh you mean the public transportation, right? (CR)*	Explicit
	7. Wh-words to request more info	*A: Anyway, can you do me a favour?* *B: What kind of a favour do you want? (CR)*	Explicit
	8. Appender (co-construction) question with candidate info	*A: Between big band and bap bip bap.* *B: Swing? (CR)*	Explicit

	9. Candidate substitution		*A: I am shame, I am shame.* *B: A little bit shy? (CR)*	Explicit
	10. Overt indication of non-understanding (rising intonation)		*A: (saying something)* *B: Pardon, what are you saying? (CR)*	Explicit
Gísladóttir (2015)	Open	1. Interjection	*Ha?*	Explicit
		2. Question word strategy	*Hvað segirðu? (What are you saying?)*	Explicit
		3. Formulaic: expression incorporating question words or interjection	*Afsakið (excuse me). Fyrirgefðu (sorry/pardon me).*	Explicit
	Restricted	4. Request (asking specification by question words)	*Who was this?*	Explicit
		5. Offer (providing a candidate by repetition or rephrasing)	*A: She is not old, sixty or something.* *B: Sixteen? (CR)*	Explicit
		6. Alternative question	*e.g. divorced or separated?*	Explicit
Manrique (2016)	-	"Freeze Look" response (notable or 'pointed' absence of response)	E.g. participant is holding their hands and body in a still position and is looking directly at the questioner	Implicit

In the structure of ordinary sequential organisation of a conversation, repairs have a particular position in the order of turns. Clarification requests indicate a breakdown (not a stop) in communication and are usually produced immediately after the problematic turn. Such turns, be they single-word turns, such as *Huh?,* or longer phrases (Sacks et al., 1974, p. 702) belong to a turn-taking sequence. When such situations arise, the speakers do not stop but continue speaking. In case of lack of comprehension or clarity in information, a listener produces a CR to retrieve the missing or misunderstood information from the speaker. This helps the interlocutors to reach and maintain a common ground. The utterance used for asking for a clarification may be *Eh?, Pardon?*, *Sorry?*, *Once again?* or, depending on the context, also *You want what?* (Purver, 2004, p. 15; Saxton et al., 2005, p. 393; Ogino, 2008, p. 8; Ding, 2012, p. 84).

Clarification requests appear in a particular position (Table 3) in a conversation sequence and are for this reason also included in the turn-taking process (Schegloff et al., 1977, p. 369; Colman and Healey, 2011, p. 1563). The position refers to when and by whom the CR is initiated. For instance, position 1 refers to whether the initiator edits, amends or reprises a part of their contribution before another participant responds to it. Position 2 refers to a contribution to a conversation that has been introduced to propose repetition or revision of another participant's contribution. Position 3 refers to whether this utterance has been used to edit, ament or reprise a previous contribution by the initiator (Colmam and Healey, 2011, p. 1564). In the context of this thesis, the CR typically appears in position 2 because it is an initiated repair by another participant (listener).

Table 3: Distribution of repair in dialogue (Colman and Healey, 2011, p. 1564).

Gloss	**Repair Protocol Category**
Repeat	Position 1 Self-Initiated Self-Repair 'Articulation'
Restart	Position 1 Self-Initiated Self-Repair 'Formulation'
Transition	Position 1 Self-Initiated Self-Repair in Transition Space
Clarification Request (CR)	**Position 2 Next Turn Repair Initiator (NTRI)**
Correction	Position 2 Other-Initiated, Other-Repair
Follow-up	Position 3 Other-Initiated, Self-Repair
Reformulate	Position 3 Self-Initiated Self-Repair

Such occurrences indeed follow a certain pattern: initiating a problem, repairing a problem, and the position of the repair (Colman and Healey, 2011, p. 1563). The focus in this thesis is therefore on CRs that helps the listener to get clarification on the missing or misunderstood information from the previous speaker. It helps to maintain the smooth flow of a conversation when communication may for some reason, technical or non-technical, be interrupted (Saxton et al., 2005, p. 394). Even though lexical problems appear to be rare and reference problems more common in human-human dialogue (Reiser et al., 2005, p.

4), in the process of L2 learning, lexical items play a crucial role in the comprehension of the target language, because words carry information included in participants' utterances (Saxton et al., 2005, p. 394; Reiser et al., 2005, p. 4). As the main research herein investigates CR utterances in Icelandic, the following section covers this topic.

2.4.2 Clarification Requests in Icelandic

To date, only one study is available on CRs in Icelandic (Gísladóttir, 2015). It focuses on repair problems in conversations between native speakers and describes only the linguistic practices of Other-Initiated Repair (OIR), which is a synonym to CR. It lists the whole sequence of turns in participants' repair process. The study highlights the interjection *ha*, the usage of which extends beyond the open type of OIR (Gísladóttir, 2015, p. 309). That study moreover looks at the ability of native Icelandic speakers to repair problems with hearing or understanding information from other speakers and described requests and practices they use towards achieving a successful conversation. In particular, the study lists two kinds of OIR sequences: minimal and non-minimal. The former is a dialogue between only two speakers consisting of a sequence of three utterances: a turn before the repair utterance (T-1), a turn containing the repair utterance (T0), and a turn after the repair utterance (T+1). This research characterises two main types of repair utterances (open and restricted), each of which has three more subcategories (Table 4).

Table 4: Types of repair initiators and their frequency in the Icelandic corpus (Gísladóttir, 2015, p. 313).

Type	**Subtype**	**Number of cases reported**	**Proportion**
Open	Interjection	51	34.7%
	Question-word	7	4.8%
	Formulaic	0	0%
Restricted	Request (asking specification)	33	22.4%
	Offer (providing a candidate)	55	37.4%
	Alternative question	1	0.7%
	Total	147	100%

According to the overview (Table 4), the most common repair utterance is an ‘open type’ interjection, *ha*. The interjection *ha* has a phonemic /h/ in onset position, followed by a low-central, unrounded vowel, with a falling pitch. This is different to other comparable interjections in various other languages that do not have a falling pitch in *ha* (Gísladóttir, 2015, p. 314). The utterance offering (providing a candidate) belongs to the most common types of restricted repair utterances. This particular repair technique helps the listener to check whether they have understood correctly what the speaker has just said. By doing so, they offer a candidate word/phrase, which is consequently either confirmed or rejected by the speaker, usually by saying “yes” or “no” or the correct word again (Gísladóttir, 2015, p. 320). Requesting clarification/asking for specification belongs to the second most common restricted type of clarification utterance in Icelandic. It is done by using content ‘wh-‘ words, such as *hvað* (what), *hver* (who), *hvar* (where), and *hvert* (where to), also with a falling pitch. In case of a partial repetition of the previously said phrase by other speaker, the one who requests repair does so with repeating a part of the sentence/word with a level intonation (Gísladóttir, 2014, pp. 317-319). This study points out the importance of intonation in Icelandic. Intonation is important for detecting and responding to differences in interpretation of received information, which seem to be a recurrent and routine problem in conversation (Colman and Healey, 2011, p. 1563). CRs are used for clarifying the meaning in such encounters. Intonation is also important to distinguish CRs from other communicative functions, such as acknowledgements (Traum and Allen, 1992, p. 4). Therefore, the present thesis includes intonation in the analysis of multimodal features of CRs. Unlike in other languages, it is the intonation contour in Icelandic, namely falling pitch, which decides for trouble-presenting repetition (Gísladóttir, 2015, p. 321).

This section has presented what linguistic types of CR are used in dialogues among native speakers of Icelandic that know each other. Even though its focus was limited to the linguistic types and forms of use, it nonetheless informs about the paralinguistic features, such as intonation, which is relevant to the current study for the purpose of comparison. Since spoken language in a face-to-face interaction is produced multimodally, the following section will also discuss the use of CR strategies from this point of view and set them into the context of human-agent interaction.

2.4.3 Multimodal Clarification Requests in Human-Agent Interaction

The clarification of meaning leads to an effective multimodal discourse because the answer one receives is usually helpful – it clarifies what was not so clear before (Skinner, 1948, p. 9). The occurrence of CRs may be, however, more difficult in human-agent interaction than in human–human interaction. Unlike in face-to-face interactions between humans where multimodal cues are mutually and instantaneously shared between participants in a conversation, the agent and the learner do not possess the same ability to detect the other participant's intention to take turn. For instance, the agent must rely on the spoken input or another command performed by a user whereas the user has more opportunities to detect whether the agent is yielding a turn. The agent can use various multimodal features to perform specific functions for giving or requesting a turn and the user can both observe and/or hear it. However, research on multimodal CR strategies is scarce. Only a small number of studies that are presented below talk about CRs in a multimodal context. One of the first studies describing CRs multimodally was conducted by Duncan and Niederehe (1974). Their requests for clarifications were, however, characterised as part of back channel behaviour by speakers in a face-to-face dialogue. At this early stage, the term *back channel* was adopted in order to cover verbalisations such as "m-hm" and "yah", and head movements, such as nods and shakes, that could be frequently observed on the part of auditors (listeners) (Duncan and Niederehe, 1974, p. 236). But, requests for clarification cannot be classified as back channels because "they have the effect of directly influencing the subject matter and the stream of talk and are very close to ordinary question/answer paired turns" (Oreström, 1983, p. 106). Even though this research took into account multimodal realisation, it was done so only in part. With the exception of head nods and shakes, all definitions were mostly based on the verbal form of utterances. The data was derived from detailed transcriptions of speech and body-motion behaviour during the first 19 minutes of two conversations that were recorded on a videotape. Speech intonation, paralanguage and all observed body-motion behaviours were carefully noted. Furthermore, they observed that in the case of longer back channels, e.g. sentence completions, requests for clarification, and brief restatements, the boundary between speaking turns and back channels became uncertain, and in particular in restatements (Duncan and Niederehe, 1974, pp. 236-237).

Another well-known study on including multimodal CRs in a spoken dialogue system between ECAs and human users is the study on the REA system (Cassell et al, 1999; Bickmore and Cassell, 2005). This system included an ECA which would use various verbal and non-verbal modalities in different communicative functions. One of these functions was also the CR. The communicative function of CR was included in the category of feedback, during which an ECA would non-verbally request feedback from the listener through gaze and raised eyebrows, or through a confused facial expression if they did not understand the speaker's input. In practice, the ECA could use the REA system to execute communicative functions with multimodal behaviour. This included asking for feedback with gaze and raised eyebrows or giving feedback with gaze and a head nod (Bickmore and Cassell, 2005, p. 26), and in this way initiate conversational error correction when the ECA misunderstood what the user has said (ibid., 2004, p. 27). The ECA could moreover generate combined voice, facial expression and gestural output (ibid., 2004, p. 27) and regulate the structure of a conversation by supporting a smoother turn-taking (Bickmore and Cassell, 2005, p. 31).

Louwerse et al. (2007) studied multimodal face-to-face computer-mediated conversations between humans in order to investigate how discourse structure, speech features, eye gaze and facial movements interrelate during a map coordination task (map drawing) (p. 1235). In particular, they examined how these modalities were aligned and whether the correct use of these channels aids comprehension. Different communicative functions were produced multimodally in a dialogue between participants. Their multimodal features, including eye tracking, were captured by video camera. Communicative functions were in this context described as dialogue acts. One such dialogue act was coded as CLARIFY and its function was described as "Reply to question over and above what was asked", e.g. "*So, you'll be between the blue and red car*" (ibid., pp. 1237-1238). An overview of facial movements with labels for each mouth, eye/eyebrows and head movement was developed, with a description of their use in average frequencies of both the information giver and the information follower. The aim of this study was to use multimodal data to develop an animated conversational agent "that can interact with a human dialogue partner and behave similarly to the human dialogue partner in terms of using modalities like dialogue structure, speech features, eye gaze and facial movements" (ibid., p. 1238). Even though their study did not provide information regarding the implementation of those into the architecture of ECAs and their effect on users, their next study does (Louwerse et al., 2009). In the next study, the results from the personal assessment questionnaire showed that the presence of facial expressions, gestures and intonation had a positive effect on the perceived usefulness of the agent and the performance at the task. Multimodal features from their previous study (Louwerse et al., 2007) had been implemented into the architecture of the Haptek avatar in their later study, which played the role of an instruction giver, whereas the user was the instruction receiver. Regarding the CLARIFY dialogue act, the ECA included behaviour such as head nodding, head shaking, and deictic gestures for giving positive feedback, and biting lip or stroke face when giving negative feedback. An experiment was conducted in which the intensity of behaviour was modified when considered too expressive (or unnatural) based on trial and error testing to achieve the desired effect (Louwerse et al., 2009, p. 1461). The results suggested that participants (users) first and foremost valued multimodal behaviour in ECAs, whereby giving the highest importance to the role of facial expressions, lesser importance to gestures and intonation when it comes to credibility and human-likeness (intonation) (ibid., p. 1463). In their view, the implementation of multimodal behaviour into the conversational architecture of ECAs had a positive effect on users when communicating with their agent. In particular, the gestures in dialogue acts had a pragmatic factor when they were general, and a semantic factor when they were specific, whereas intonation played only a semantic role (ibid., p. 1464).

Healey et al. (2015) examined whether and how speakers and non-speakers provide concurrent feedback in a conversation. According to their results, both actively contribute to the production of each turn, and although the non-speakers produce fewer hand gestures than speakers, "(they) provide frequent concurrent feedback to speakers and sometimes use non-speech signals to engage directly in helping the speaker to produce their turn" (Healey et al., 2015, p. 28). They suggest that clarification sequences should be separated from the rest of the dialogue because they involve a distinctive use of non-verbal resources, in which both speakers and non-speakers change their (multimodal) behaviour (ibid., p. 28). Even though this study informs about the frequency of occurrence of hand gestures and their general characterisation as content-specific (iconic, metaphoric, deictic, pantomime and abstract deictic) and feedback (contact perception, comprehension, attitudinal/emotional) in clarification sequences, it unfortunately lacks to inform during which particular types of clarification sequences that particular hand gestures and feedback are produced.

The commonality of selected studies above is that they lacked to inform about multimodal features that occur in various types of CRs. This thesis focuses aims to catalogue various CR types and their features in order to offer a variety of options for how ECAs can execute clarification strategies in context-specific situations. The following section will discuss the role and function of ECAs and their contribution to creating a more realistic in human-agent interaction.

2.4.4 Embodied Conversational Agents

Embodied conversational agents are animated virtual characters, with shapes ranging from human through animal and to other kinds of creatures. In the context of this thesis, however, the ECAs have a shape of human speakers that can recognise and respond to verbal and non-verbal input from human users. The form of the verbal input is in a form of speech that is recognised through the means of an ASR. The form of the non-verbal input is the proximity of the learner's avatar (how close the agent stands to the learner's avatar in the game) and the direction of looking (whether the learner's avatar is looking at the agent, which will inform the agent about the direction they need to look in order to look at the learner's avatar, and be prepared for interaction). Generally, ECAs can generate verbal and non-verbal output, deal with different conversational functions such as turn-taking, feedback and various repair mechanisms, and give signals indicating the state of the conversation, as well as contribute with new propositions to the discourse as human speakers would in real life (Cassell et al., 2000, p. 29). In order to design a realistic human-agent interaction in *Virtual Reykjavik*, it is important to firstly understand the characteristics of human-human interactions (Jokinen and McTear, 2010, p. 97) and endow the agents with intrinsic properties to make them more believable (Cassell et al., 2000, p. 31). This may assist to generate realistic verbal and non-verbal output and deal with different conversational functions in an authentic way. As embodiment is increasingly becoming a part of the design of many intelligent systems (Hasegawa et al., 2010, p. 11), it is important to mention some examples to support why it would be important in *Virtual Reykjavik*.

Believable ECAs are often used in serious computer games that simulate real-world situations. Three-dimensional serious games with L2 learning purposes include the tactical language and culture training system *Tactical Iraqi* and *The Danish Simulator*. Other serious games, such as the Danish digital/video game *Mingoville*[29] for teaching English to children, or *Adventure German*[30] for teaching German as L2 to adults, use animated 2D agents, in the shape of humans or creatures, that do not have features of ECAs with human-

[29] http://www.mingoville.com/

[30] https://www.goethe.de/en/spr/ueb/him.html

like behaviour. They only produce simple movements, e.g. walking, sitting down, and opening mouths when speaking. However, the *Tactical Iraqi* and the *Danish Simulator* are similar to *Virtual Reykjavik* in that they are populated with 3D virtual agents in human-like ways. They use ECAs whose multimodal behaviour consists of a combination of different verbal and non-verbal modalities that occur in real time (Cassell et al., 2001, p. 477; Kopp et al., 2006, p. 205). This thesis nonetheless shows the importance of including multimodal behaviour in specific communicative functions to simulate a realistic conversation in a 3D computer game, and in this way contribute to a better L2 learning experience. The following section describes how this can be achieved.

2.4.5 Communicative Functions and Behaviour

From the point of view of linguistics, communicative functions belong to the pragmatic use of language (Stadler, 2013). They are defined as speech acts that convey a communicative purpose in spoken interactions, e.g. making suggestions, agreeing, disagreeing, and asking for information (Koester, 2002, p. 168). This view is, however, limited because it refers only to the verbal part. Apart from linguistic cues, further research shows that communicative functions also include other cues, such paralinguistic (e.g. intonation, pitch) and non-verbal (e.g. gestures, body posture, etc.) (Duncan, 1972, pp. 287-288). Participants in a conversation often use a combination of these. For instance, when someone wants to yield a turn, i.e. to make other speakers aware that they want to say something, they use the communicative function of turn-yielding. It is done similarly with CRs when a listener requests the other speaker to clarify what he/she have just said.

From the perspective of modelling conversational behaviour of ECAs, communicative functions with multimodal behaviour are used to specify the communicative intent behind agent's behaviour (Heylen et al., 2008, p. 270). From the technical point of view, when designing multimodal communicative functions for ECAs, e.g. a multimodal CR, one has to work with two types of technical extendable markup languages (XMLs). Each type of a communicative function with a specific multimodal behaviour can be translated into these markup languages. They provide a straightforward machine-readable format that allows information to be annotated, or marked up (Bateman, 2012, p. 4) and translated into the function markup language (FML) and the behaviour markup language (BML). The FML describes intent without referring to physical behaviour, i.e. it represents a particular communicative function. According to Heylen et al. (2008), it represents what the agent wants to achieve, i.e. its intentions, goals and plans

(p. 270). For this reason, it is divided into different parts. The first part includes information about the aspect of context, i.e. information about the participant. The next part includes information about other dimensions, such as communicative actions (turn-taking, grounding, speech acts), content (elaborate, summarise or clarify, convince or find-plan, i.e. finding the appropriate speech act from the pre-stored options in the system). Also, information about belief-relation, i.e. whether the act is general or specific (gen-spec), a cause or effect (cause-effect), or solutionhood (i.e. finding a solution within a dialogue plan to give an appropriate answer), suggestion, modifier, justification, or contrast are important. Further parts consist of mental state especially in processes that accompany gaze behaviours of agents, e.g. planning, thinking, remembering, and social-relational dimension or goals (Heylen et al., 2008, pp. 273-275). All of this information helps to specify what purpose the content of performed speech acts should serve (Kopp et al., 2006, p. 210).

The BML, on the other hand, describes the behaviour that supports or carries out communicative functions. Some communicative behaviour, e.g. nodding, requires the head to be moving in a particular way and direction. According to Kopp et al. (2006), this behaviour also requires the synchronisation of various other elements, such as head, torso, face, lips, gaze, body, even legs when the agent's whole body is visible. Each of these BML elements contributes to the visual appearance and movement of the whole behaviour. With all of them employed together, a particular expressive effect can be achieved (Kopp et al., 2006, pp. 210-213). The following characterisation of those two types of behaviours comes from the SAIBA effort (situation, agent, intention, behaviour, animation) to unify multimodal behaviour framework for ECAs (Kopp et al., 2006; Vilhjálmsson et al., 2007; Heylen et al., 2008).

In the context of this thesis, it is both the FML and BML that are part of the CR communicative function which is analysed in the main study of this thesis. Both of these markup languages are agent (programming) languages (Bevacqua et al., 2010, p. 4). The context of a communicative intent, which is expressed by the FML, can be constructed in an arbitrary XML structure (Vilhjálmsson et al., 2007, p. 110). In contrast, BML is an XML-based markup language that can be embedded in a large XML message or document by starting a <bml> and then filled in with behaviours that should be realised by an ECA (Vilhjálmsson et al., 2007, p. 100). Both of these programming languages started to take form in Cafaro's (2014) PhD thesis, in which he presented a theoretical framework for analysing and modelling human non-verbal behaviour for managing impressions. He furthermore demonstrated how relational agents can exploit FML and BML in their first encounters with human users, which is used in the CR scenario in *Virtual Reykjavik*. The main focus of that research was, however, on smile, gaze and proxemics that help to exhibit personality and interpersonal attitudes. But present thesis here focuses on features that need to be included in the BML of the CR function.

Both FML and BML depend on the context in which a conversation proceeds. This means that one may use the same function for known encounters (people know each other when they meet) as well as for unknown first encounters (people do not know each other when they meet for the first time), but the multimodal behaviour involved in each of the encounters will be different. Otherwise it may not be appropriately used and cause confusion, e.g. when meeting a stranger and greeting him/her as one's friend. For this reason, the CRs have been studied in a specific context in real life, which will be similar to the one learners experience in the game, namely first unknown encounters asking for directions.

2.4.6 Keeping Leaners in Flow Through the Use of Multimodal Clarification Requests by ECAs in **Virtual Reykjavik**

Multimodal CRs can better demonstrate to learners in *Virtual Reykjavik* how language is produced in real life by native speakers. Apart from keeping a smooth flow of a conversational between the learners and virtual characters in the game, the multimodal CRs also keep the learners engaged in speaking and therefore playing the game without interruptions and practicing the language when technical problems occur, e.g. the ASR not recognising the speech input correctly due to the learners' speaking silently or unclearly into the microphone. By doing so, it can increase the learners' presence in the game and keep them engaged. In this way learners' immersion in the game can be enhanced. The learners can feel motivated to further play in the future, which will support the learning of language and culture in a game based VLE. By engaging learners with virtual characters in the game, the learners' cognitive and affective learning, immersion in the environment, and interactions of the game world known as presence can be enhanced (Bachen et al., 2016, p. 77). In general fashion, learners can enter a flow state, a kind of immersive experience, when they enjoy playing the game, find it useful for their personal goals by getting feedback, and are motivated to stay and re-enter the game because their required skills and challenges are in balance (Weibel and Wissmath, 2011, p.12). The flow state can be used to predict experiences and outcomes of using computer games in learning and education (Nah et al., 2014, p. 94).

The theoretical background for how users can experience the flow state was derived from the Concept of Flow (Nakamura and Csikszentmihalyi, 2002). The Concept of Flow is a phenomenon studied in the field of positive psychology and provides "understanding of experiences during which individuals are fully involved in the present moment" (Nakamura and Csikszentmihalyi, 2002, p. 89). In the context of this thesis, the flow concept helps to understand the nature and conditions of enjoyment by learners pursuing playing *Virtual Reykjavik* for practising spoken language. The flow concept can shed light on both the users' general perception of the game and also on how natural the interaction with ECAs, e.g. endowed with multimodal features in CRs, is. The contribution of the main study in this thesis is to deliver multimodal models for CRs that would help keep a natural flow of a conversation between the learners and the virtual characters, and assist the learners with ways how locals ask for clarifications in real life. Learners, by using embodied condition, will be able to embody experiences by viewing, playing, and feeling perceptual cues linked to those experiences (Shin, 2018, p. 68). Based on these experiences, i.e. how ECAs asked for clarifications, the learners can later on use similar verbal and non-verbal ways of asking for clarifications when speaking to native speakers in real life. This may lead to a better immersion of learners in the game. But while immersion influences presence and flow to a certain level, user's empathy, i.e. meaning to stories or objects encountered in a mediated environment (embodiment), depend on individual users (Shin, 2018, p. 69). In order for the users to achieve flow, it is not only the embodied experience but also "[p]erceived challenges, or opportunities for action, that stretch (neither overmatching nor underutilizing) existing [learners'] skills; a sense that [the learner] is engaging challenges at a level appropriate to [their] capacities; and [it is necessary to have] clear proximal goals and immediate feedback about the progress that is being made" (Nakamura and Csikszentmihalyi, 2002, p. 90). Identification with a character in the game, in this case the learner's avatar, and the role in the game scenario can also lead to positive educational outcomes, "including greater attention to and retention of messages associated with those characters" (Bachen et al., 2016, p. 82). These elements are associated with engagement and subsequent learning in games. Additional game elements, that should be implemented into computer games to support flow during game play, are a narrative plot and storyline, interactivity with characters, a reward system, short-term goals of game, social interaction and feedback from other players in the game, and devices for controlling the navigation in the virtual environment (Nah et al., 2014, pp. 109-110).

Achieving a natural flow in a conversation between agents and human learners must be viewed in a context of the whole computer game. When users are engaged in playing the game, it is the flow of the whole process of engagement when they are playing and are focused, motivated and concentrated on the tasks and activities. If the conversation proceeds naturally without stopping, and if the feedback gathered while playing is positive as well as educational, it then may naturally encourage the learner to proceed and explore the game even further and use it for learning, e.g., new oral language skills. In addition, it may increase the learner's awareness of being part of the game as an actor who without difficulties can deal with the situation and respond adequately to the agent when playing the game. Through this, a positive experience of the activity can be achieved, and the learner may want to set further goals in fulfilling additional tasks and conversing with other agents in the game while practising the language. Constant feedback is therefore very crucial in this activity and should positively stimulate. Multimodal CRs can be also viewed as an indirect but effective feedback when directing the user to repeat what they have just said and assist them to continue the conversation with the agent without interrupting the conversation. Nakamura and Csikszentmihalyi (2002) define the following characteristics of one being in flow: "Intense and focused concentration on what one is doing in the present moment; merging of action and awareness; loss of reflective self-consciousness (i.e., loss of awareness of oneself as a social actor); a sense that one can control one's actions; that is, a sense that one can in principle deal with the situation because one knows how to respond to whatever happens next; distortion of temporal experience (typically, a sense that time has passed faster than normal); experience of the activity as intrinsically rewarding, such that often the end goal is just an excuse for the process" (ibid., p. 90).

The multimodal features in CRs play an important role in that they keep learners focused on the dialogue; they raise the level of awareness because learners may observe how these CRs were produced by the ECAs as well as why they were produced. The ECAs can use different verbal and non-verbal features in various other types of CRs in a given conversational setting. Moreover, the use of the CR function gives the learners a corrective feedback. The learners can start articulating and speaking more clearly to the microphone, pronouncing words correctly and move toward a better experience in playing the game with less distortion and more rewarding feedback. The *Virtual Reykjavik* game makes it possible for learners to have a conversational practice in an interim virtual learning space allowing them to make mistakes and therefore minimising anxiety and confrontation with real native speakers. In this environment, feedback is stimulating and educational. Achieving the flow in this game depends on establishing a balance between challenges and skills of learners.

There are several challenges for including multimodal features into CRs in order to design a more natural human-agent interaction. Some of these features have already been mentioned above, i.e. conducting natural language research with real people in real life, implementing results into building theoretical models for multimodal CRs, and work associated with designing and programming the ECAs so that they can execute all verbal and non-verbal features believably and naturally when speaking to human users (learners) in the game. These challenges are therefore associated with the development and design. On the other hand, there are other challenges associated with users' perception of these multimodal CRs in the game, i.e. whether the users can perceive them as what the intention is by the designers. One way of finding out is to conduct a rigorous study with multiple users. After evaluating it, the results can advise on the successfulness of the design work as well as the general perception of multimodal behaviour of ECAs endowed with such features in playing the game. In addition, such study can also provide information about how immersed the learners would feel in the game and whether the learners achieved a state of flow when playing the game. The theoretical background for gaining insight into the users' experience with playing the game and evaluation of the study connected with playing *Virtual Reykjavik* is presented here below.

The main task is to keep an uninterrupted conversation between the ECAs and learners. The ECAs should look and sound more believable, i.e. less robotic, with appropriate multimodal features in conversation, and provide feedback to learners to improve their spoken language skills. The challenge of practising spoken conversation with ECAs in a given task with an appropriate language level will need to be kept in balance in order to imply to the flow concept. This can be graphically demonstrated on the model of the flow state presented in Figure 7. It shows both the classical and the current model of the flow state (Nakamura and Csikszentmihalyi, 2014). In the classical model presented in (a), a user experiences the flow represented by an action, in this case it would be playing a computer game. When the user is playing the game, the opportunities and player's capabilities for the action are in balance. This means that perceived anxiety and boredom "meet" in the middle. The current model of the flow state (b), on the other hand, features perceived challenges and skills in an equilibrium state, i.e. when all main factors "meet" in the middle of circulating rings symbolising the distance from the middle. The learner experiences a state of flow, or in other words immersion, when both his/her skills and the challenges are above average. However, when they are below average, the learner can experience apathy, boredom, and too much relaxation. The higher the challenge and the required skill, the more intensive is the learner's experience. When learners experience a flow by playing a game, it may encourage them to persist and return to the activity "because of the experiential rewards it promises, and thereby fosters the growth of skills over time" (Nakamura and Csikszentmihalyi, 2014, p. 249). When learners are in the flow, it can motivate them to use the activity again for further enhancement of their language skills. In section 2.2.3.5, the term motivation for learning was viewed as a complex and non-linear interaction of variables that are internal and external to the cognitive system of the learner (Tennyson and Jorczak, 2008, p. 5). The various components of motivation for learning constantly interact with each other in all directions. In combination with sensory information from an external source, they can help improve the individual's knowledge base. For instance, graphic design in computer games supports visual senses, sound and other audio affect hearing senses. Moving objects around in the game environment simulates tactile senses. The more objects with physical/living characteristics are implemented into the game environment and interact with the player, the more realistic it becomes for the learner to play, which eventually leads to a more natural experience in virtuality (Bossomaier, 2012, p. 32).

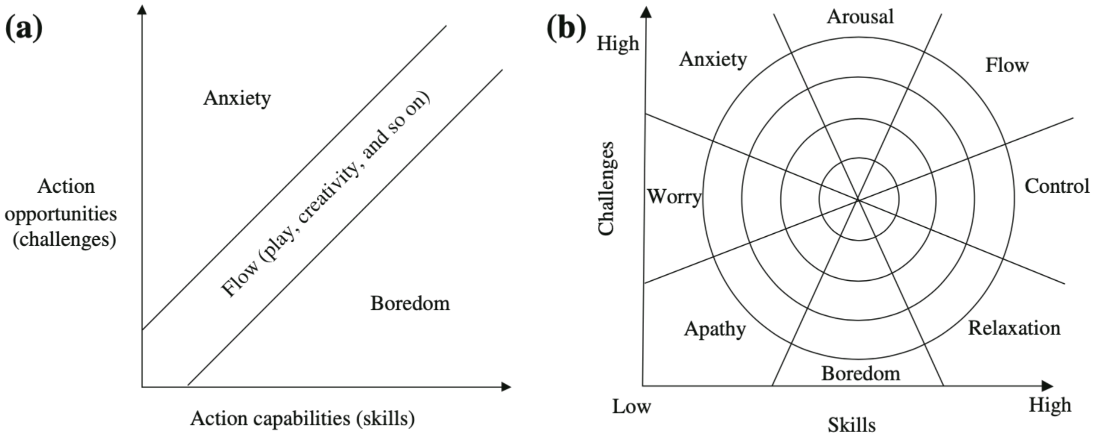


Figure 7: The original and the current model of the flow state. (a) The original model of the flow state. Flow is experienced when perceived opportunities for action are in balance with the actor's perceived skills. Adapted from Csikszentmihalyi (1975/2000); (b) the current model of the flow

state. Flow is experienced when perceived challenges and skills are above the actor's average levels; when they are below, apathy is experienced. Intensity of experience increases with distance from the actor's average levels of challenge and skill, as shown by the concentric rings. Adapted from Csikszentmihalyi (1997). (Nakamura and Csikszentmihaly, 2014, p. 248). Reproduced with permission from Elsevier with license number 4899381231070.

The inclusion of many senses to successfully promote learning of L2 is very important. Each learner perceives these senses or processes as motivational components differently. The theoretical model of the Concept of Flow should help shed light on the importance of including multimodal features into CR models executed by the ECAs in the game. These features will add to the scope of senses learners will use in the game to learn and practise the language. Even though the multimodal CRs represent only one communicative function described at length in this thesis, there are many other communicative functions with verbal and non-verbal features that ECAs execute during an interaction with learners, but these should be subject of further research.

Although there are various theoretical models for evaluating users' general experience in computers, the Integrated Model of Training Evaluation and Effectiveness (IMTEE) (Alvarez at al., 2004) has been chosen as a theoretical reference for evaluating learners' needs (expectations) and the state of flow (experiences) of L2 learners of Icelandic playing *Virtual Reykjavik* with ECAs endowed with multimodal CRs. This is model is very complex. Very briefly, however, this model was developed based on a review of literature on training evaluation and effectiveness published during a period of ten years including seventy-three studies ranging from the field of applied psychology through management to computer science (Alvarez et al., 2004, p. 391). In the context of this thesis, the IMTEE model (Figure 8) suggests that the needs of future users of the game and their experiences should be analysed to gain a better picture of how successful the game is by users. In general, results from the needs analysis can be used in the future to help develop a suitable content and design of the computer game that will enhance changes in learners, i.e. support learning of new language and culture skills. All of this can, however, be influenced by individual characteristics of learners. Factors such as personality traits, attitudes, abilities, demographics, experience, expectations, self-efficacy, goal orientation and motivation that trainees bring to the situation, as well as the context in which training is implemented, can influence the performance and user experience leading to the state of flow (Alvarez et al., 2004, p. 389).

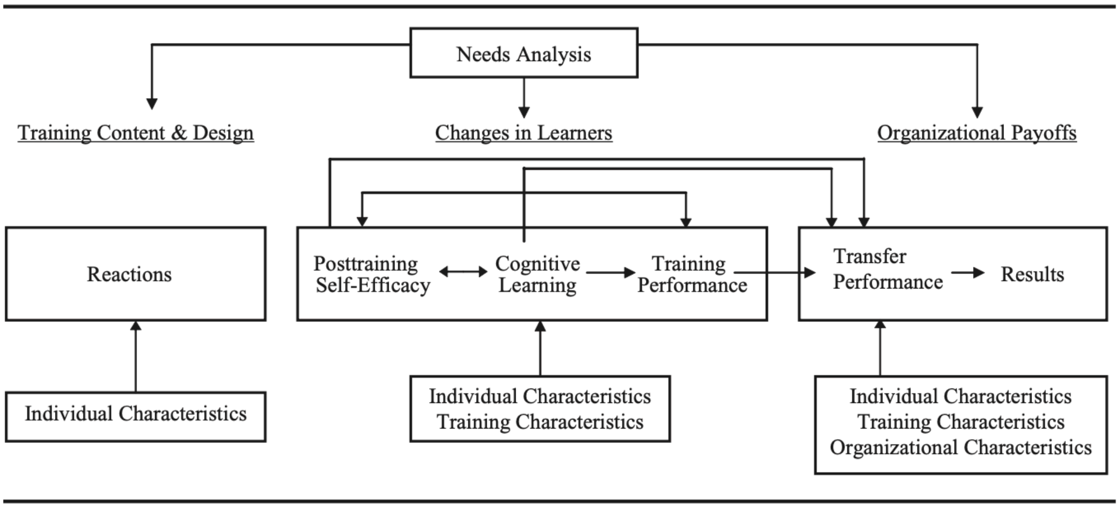


Figure 8: The Integrated Model of Training Evaluation and Effectiveness (Alvarez et al., 2004). Reproduced with permission from SAGE under gratis reuse for doctoral dissertation.

In the context of this thesis, two auxiliary studies were conducted, a survey about learners' expectations (section 3.2) and a pilot study about learner's experiences (section 3.5). These two studies correspond with the general idea of this IMTEE model in that they inform about the needs and experiences of learners in *Virtual Reykjavik*. The IMTEE model advises to conduct further empirical studies. The two auxiliary studies were, however, conducted only with a small number of participants and therefore cannot be considered as part of an empirical research. They can, nonetheless, inform about the background and motivation for the larger project *Virtual Reykjavik* and offer a preliminary evaluation of users' needs and experiences with the project. In the context of including multimodal CRs into the ECAs behaviour, evaluating a general experience of users when playing the game can shed light on whether the two CR models implemented met the goals intended, i.e. whether learners found them useful when practising the language with the agents in the game. Although the pilot user study is a preliminary study and provides only a micro view of training results, it nevertheless provides some information about the learners' experiences with playing the game populated with ECAs endowed with two multimodal CRs. Further training evaluation with a more improved game and user study is necessary in the future. It will contribute to the methodological approaches used for measuring learning outcomes in Second Language Acquisition (SLA) field. A combination of originally developed questions for qualitative and quantitative assessment tools for measuring particular learners' language skills in and outside of a classroom setting (Pellettieri, 2011; De Paepe, 2018) can help inform about the learners' progress in *Virtual*

Reykjavik. Combining the above model with the traditional assessment approaches in the SLA field would help inform about both the learners' progress in learning while playing the game and the reactions to the game design and its functionalities compared to the learners' needs.

2.5 Summary

This chapter provided an overview of various approaches, theories, characterisations and definitions that have informed about the theoretical framework for designing a realistic human-agent interaction in *Virtual Reykjavik*. This thesis has placed the game within the ICALL domain. Relevant teaching and learning approaches were discussed followed by the challenges of designing a realistic human-agent interaction in serious computer games. The challenges included both the technologies that are necessary to enable such interaction and the research in natural language which provided data on the use of multimodal features in specific communicative functions. In this context, natural language was described as a complex phenomenon with multimodal cues. The theory of embodied cognition and multimodal grounding were mentioned as a supportive framework for designing a realistic human-agent interaction. Definitions of CR functions and its various types were discussed, which introduced the main research focus in this thesis. Throughout this CR section, the emphasis was on natural language, which is a necessity for collecting authentic data. The CR strategy may be produced and perceived differently in different languages and cultures, and previous CR research in Icelandic was reviewed. Afterwards, the multimodal realisation of CRs in 2D and 3D interfaces was discussed and the definition of ECAs introduced. However, in order to endow ECAs with multimodal CRs in a realistic way, one has to work with two types of technical extendable mark-up languages (XMLs), FML and BML. In order to understand the motivation of learners for using *Virtual Reykjavik* to enhance their spoken language skills, the Concept of Flow was introduced. The theoretical model of this Flow Concept should shed light on the importance of including multimodal features into CR models executed by the ECAs in the game, because the more objects with physical/living characteristics are implemented into the game environment and interact with the player, the more realistic it becomes for the learner to play, which eventually leads to a more natural experience in learning spoken language skills in virtuality. In order to evaluate general experiences of users when playing the game can inform about whether the two multimodal CR models implemented met the goals intended, i.e. whether learners noticed multimodal features and whether the learners found these features useful when practising the language with the agents. The IMTEE model is used as a theoretical background for doing a brief needs analysis and evaluation of the pilot study of *Virtual Reykjavik* included as two auxiliary studies in this thesis. The next chapter will focus two auxiliary studies and one main study on multimodal CRs that informed the effort behind creating *Virtual Reykjavik* and a realistic human-agent interaction.

3 The Studies

3.1 Introduction

This chapter presents three studies, two auxiliary studies and one main study. The theoretical background guiding the two auxiliary studies is based on the Integrated Model of Training Evaluation and Effectiveness (IMTEE) (Alvarez at al., 2004) (presented in Ch. 2.4.6.). The theoretical background for the main study is the Multimodal Approach (presented in section 1.2 and section 2.2.3.4). The first study is an auxiliary study that was conducted in a form of survey to determine the needs learners of Icelandic have in the country of the target language - Iceland, and what expectations they have from a 3D computer game for teaching Icelandic. This survey supports the rationale of this thesis presented in the Introduction chapter. The second study informs about research in natural language during which CRs were examined in the context of first unknown encounters, by asking for directions. It presents findings about the types of CRs and the multimodal features used in this conversational context. Results will be presented in a separate section. Based on the results from this study, six multimodal CR models will be suggested and described in another section. The third study is an auxiliary pilot study about learner's perception of and experience with playing *Virtual Reykjavik*. This study informs about the general user perception of *Virtual Reykjavik*, as well as the perception of two multimodal CRs that were implemented into the multimodal behaviour of ECAs in the game. This will be presented in a separate section.

3.2 Survey about Learners' Expectations from *Virtual Reykjavik*

This survey is the first auxiliary study, or a preliminary needs analysis, which is based on the theoretical notion of the IMTEE model (Figure 8). This model suggests that the needs of future users of artificial systems, in our case the game *Virtual Reykjavik*, should be analysed to gain a better insight into what needs and expectations do the users have from the game. Due to the fact that the project *Virtual Reykjavik* had already started without conducting any such study, there was a gap caused by missing this information. The present survey was done by a personal initiative. Due to time and work constrains, it was short in length and included only a small sample of participants. This survey can be considered a pre-study getting a preliminary response form possible future user of *Virtual Reykjavik*.

The questions for the survey were developed in order to find answer about the needs for such application. The questions included in this survey will contribute to the developing of a larger study in the future, which would examine the needs and expectations of L2 learners of Icelandic in a more rigorous way. The current survey is a first step toward it.

3.2.1 Methodology

This survey was conducted in order to find out what L2 learners of Icelandic at the University of Iceland expect from a 3D computer game for teaching Icelandic language and culture, and to determine the needs learners of Icelandic have for practising the language in 3D computer game. This short survey also pointed at the difficulty of practising spoken language skills in real life in Iceland, and gathered information about the language skills the learners expect to practise in the game. This survey included beginner to intermediate learners of Icelandic attending a traditional language course at the University of Iceland. The purpose was to determine the following:

7. Learners' country of origin;
8. What language skills learners want to practise in the language course;
9. What language skills they expect to practise in a 3D computer game for Icelandic, which has virtual characters that are able to speak;
10. What elements such a 3D computer game should contain to make it enjoyable for them to play and learn;
11. What would be the advantages and disadvantages of such a game;
12. How do learners feel about using Icelandic in a face-to-face conversation with local native speakers?

These questions helped to design the survey which was distributed online in the early stage of the project *Virtual Reykjavik*. The design and questions used in the survey were not based on any standardised set of questions used in a previous research. Instead, an original set of questions were created. Appendix A provides the list of all questions used in the online survey. The reason why the questions were created and not adapted from a previous study was based on the purpose of this survey. Using another standardised set of questions form a reliable and valid research would be practical, however, it would not measure exactly the intention of this study. The questionnaire in this auxiliary study was designed so as to include open ended questions that are suitable for a qualitative research and two questions with multiple choice answers that included the exact selection of choices we wanted the learners to compare. Creating questions about learners needs for a future

ICALL application would produce evidence about learners' opinions concerning the value of the tasks in *Virtual Reykjavik* relative to what they need to be learning in the classroom (Chapelle, 2001, p. 90). By creating specific questions, one can more precisely focus on the concerns and theory in one's research, and can test the ideas or methods used in that research (Maxwell, 2012, p. 72). Creating the first set of questions in this questionnaire was done as the first step on a long journey of developing, reflecting upon, and refining questions to achieve a more valid and reliable questionnaire in the future (Agee, 2009, p. 432).

3.2.2 Data collection and analysis

As the purpose of this study was to collect information from participants about their preferences and opinions, and generalise these results based on the responses, the design of a quantitative study (Creswell, 2009) was chosen. The online survey included both closed and open questions. For the pragmatic part, it was administered through the Internet Google Forms and distributed *via* email to undisclosed recipients of group of students of Icelandic L2 at the University of Iceland. The following group of participants was addressed: L2 Learners of Icelandic at University of Iceland that are enrolled in the practical diploma and the BA programme for Icelandic for foreign students. A group of 202[31] students was addressed. Learners could participate anonymously. The results were analysed by using a mixed-method perspective. This means that charts with values of participants' preferences were created and their written answers were arranged into thematic circles, towards helping to better interpret the results. This survey was conducted only during one point of time (the summer semester of 2012) and lasted for two weeks before closing.

A random sampling method (Creswell, 2009, p. 148) was applied. Here, participants within this group had an equal chance to answer the questionnaire. All data were anonymous, i.e. without any personal details. The online survey aimed at addressing 10% of the students and was achieved. Twenty-one adult learners took part in this online survey. Respondents were from 20 different countries, comprising 18 females and 3 males, aged 19-40, either beginner or intermediate learners of Icelandic. Answers were collected online using Google Forms, with protected access. Answers were automatically structured

[31] This information is based on confidential data from the Student's Registry Office at the University of Iceland.

into Google Sheets, from which an Excel document was created. A quantitative approach using a questionnaire was used for data analysis. The data from the first two questions were ordered into an Excel document and charts were created to view and compare the responses. The data were interpreted according to the responses from the survey and percentage to each value was assigned in the chart. The answers from open questions were analysed qualitatively. They were ordered in one document and categorised according to common concepts.

3.2.3 Results from The Survey

For a better orientation, the results were divided into three categories that corresponded with the questions in the survey: 1) what elements learners expect from the game, 2) what advantages and disadvantages they expect the game to have in learning Icelandic, and 3) how do they feel about using Icelandic when speaking face-to-face with native speakers.

The first was an open-ended question that provided a long list of answers. These can be summarised as follows. The survey indicated that learners, among others, expect the game to have a good storyline with particular conversational scenarios for speaking practice, a voice recognition in order to be able to communicate with virtual agents that would assume the role of native speakers and interact with others in the game. They moreover expect the agents to be funny and entertaining, perhaps be able to tell a joke, and to give feedback on grammar and the learners' choice of vocabulary, and to learn practical information about the city.

The second part provided a list of advantages and disadvantages the learners expect from the game. Building confidence by interacting with virtual agents and an ability to focus on Icelandic without having to switch to English were among the advantages given for the game. However, the ability of the agents to give less feedback on grammar and the choice of vocabulary compared to the teacher, resulting in more laborious work with a dictionary and individual learning in the traditional class was seen as a disadvantage.

The third part informed about learners' using Icelandic face-to-face with native speakers. Fourteen out of twenty-one learners (67%) felt negatively about it, but seven out of twenty-one (33%) felt good and comfortable about it. Even though most of the respondents reported being insecure, distressed and intimidated (to mention selected phrases) when using Icelandic with native speakers in real life, they did indicate that the game would be a practical tool for practising Icelandic with virtual agents before using the language with real speakers in real life, and thus serve as good preparation. In the chart below, learners indicated what language skill they expected to practise in a language classroom compared to the game. They expect to practise more vocabulary, grammar, listening and cultural understanding but less speaking in *Virtual Reykjavik* than in the traditional classroom (Figure 9). The results about speaking indicate only a small difference, which is a positive sign for the game.

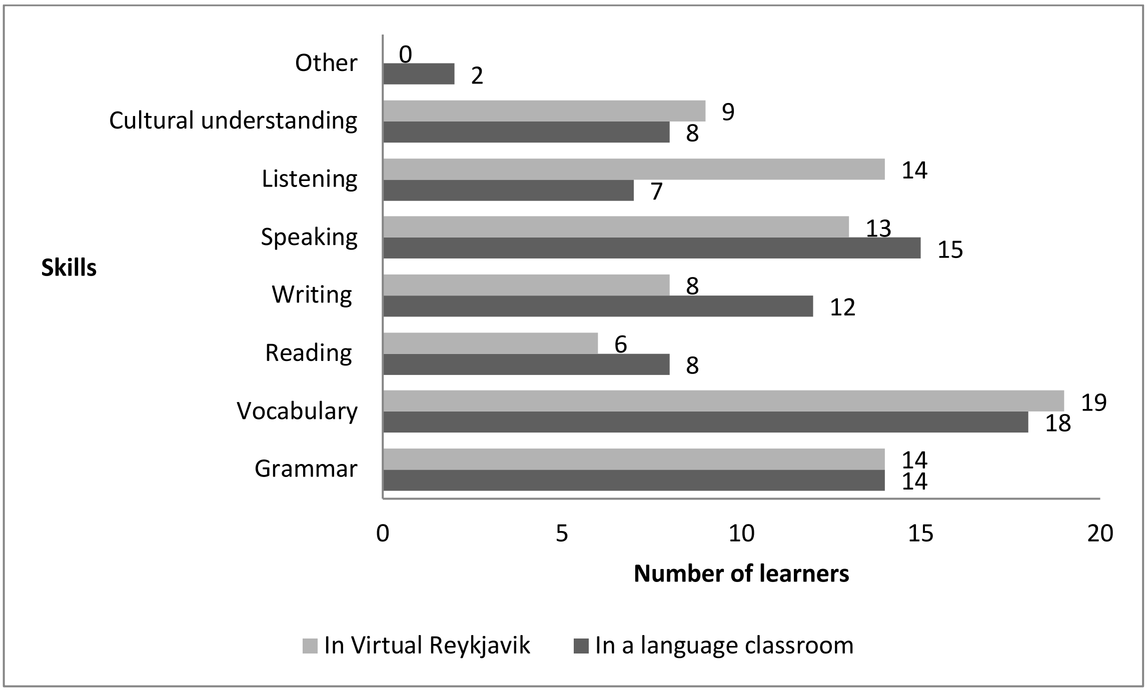


Figure 9: What language skills learners expect to practise in a language classroom compared to *Virtual Reykjavik*.

3.2.4 Discussion of Results from The Survey

This study informed about including virtual agents into the game that would act as local people and the agents be humorous. Importantly, learners expected to focus on speaking Icelandic without switching into English.

Twenty-one students of Icelandic participated in the survey. This means that the target of addressing 10% of the students enrolled in the Icelandic courses (202 students in the academic year 2012-2013) was reached, namely 10,4%. The results indicated that learners expect to practise similar language skills in *Virtual Reykjavik* as in a traditional language classroom. In the game, however, they expect more to practise cultural understanding, listening, and vocabulary. The learners expected the virtual characters to be funny and be able to tell a joke. There should be a speech recognition system implemented to enable learners to communicate with virtual characters that would take on the role of native speakers. Overall, the game should be enjoyable for playing, while helping learners to keep focused on the target language without switching into English. As most of the learners indicated they feel insecure, distressed and intimidated (to mention just a few examples) when speaking Icelandic with native speakers in real life, the game should bridge this gap and become a practical tool for practising Icelandic spoken language with virtual agents. This will prepare the learners better for using the language in real life. This auxiliary study shed light on the need for having realistic agents in the game who can act as local people, be funny, but most of all, enable learners to be focused on speaking Icelandic without switching into English.

This study, however, did not include any questions about multimodal behaviour, since it was not the purpose of it. Questions regarding perception of multimodal behaviour of ECAs will be included in the auxiliary pilot study. The study in the following section examines multimodal features in CRs that will contribute to a more realistic design of ECAs in the way they speak in *Virtual Reykjavik*.

3.3 The Study on Multimodal Clarification Requests

This section presents the study on the multimodal realisation of CRs in real life interactions. The data provides a foundation for endowing ECAs with human-like characteristics in one communicative function, the CR. In order to model first encounters between ECAs representing native speakers of Icelandic and the players representing L2 learners of Icelandic, data from real-life interactions is needed to help model a realistic human-agent interaction in *Virtual Reykjavik*. A study in a natural language setting will shed light on what kind of CRs native Icelandic speakers produce in a specific conversational setting, and which multimodal features they use in their production. The specific setting represents an unknown first encounter, native and non-native speakers, men and women, aged between 18-70, asking for directions to a particular place in central Reykjavik. Actors, who were both native and non-native speakers of Icelandic, were engaged to ask native speakers for directions in Icelandic. Participants were divided into two groups, native-non-native (focus group) and native-native (control group) speakers. This enabled examination and comparison of different speaker pairs in order to document the nature of production of multimodal CRs. The following two variables were chosen as a basis for the data set: 1) gender and 2) whether the initiators had Icelandic as their first or second language. The task of the actors was to select a “likely subject”, who would be a native Icelandic speaker, male or female. Short social interactions were captured by video camera. Each actor followed one scenario - to ask local people for directions to a particular location in central Reykjavik. This represented the same conversational setting which was designed in the first scenario in *Virtual Reykjavik*. Participants, who were not native speakers of Icelandic, but spoke fluent Icelandic, were excluded from this research because the aim was to collect data on CRs solely from native speakers. In order to determine that local people were indeed native speakers, they were asked whether they were local from Reykjavik. It was expected that different actors would ask in different ways for directions. This contributed to the variety of questions and answers used, and thus, possibly, to different kinds of CRs. Subsequently, the responses of native speakers to the actors' questions, when the native speakers did not understand what the actors have said, were multimodally analysed. In this way, this study features both the linguistic forms of CRs and their multimodal realisation. Figure 10 demonstrates how the data was collected in the field and how the scene in *Virtual Reykjavik* appeared. Note that both feature the same place, Austurvöllur Square, in central Reykjavik. In the upper panel, an actor is approaching native Icelandic speakers and asking for directions to a particular place in central Reykjavik. In the lower part of the picture, this situation is simulated by the learner's avatar in the game scenario in *Virtual Reykjavik* (lower panel).

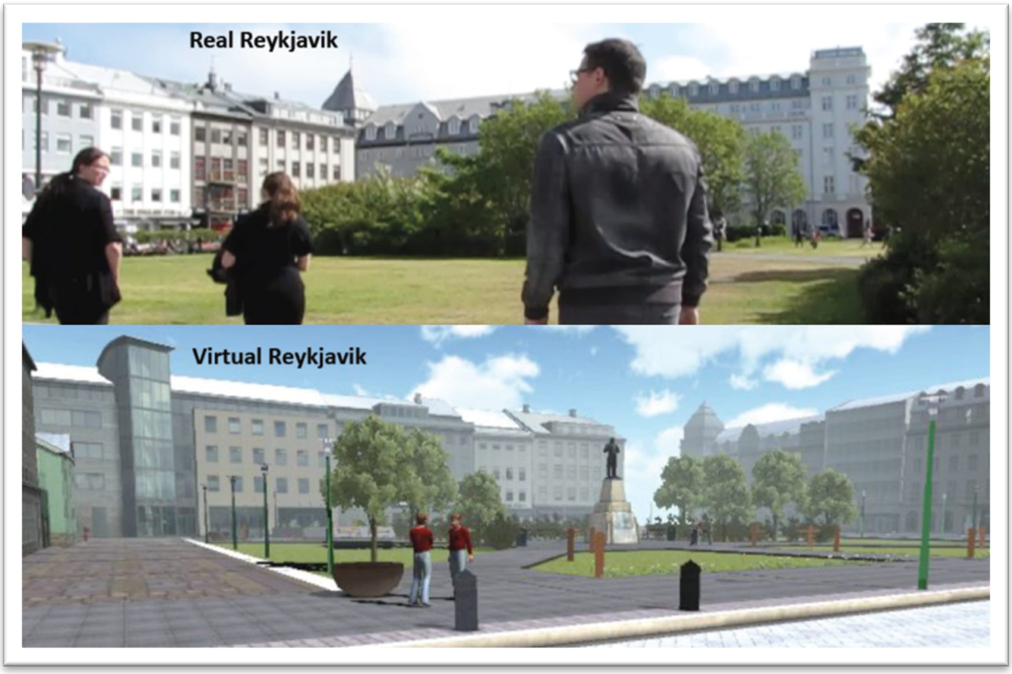


Figure 10: Data collection in real Reykjavik (top) and the implemented data in *Virtual Reykjavik* (bottom).

3.3.1 Research Questions

Three multifaceted research questions guided this study:

1. In a given scenario of asking for directions in central Reykjavik, what are the commonly used clarification requests native Icelandic speakers (random men and women, aged 18-70) make in a face-to-face interaction with other native and non-native speakers (actors, men and women, aged 18-70), in order to initiate speech repair?
2. Are there any differences between gender[32] and native speakers (NS) and non-native speakers (NNS), or in other words, can a difference be identified in clarification requests when considering the following pairs:

   Focus group:

   - Male native speaker (MNS) – male non-native speaker (MNNS)
   - Male native speaker (MNS) – female non-native speaker (FNNS)
   - Female native speaker (FNS) – female non-native speaker (FNNS)
   - Female native speaker (FNS) – male non-native speaker (MNNS)

   Control group:

   - Male native speaker (MNS) – male native speaker (MNS)
   - Male native speaker (MNS) – female native speaker (FNS)
   - Female native speaker (FNS) – female native speaker (FNS)
   - Female native speaker (FNS) – male native speaker (MNS)
3. Which of the verbal and non-verbal features used by native Icelandic speakers during CRs are critical for implementation into the multimodal behaviour of virtual humans in order to simulate authentic interaction? Or in other words, which of the multimodal features found in human CRs are most representative of natural interaction and therefore candidates to build a model for CR that can be implemented into the AI of ECAs, in order to simulate natural conversation in virtuality?

[32] Carli (1989) suggests that difference in gender affects partner behaviour in interaction, e.g. men interacting with other men have less effective influence on each other than when interacting with women (p. 565). Similarly, Ridgeway and Smith-Lovin (1999) see difference in interaction between genders: gender difference perpetuates status beliefs leading men and women to recreate the gender system even in everyday interaction (p. 191).

3.3.2 Methodology

3.3.2.1 Sampling Method. The sampling method was chosen in line with the research design. According to Schatzman and Strauss (1973), the Selective Sampling Method is a 'practically-oriented' method. This allows researchers to select representative sample of participants according to the aims of the study, or the researcher's available time, the pre-set framework, starting and developing interests, or by any restrictions that are placed upon the observations by his/her hosts. They furthermore suggest that after several visits to the sites, the researcher will know how to proceed with sampling, i.e. who to sample, during which time and at which location, what events and people to consider (Coyne 1997, p. 624). Even though the description of selective sampling resembles Patton's (1990) Purposive Sampling, Coyne (1997) argues that it is different in how objects are chosen. For example, in purposive sampling, information-rich cases are selected for in-depth analysis in order to help the researcher learn about the issues of central importance, e.g., in interviews. In selective sampling, specific locales are selected according to an initial set of prerequisites, such as time, space, gender, etc. (p. 624).

The sampling strategy here was executed according to the initial set of prerequisites (Table 5). The task was to capture the micro social interaction on video by using an automatic photo/camera Canon EOS and select participants, who were native speakers of Icelandic and local to Reykjavik, and ask them for directions based on the following criteria:

1. The actor was asked to remain natural throughout the whole time; his/her task was to ask local persons for directions in central Reykjavik in whichever way he/she would normally feel in that moment or do in real life. The actor was told to select a native speaker that "looked Icelandic" (i.e. was not wearing Gor-Tex® or alpine clothing or a large travel backpack that is more typical of tourists), or headwear and sunglasses that would prevent the camera from capturing facial features, eye and head movements;
2. The native Icelandic speaker would appear available to talk to (i.e. would not seem to be preoccupied by any other activity, such as using a mobile phone, smoking a cigarette, encountering a different person, or did not seem busy in any other way;
3. The native Icelandic speaker would appear sober, i.e. not under influence of drugs or alcohol, and capable of holding a conversation.

Table 5: Prerequisites for data collection and sampling for the second study on multimodal CRs.

Data Collection and Sampling		
Situation	Unknown first encounters	
Participant	Speaker	Actor
Speaker of Icelandic	Native	Native & non-native
Gender	Male	Male
	Female	Female
Age	18-70	18-70
Role	Local speaker	Actor
Task	Being asked for directions	Asking for directions
Sampling preference	A local person "Icelandic looking" (not a tourist with alpine clothing or a backpack with training shoes on), preferably not wearing sunglasses or a head cover, not using a mobile phone, not smoking a cigarette	Native and non-native speakers of Icelandic, male and female, hired for asking for directions had eventually age range between 20-40 years

3.3.2.2 Participants. There were two groups of participants in this study: native Icelandic speakers and the actors. The native speakers were adult native speakers of Icelandic, men and women, aged 18 to 70, who were approached by actors chosen to initiate conversation with the native speakers. The actors were volunteers, who were both native and non-native speakers of Icelandic, men and women, aged 20-40. Their ages were not registered and therefore an official average age cannot be precisely given. Nonetheless, based on a conversation with each of them the average age range would be estimated to 30 years. For this reason, the age is not further analysed. Even though originally the plan was to have actors in various ages representing age groups in the scale between 18-70 years, only those actors, who accepted voluntary work, were selected to participate and they had a narrower age range. The age range 18-70 was chosen because 18 marked the legal age, which allowed us to ask for permission of recording directly and not a legal representative of the person in case they were minors. The upper age 70 indicates persons over retirement age in Iceland. This age range had been given for visual purposes to concentrate on finding speakers of various age groups. The ages of participants were not registered and therefore an official average age cannot be precisely given. Nonetheless, based on appearance the average age range would be estimated to 35 years. For this reason, the age is not further analysed. The focus, however, was only on the approached native speakers, because the aim was to analyse how they produced CRs in interactions with the actors. Even though it would have been interesting to analyse the multimodal behaviour of both participants at the same time when interacting with each other, it was not the focus of this research because only the multimodal behaviour observed on the native speakers would have been used for modelling the behaviour of the ECAs in the game. The ECAs will not have any detection of user's face or body. Table 6 shows how different speaker pairs, i.e. male native speaker (MNS), male non-native speaker (MNNS), female native speaker (FNS), and female non-native speaker (FNNS), were divided into two groups. All speaker pairs were colour-coded throughout the research towards keeping a clear track of them.

Table 6: Native and non-native speaker pairs divided into two groups, focus and control. Number of aimed and achieved video recordings is presented for each of the groups.

Speaker pairs					
Focus group			**Control group**		
Speaker pairs	No. of video recordings		Speaker pairs	No. of video recordings	
	Aimed	Achieved		Aimed	Achieved
MNS-MNNS	25	22	MNS-MNS	10	17
MNS-FNNS	25	32	MNS-FNS	10	10
FNS-MNNS	25	23	FNS-MNS	10	20
FNS-FNNS	25	31	FNS-FNS	10	10
Total	100	108	Total	40	57

3.3.2.3 Data Collection. The data were collected in central Reykjavik. Firstly, a pilot study was conducted with an audio recording device iPhone 4 using the application for data organisation 'Audio Memos/Voice Memos'. The aim of the pilot work was to map the local area, look for the optimal places and times for recording, to gauge people's willingness to be recorded, their reactions before and after the recording, detect what language they use and how their reaction is towards the actor and the cameraman. 32 anonymous audio recordings were collected that are excluded from the proper data analysis. A journal was kept throughout the whole time, and field notes were written after each trip to the field, which helped to understand the situation better. As a result, the optimal area for recording was found to be either at Austurvöllur Square and its surrounding streets, or in the main pedestrian zone Bankastræti Street and Laugavegur Street. In this area, numerous buildings could serve as a landmark for asking for directions. Moreover, when moving around Reykjavik, there was more of a chance to find participants than when only remaining in one place. Each field trip lasted about two hours; therefore, it was important to get as many recordings as possible. According to the pilot data collection, the optimal time for recording was found to be around lunch time (±2 hours), because several local people left their workplace and moved around Reykjavik. Generally, people showed a great willingness to participate in the study. Many of them also showed their moral support and wished the researchers good luck with the project. Their reaction was most of the time very positive. There were, however, only a few participants who did not want to participate because they did not have the time to stop and talk to us or did not agree to being recorded. Figure 11 shows the area in central Reykjavik where recordings were collected. Marked are pathways where recordings took place. For a better orientation, Austurvöllur Square is also highlighted in the image.

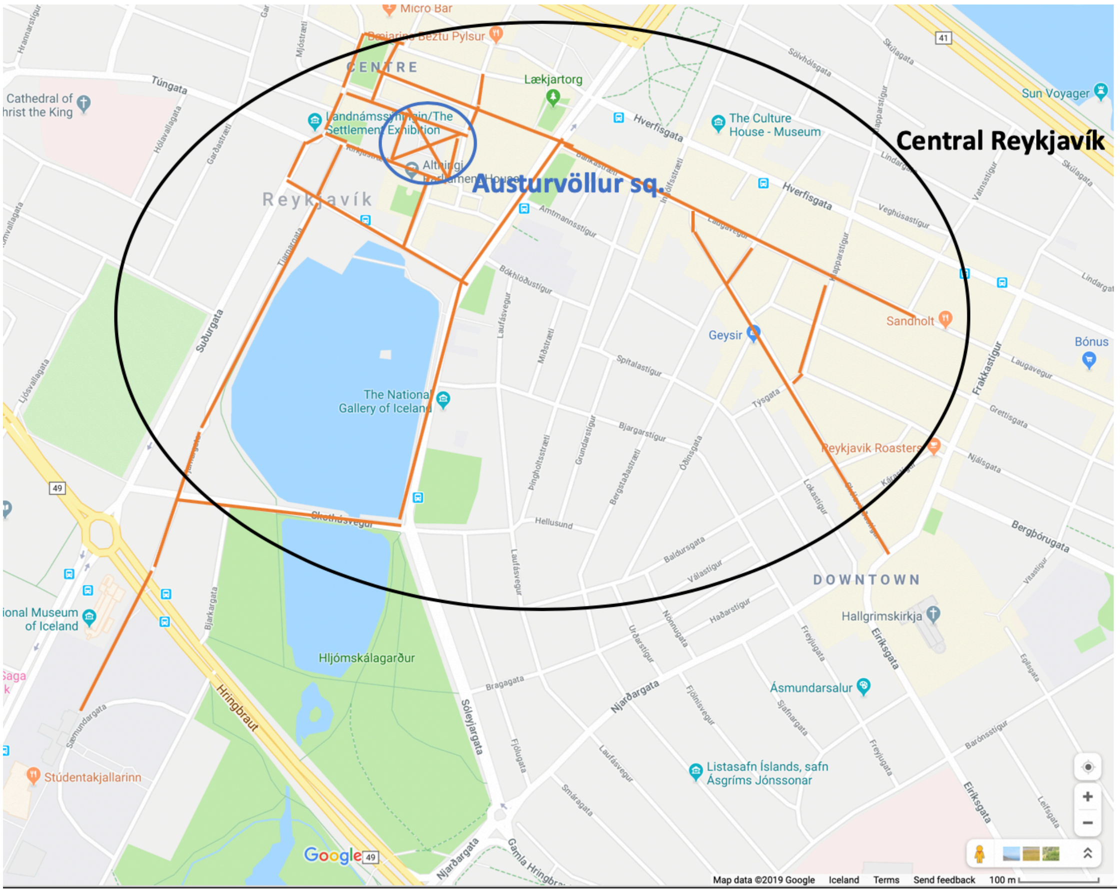


Figure 11: Pathways for data collection in central Reykjavik. The map itself is a screenshot from Google Maps. The red lines mark the pathways for data collection in central Reykjavik; the blue circle marks the Austurvöllur sq. where the game is situated; and the black circle marks the area of central Reykjavik.

This pilot work showed that when people know beforehand that they are being recorded, they immediately change their body posture to some kind of a 'being-prepared-and-ready-to-talk' one; they become friendlier, smile, and are unnaturally very polite. Consequently, this changes their tone of speech and choice of words to a more formal one. Some participants even expressed their concern about making possible mistakes when being recorded, because they wanted to sound and look in an ideal way. This brief pilot study offered an insight to both the geographic and social environment, which was necessary to understand. No results from the pilot study are presented here because the evaluation consisted of comparing the research journal and field notes with the audio data.

The proper data collection was carried out immediately after the evaluation of the pilot study. Video recordings were made with the Canon EOS video camera. In the beginning, the consent from participants was collected before the recording, but soon it was realised that this action affected people's behaviour. Consequently, the consent was collected after the recording. The participants had the opportunity to see or hear the relevant section of the recording, and, if they wished, they could let the recording to be immediately deleted from the camera device. All participants received information about the purpose of this study and that the use of the material only for the research purposes without publishing their photograph that would permit identification. The consent from the native speakers was collected verbally and recorded on camera. The consent from the actors was collected in writing.

The data was collected according to the preliminary number of video recordings assigned to each speaker pair, i.e. 25 for each NS-NNS pair and 10 for each NNS pair (Table 6). This would represent 100 video recordings for the NS-NNS pairs and 40 video recordings for the NNS pairs. The number of video recordings achieved for NS-NNS pairs was 108 and for NNS pairs was 57. After reviewing the recordings and seeing the quality of captured data, which was clear enough in both picture and sound for analysis, 22 video recordings for each speaker pair in the focus group were deemed sufficient and for this reason there was no need to collect any further data. In the control group, however, the number of videos for each subject pair was lowered to 10 as advised above. A great volume of sampling data was not needed in this group because even only a small data set could determine whether there are any differences compared to the focus group.

According to Jewitt (2012), video recordings can lead to overwhelming amounts of rich video data and there is no universal 'right amount' of video data to collect, because it depends on the chosen research approach, aim and questions of a study, and pragmatic questions of time and resources (p. 18). Learning also from other researchers, Veer (2013) suggests that the aim is to collect data suitable for analysis, rather than a large amount of data that is difficult to analyse (p. 218). Similarly, Parry (2010) suggests that even though video recordings permit collection of large data sets, it is not imperative to analyse all of the data in depth, but focus on a sufficient volume to allow subsequent sampling from within the dataset depending on emerging questions and issues (p. 379). Similarly, Heath et al. (2010) address *How much video data is enough?* by saying that it depends on the nature and demands of the setting, the action and activities that are being addressed and, most importantly, the methodological commitments that inform the collection and analysis of data (p. 59). In this view, data saturation is reached when the following three points are reached: (1) there is enough information to replicate the study, (2) when additional new information is obtained, and (3) when further coding is no longer feasible (Fuchs and Ness, 2015, p. 1408). Here, data saturation was reached after the video recordings had been analysed. Several video recordings obtained were rich in quality in both sound and picture, which helped to perform an in-depth multimodal analysis of participants' behaviour when producing CRs. A repeating pattern in the production of different types of CRs was detected. Moreover, the data from the focus group was compared with the control group and it showed similarities in the native speakers producing CR. The number of collected video recordings for each group and the analysed data provided a sound representation of saturation. The following section describes the data analysis.

3.3.2.4 Data Analysis. Multimodal Interaction Analysis (Norris, 2004, 2013) was chosen as a method for data analysis, since it is mainly concerned with the human being in an interaction. How a person displays structures of various higher-level actions, such as meeting and greeting someone that consists of a chain of different utterances in a conversation between people, in hierarchical order through the employment of communicative modes (Norris, 2004, p. 106). Interaction Analysis from the beginning concentrated on both linguistic and non-linguistic aspects of spoken language and "attempt(ed) to articulate links between the linguistically-focused rhetorical routines and social aspects of interaction" (Nunan, 2005, p. 161).

As a method for data collection and analysis, videography (Knoblauch, 2012; Jewitt, 2012; Knoblauch and Tuma, 2011), also known as videoethnography (Veer, 2013), is used. This method combines video interaction analysis (Kissmann, 2009) with ethnography (Baszanger and Dodier, 2004) that collects observations *in situ* of concrete sequences of activities. Nowadays, it focuses on demonstrating the relationship between forms of heterogeneous actions, rather than trying to identify a culture as a whole (Baszanger and Dodier, 2004, p. 9). For instance, in order to collect large amount of data and analyse from different perspectives, Veer (2013, p. 215) suggests combining video recordings and ethnography for the following reasons:

> *(…) a narrative analyst would be able to incorporate tonality and inflexion far more effective into his/her analysis; a body language analyst might focus on the role of subtle body movements into their analysis; an ethnographer can develop a fuller appreciation for a culture as a whole, by taking a varied approach to the site and analysing data from multiple perspectives.*

This approach to video-recorded micro-social interaction in various natural settings can also be characterised as interpretative. Veer (2013) suggests that "[a]n ethnographer needs to be able to take an entire cultural setting, understand the setting and focus in on the nuances that make the site both interesting and relevant to a wider audience" (p. 218). In order to demonstrate the relationship of heterogeneous action, such as multimodal production of CRs in conversations between native and non-native speakers of Icelandic asking for directions, this approach was selected as the most suitable. Moreover, observations and ideas gathered through ethnography and fieldwork are very informative and insightful, and that is why Heath et al. (2010) also suggest that if a researcher wants to show their relevance to analysis, he/she has to do so within the situated and interactional accomplishment of the participants' actions (p. 107). For this reason, videos and observational notes, together with a research journal, support the data analysis in this study by describing the context in which an action was taken.

In order to start with the analysis, all video recordings were labelled. The label showed the video corpus sequence number, the actor's initials, the speaker pair, the type of file, the recording sequence, and the file format, e.g.: 1_BB_FNS_MNNS_MVI_0001.MOV. The content of each video was transcribed with a professional tool for the creation of complex annotations on video and audio resources, called ELAN (Sloetjes and Wittenburg, 2008). In the particular tier for intonation, it was marked by words: fall, rise, rise-fall, fall-rise, or level. In order to mark particular parts of a dialogue with codes that best represented what was going on in the recorded section, Saldana's (2009) coding system was adopted. It consists of simple words and phrases, and allows to use particular codes in different data files, such interview transcripts, participant observational field notes, journals, documents, literature, artefacts, photographs, video, websites, e-mail correspondence, and is defined by Saldana (2009, p. 3) as follows:

> *A code in qualitative inquiry is most often a word or a short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data.*

On the basis of this, a simple coding system was applied to the transcribed video sections in Elan, as well as in the observational notes and the research journal of this study. This allowed for tracking the CR sequence throughout the different documents. For example, the code *Question* was used to mark the preceding part of the repair turn that was marked with the code *CR* (clarification request), and the code *Answer* was used to mark the turn, which immediately followed after the repair turn CR. In other literature, however, this sequence is categorized as T-1, T0, T+1[33] (Enfield et al., 2013, p. 346) representing the Other-Initiation Repair (OIR) sequence. The OIR sequence will serve here only for reference purposes (Table 7).

Table 7: Codes in a dialogue to mark different turns surrounding the CR.

Transcription of a dialogue sequence (English translation of Icelandic original)		**Codes used in this study**	**OIR sequential set** (Enfield et al., 2013)
Actor	Fyrirgefðu, veistu hvar Hitt húsið er? (Excuse me, do you know where Hitt húsið is?)	Question	Trouble source (T-1)
Speaker A	Hitt húsið?	CR	Repair initiation (T0)
Actor	Já. (Yes.)	Answer	Repair (T+1)

In the present study, the real-world data consists of micro-social interactions (Knoblauch and Tuma, 2011) between different speaker pairs depending on gender and speaker background. The focus is on native Icelandic speakers only, because the CR communicative function (tech. ReqRepair) and its appropriate multimodal behaviour stand

[33] According to Enfield et al. (2013), the anatomy of other-initiation of repair defines T0 as turn pointing back to a problem in T-1 and point forward to a next turn T+1 where the problem can be repaired.

as a model for ECAs. Attention is given to speech acts that include CR utterances. These are collected and multimodally analysed according to the following features: linguistic types of CRs, suprasegmental features, i.e. intonation, and non-verbal features, i.e. facial expressions, hand gestures, and body posture.

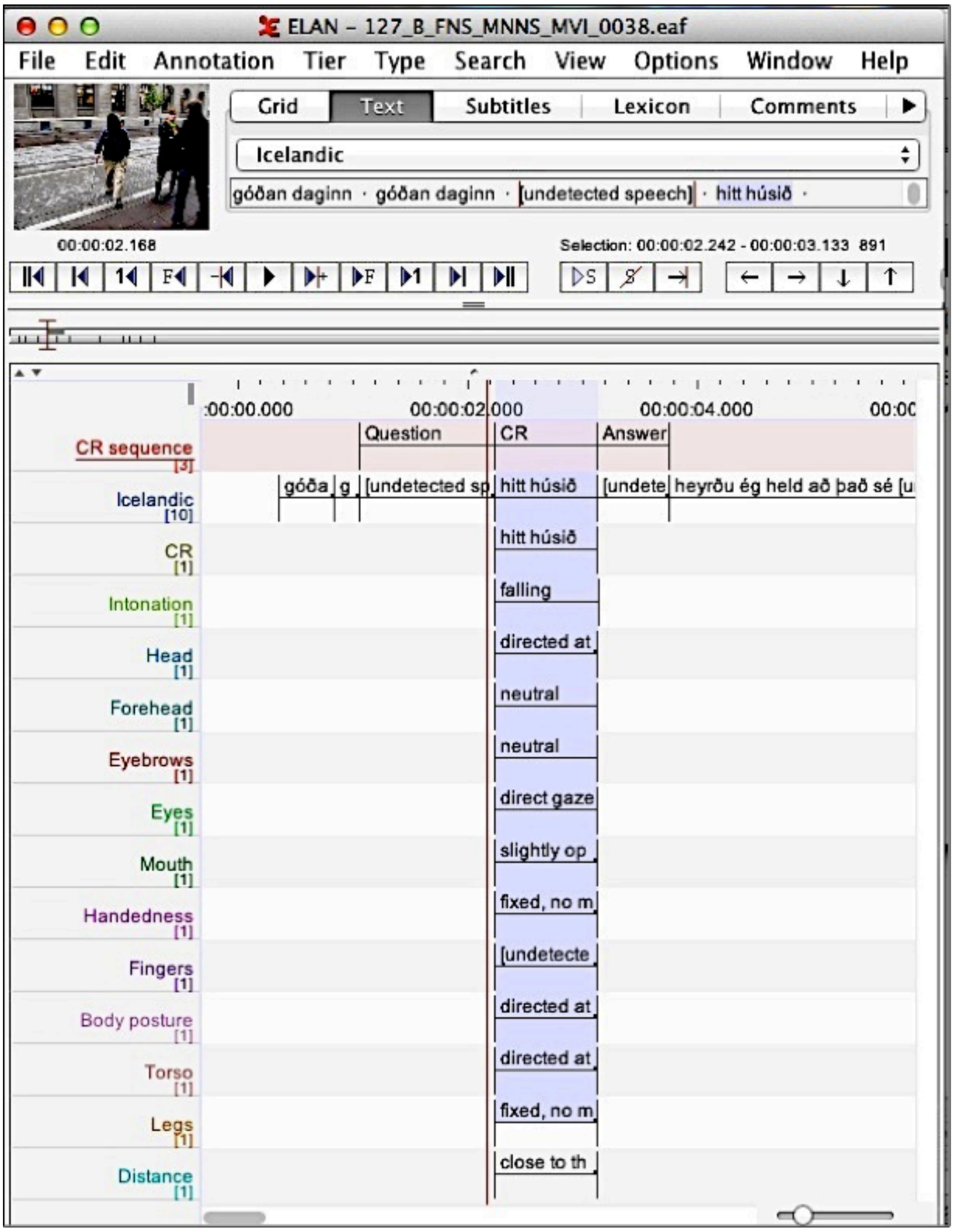
The conversational context is as follows: actors in the recordings had to ask for directions to a particular location. The initial question was not always asked in the same prescribed manner, but the purpose was achieved: ask for directions. Even though the manner of asking a question could have an effect on the native speaker's response, throughout the data analysis, no obvious differences in native speakers' responses in CR utterances were detected that would point at some different way of asking for directions.

Micro-social interaction of only those videos that contained CR utterances was analysed. Videos containing CRs were segmented, transcribed and annotated in Elan. Two coding schemes were developed: one to mark the CR sequence and the other one to provide a scheme for a multimodal annotation. The first coding was a very basic system to mark the CR sequence and was applied in all documents including the Excel overview document, which listed all transcribed dialogues, field notes, the research journal, and in Elan files. The multimodal annotation scheme for *Virtual Reykjavik* (see Appendix B) was used to help annotate multimodal features in Elan. This coding scheme was especially developed for the needs of this study and included both the Behaviour Markup Language (BML) (Kopp et al., 2006) and the Function Markup Language (FML) (Cafaro et al., 2014) as described by the SAIBA framework community. Other multimodal coding schemes, such as the MUMIN (Alwood et al., 2005), SmartKom multimodal corpus (2002) or the HuComTech multimodal corpus annotation scheme (2011), and several others (e.g., Quek et al. 2002; Abrilian et al., 2005; Zwitserlood et al., 2008; ISO/DIS, 2010; Lücking et al., 2011; Abuczki and Ghazaleh, 2013) were used for reference. The *Virtual Reykjavik* Coding Scheme includes various other behaviours that were observed in the video analysis of this study and that were missing in the above literature. This coding scheme was created to meet the needs of this research.

The simplified coding scheme represented following tiers that were needed for the annotation of multimodal data in ELAN (Figure 12). Each tier in the annotating programme ELAN represents a layer, which is annotated by inserting a certain tag. For instance, the tier (layer) for the CR sequence is first segmented into three parts. Each part represents a speaker's turn and is annotated with a particular code, or tag, e.g. “question”, “CR”, and “answer”. Another tier is called “Icelandic”, which represents the transcription of spoken

Icelandic. This tier is segmented into speaker turns and each segment contains the transcribed words. This means that when one layer is put on top of the other, the segments for CR align showing the code for the "CR" utterance and the transcribed speech (words). A similar process follows with other tiers. The multimodal annotation in the ELAN program thus makes it possible to insert tags (various descriptions or values) to each segment of interest. Below is the description of each tier, which was included in the multimodal annotation in ELAN (Figure 12):

- CR sequence: to mark the OIR sequence according to Enfield et al. (2013) as Question (T-1), CR (T0), Answer (T+1);
- Icelandic: for transcription of a dialogue and segmentation according to speaker turns;
- CR: to mark the clarification according to its verbal (transcription of speech) or non-verbal realisation;
- Intonation: to describe the intonation in a CR;
- Head: to describe the movement of the speaker's head;
- Forehead: to describe the movement of the speaker's forehead;
- Eyebrows: to describe the movement of the speaker's eyebrows;
- Eyes: to describe the movement of the speaker's eyes;
- Mouth: to describe the movement of the speaker's mouth;
- Handedness: to describe the movement of the speaker's hands;
- Fingers: to describe the movement of different fingers including the position of palm of the speaker;
- Body posture: to describe the movement of the speaker's body posture;
- Torso: to especially describe the movement of the upper body of the speaker;
- Legs: to describe the position and movement of the speaker's legs;
- Distance: to describe the distance between the NS and the actor according to subjective observation, e.g. very close, close, further away, far away.



Icelandic

góðan daginn · góðan daginn · [undetected speech] · hitt húsið ·

00:00:02.168

Selection: 00:00:02.242 - 00:00:03.133 891

Tier	Annotations
CR sequence [3]	Question CR Answer
Icelandic [10]	góða g [undetected sp hitt húsið [undete heyrðu ég held að það sé [u
CR [1]	hitt húsið
Intonation [1]	falling
Head [1]	directed at
Forehead [1]	neutral
Eyebrows [1]	neutral
Eyes [1]	direct gaze
Mouth [1]	slightly op
Handedness [1]	fixed, no m
Fingers [1]	[undetecte
Body posture [1]	directed at
Torso [1]	directed at
Legs [1]	fixed, no m
Distance [1]	close to th

Figure 12: Example of a multimodal annotation for a CR in Elan.

All annotated data from Elan were exported into a separate Excel document and divided into different sections, each section representing a different speaker pair. This provided a better overview of the annotated data and allowed comparing results between different speakers and speaker pairs. The annotated data were analysed, and the multimodal behaviour interpreted according to how it was observed on speakers in the video. Different categories of CRs were created according to the types of their linguistic realisation found in the study. Each linguistic type included also a description of multimodal behaviour. The multimodal behaviour from all examples listed in each linguistic category was studied and a common pattern was detected for each category. According to the common pattern, a multimodal model for each CR category was proposed.

3.3.2.5 Video Corpus. Video recordings were collected in the field over a period of two years (spring 2014 – spring 2016). All video recordings were downloaded onto a separate hard disc and the access to it was coded, allowing only the researcher to access the data. All video recordings were sorted to particular folders, each folder containing videos for a specific speaker pair. An Excel sheet was created in order to list all video recordings and their transcription of CR utterances. For a better orientation in this Excel sheet, the speaker pairs were colour coded. Table 8 below gives an overview of all the video recordings according to the speaker pairs and the frequency of CRs. The video corpus consists of 165 videos. The total time is 1 hour 59 minutes 2 seconds (01:59:02) and the mean video time is 57 seconds (00:00:57). Of the video corpus, there are 86 videos containing CRs. Some video recordings contain only one CR instance, others two CR instances. The total time of utterances containing CRs is 1 minute and 35 seconds (00:01:35). The mean length of time of a CR utterance is 0.6 seconds (00:00:06).

Table 8: Overview of the video corpus for CRs.

Video Corpus for CRs (total time 01:59:02)							
Group	Speaker pair	No. of collected videos	Video sequence no.	No. of videos containing CRs	Video sequence no. containing CRs	No. of CRs in videos containing CRs	Proportion of videos containing CRs compared to all collected videos
Focus group	MNS-MNNS	22	124, 130, 131, 134, 154, 157, 162, 177, 178, 179, 180, 181, 182, 183, 185, 186, 187, 188, 189, 190, 191, 192	12	125, 157, 177, 179, 181, 182, 183, 185, 187, 188, 189, 190	13	54.55%

	MNS-FNNS	32	48, 49, 50, 53, 56, 57a, 58, 59, 61, 64, 69, 70, 71, 73, 78, 79, 80, 81, 87a, 89, 90, 93, 94, 96, 97, 99, 100, 102, 106, 107, 108, 109	23	48, 49, 50, 57a, 58, 61, 64, 69, 70, 73, 79, 80, 80a, 89, 90, 93, 94, 96, 99, 100, 106, 107, 109	32	71.88%
	FNS-MNNS	23	125, 126, 127, 128, 129, 132, 133, 153, 155, 156, 158, 159, 160, 161, 163, 164, 165, 166, 167, 168, 169, 171, 170	18	125, 126, 127, 129, 132, 153, 155, 156, 158, 159, 160, 161, 163, 165, 166, 167, 168, 171	20	78.26%
	FNS-FNNS	31	51, 52, 54, 55, 57, 62, 63, 65, 66, 67, 68, 72, 74, 75, 76, 77, 82, 83, 84, 84a, 85, 86, 87, 88, 91, 92, 95, 98, 101, 103, 104	22	51, 52, 54, 57. 63, 66, 67, 68, 72, 74, 75, 84, 84a, 85, 86, 87, 91, 95, 98, 101, 103, 104	23	70.97%
	Total:	**108**	**-**	**75**	**-**	**88**	**Mean 69.44%**
Control group	MNS-MNS	17	44, 45, 46, 47, 113, 114, 116, 118, 122, 135, 137, 139 144, 145, 146, 150, 151	4	46, 122, 135, 151	5	23.53%
	MNS-FNS	10	33, 36, 38, 40, 41, 42, 173, 176, 196, 197, 198	2	173, 198	3	20%

	FNS-MNS	20	43, 110, 111, 112, 115, 117, 119, 120, 121, 123, 136, 138, 140, 141, 142, 143, 147, 148, 149, 152	4	43, 112, 136, 142	4	20%
	FNS-FNS	10	34, 35, 37, 39, 172, 174, 175, 193, 194, 195	1	174	1	10%
	Total:	**57**	**-**	**11**	**-**	**13**	**Average 19.30%**
Both groups	**Grand total:**	**165**	**-**	**86**	**-**	**101**	**-**

Utterances that contained CRs were defined and categorised into different linguistic types. Based on Schegloff (1977), Purver (2004), Cho (2007) and Gísladóttir (2015), five CR categories were developed here. However, a new category of CR was found in this study – the non-verbal CR which was added to the five previously known. This new category is similar to the one found in the Argentine Sign Language (LSA) and described as an implicit type of repair initiator "freeze-look", which is a response performed by an addressee by a non-manual action (eyebrow, forehead, eye gaze, nose, mouth, tongue and cheek) after a question has been asked (Manrique and Enfield, 2015; Manrique, 2016). Since the video corpus is very large (165 recordings in a .mov and .eaf (ELAN) format), and since the transcription of all CR utterances is listed in a separate Excel, only an overview of the discourse analysis of CR is presented here. According to the transcription of dialogues, specifically of CR utterances in the videos, different discourse types of CRs were characterised (Figure 12). The sequential number of each recording was assigned to each CR type that contained it. The frequency of their occurrence can be found in Table 9.

Table 9: CR types in the *Virtual Reykjavik* corpus.

Speaker pair	**No. of collected videos**	**No. of videos containing CRs**	**Frequency no. o. of**	**Video sequence no.**	**CR**	**CR Type**
MNS-*MNNS*	22	12	11	95, 128, 148, 152, 153, 154, 156, 158, 159, 160, 161	Hitt húsið (Hitt húsið)	Ellipsis (Restricted)
			1	150	Hvað, Hitt húsið (What, Hitt húsið)	Partial Explicit Query (Open)
			1	160	Þú meinar Kiki (You mean Kiki)	Offering a Candidate (Restricted)
MNS-*FNNS*	32	23	24	48, 50, 57a, 58, 64, 69, 70, 73, 79, 80a, 90, 93, 94, 99, 100, 106, 109	Hitt húsið (Hitt húsið) / Hvar (Where)/ Hvaða hús (Which house)	Ellipsis (Restricted)
			3	80, 89	Hvað hitt húsið (What, Hitt húsið)/ Hvað segirðu (What are you saying) / Hvaða hús (Which house)	Full / partial explicit queries (Open)
			2	61, 80	Postulinn þú meinar (The apostle, you mean) / Hitt húsið, það er auglýsingastofa, er það ekki (Hitt húsið, that is an advertising agency, isn´t it)	Offering a Candidate (Restricted)
			1	49	Hvar er Hitt húsið (Where is Hitt húsið)	Repetition (Restricted)

			2	96	(no speech, interplay of gaze, facial expressions, handedness, body posture)	Non-verbal (Open)
FNS-MNNS	23	18	17	125, 126, 127, 129, 153, 155, 156, 158, 159, 160, 161, 165, 166, 167, 171	Hitt húsið (Hitt húsið)/ Hvar (Where)	Ellipsis (Restricted)
			2	132, 163	Hvað segirðu (What are you saying)/ Veit hvað (Know what)	Partial query (Open)
			1	168	Ha (Huh)	Fragment/ Interjection Strategy (Open)
FNS-FNNS	31	22	18	51, 52, 54, 57, 67, 68, 72, 74, 75, 84a, 86, 87, 91, 98, 101, 103, 104	Hitt húsið (Hitt húsið)	Ellipsis (Restricted)
			2	84, 85	Ha (Huh)	Fragment (Open)
			1	63	Fyrirgefðu, hvar er (Excuse me, where is)	Partial query (Open)
			1	66	Hitt húsið er (Hitt húsið is)	Repetition (Restricted)
			1	95	Við Hringbraut (By Hringbraut)	Offering a Candidate
MNS-MNS	17	4	4	46, 122, 135, 151	Hitt húsið (Hitt húsið)	Ellipsis (Restricted)
			1	135	Svona fjölþjóða eitthvað (Something like the multinational)	Offering a Candidate (Restricted)
MNS-FNS	10	2	2	173, 198	Hitt húsið (Hitt húsið)	Ellipsis (Restricted)

			1	173	Ha (Huh)	Fragment / Interjection Strategy (Open)
FNS-MNS	20	4	4	43, 112, 136, 142	Hitt húsið (Hitt húsið)	Ellipsis (Restricted)
FNS-FNS	10	1	1	174	Hitt húsið, er það ekki fyrir ofan pósthúsið? Veistu það? (Hitt húsið, isn't it above the Post Office? Do you know?)	Full Explicit Query (Open)
All speaker pairs:	**165**	**86**	**101**			

3.3.2.6 Ethical Issues. There are several ethical issues that pertain to this study that affected the data source:

a. Video recordings of participants (speakers A), who seem to be under the influence of alcohol or drugs, were deleted.

b. Participants may be approached only in a public place. This means that is not possible to approach them in any other institution and its premises (e.g. schools, offices, shopping malls, cafés or restaurants, etc.) without asking a special permission from the management of such premises.

c. Oral consent from the participants was collected after the recordings. This approach was not entirely ethical; however, it was decided to proceed in this way in order to get authentic data. During a pilot data collection, the consent was always asked before the recording took place, but this approach influenced the way participants behaved. Participants tended to be more formal and kinder, which had an effect on how they used their hands, eyebrows, eyes, mouth, body posture, etc. During the actual data collection, once the participants gave permission to us to use their recording for the purpose of this research, it was possible to withdraw only by contacting the researcher *via* the University of Iceland.

d. The participants were given a guarantee of confidentiality; i.e. no voice, picture or video of their face or body will be published or distributed anywhere. That is why drawings have been applied instead of photographs of real speakers describing the participant's non-verbal behaviour.

3.3.3 Results from the Study on Multimodal Clarification Requests

This section describes the results of the study on multimodal CRs. It is divided into three parts. Part one presents the linguistic analysis of CR types. The second part presents results of multimodal analysis of each of the six linguistic types of CRs, and the third part presents the multimodal CR models based on the results from the previous analysis. The summary of results is presented at the end of this subchapter.

3.3.3.1 Linguistic CR Types. According to the video corpus that provides the data source for this study, native speakers of Icelandic (men and women, aged 18-70) produce six types of CRs in face-to-face interactions with other non-native and native speakers. In this case, initiators were non-intimate actor volunteers (men and women, aged 20-40) in first-time encounters asking for direction. These CR types, which will be discussed below, are ordered according to the frequency of occurrence: Ellipsis, Full or Partial Explicit Query type, Offering a Candidate, Fragment/Interjection Strategy, Repetition, and the Non-Verbal 'Freeze Look' CR. The Ellipsis was the most common CR occurring in 79.21% of the collected corpus, whereas the other types only under 10%. No difference was found in production of this particular type by native speakers in the following speaker pairs: native-native and native-non-native. However, the only difference was in the frequency of CR occurrences, i.e. more than six times as many CR occurrences were found in native-non-native speaker pairs (88 instances) than in native-native speaker pairs (13 instances). Gender did not seem to affect the types of responses given, as no preference was detected by neither men nor women in regard to the choice of linguistic CR types. These seemed to vary randomly. However, no comparison can be done with the last CR type, the non-verbal CR, because it occurred only twice within the same speaker pair (male native – female non-native). One cannot say that a particular actor's behaviour caused a given CR type. From notes of observations, different types of CR were used in order to request a clarification from that part of information which was not understood. For instance, when the actor asked: "Excuse me, do you know where Hitt húsið is?", then the native speaker would answer with a CR Ellipsis ("Hitt húsið?"), or with the Full or Partial Explicit Query(Do I know where what is?"), or another type of CR, depending on the kind of information they needed the actor to clarify for them. For a better overview, the linguistic types and their frequency is presented in Table 10.

Table 10: Types of CRs and their frequency in the *Virtual Reykjavik* corpus.

No.	Type		No. of instances	Proportion	Selected example video sequence no.
1.	Ellipsis (Restricted)	Explicit	80	79.21%	#187
2.	Full or Partial Explicit Queries (Open)	Explicit	8	7.92%	#89
3.	Offering a Candidate (Restricted)	Explicit	5	4.95%	#61
4.	Fragment / Interjection Strategy (Open)	Explicit	4	3.96%	#85
5.	Repetitions (Restricted)	Explicit	2	1.98%	#49
6.	Non-Verbal ‘Freeze Look’ (Open)	Implicit	2	1.98%	#96
	Total		**101**	**100%**	**-**

The conclusion therefore is that speakers chose to use different types of CRs, proportionally 79.21% Ellipsis, 7.92% full or Partial Explicit Query, 4.95% Offering a Candidate, 3.96% Fragment/Interjection Strategy, 1.98% Repetitions, and 1.98% non-verbal, according to what information they needed to clarify, due to mishearing or possibly misinterpreting or misunderstanding the actor’s question, or some parts of the question. Below the different multimodal CR types identified in this study are presented.

3.3.3.2 Multimodal Description of CR Types. As mentioned above, six linguistic CR types were identified in the corpus: Ellipsis, Full or Partial Explicit Query, Offering a Candidate, Fragment/Interjection Strategy, Repetition, Non-verbal (see Table 10). Their occurrence in frequency, as well as their multimodal realisation, is described in an overview table in each subsequent section. The multimodal analysis revealed no difference in the multimodal realisation of different CR types between native-native and native-non-native speaker pairs. The group of native-native speakers, the control group, whereas the group of native-non-native speakers represented the focus group. It must be mentioned that most of the CR types did not have sufficient examples in the control group. It was only the CR Ellipsis which had examples of several instances between speaker pairs in the control group for comparison purposes. The other five had only one or two examples. The CR type "freeze-look" had no examples in the control group. The focus in the research was on collecting data from various speaker pairs in both groups of speakers, however, it turned out that in the video recordings, the native speakers in the control group did not really use CRs. For this reason, there were no CRs registered in some of the speaker pairs in the control group. Based on this finding, it can be concluded that conversations between native speaker pairs did not require many CRs, most probably due to the fact that the speech was clear and there was no significant disruption coming from the surroundings. Moreover, no significant difference in CR realisation was found between men and women. The multimodal data shows that both genders use CR strategies in similar ways. For this reason, the present research concludes that CR strategies are multimodally produced similarly between genders. One representative example of each linguistic CR type has been chosen from the corpus following these criteria: optimal video quality, optimal position of speaker A and the actor in the video, clearly visible multimodal features on camera allowing for their clear description (or because there was only one example available). Each subsequent section is dedicated to one CR category. The CR categories are presented in order according to frequency of occurrence in the data from the most frequent to the least frequent.

3.3.3.1.1 Ellipsis (Restricted). Extract 1 is an example of the Ellipsis (restricted type) of CR which is demonstrated in a transcription of a dialogue between speaker A (MNS) and the actor (MNNS).

Extract 1: Video 187_B_MNS_MNNS_MVI_0175

1 Actor: Afsakið, fyrirgefðu, vitið þið hvar Hitt húsið er? Question
(Excuse me, do you know where Hitt húsið is?)
2 Speaker A: Hitt húsið? CR
(Hitt húsið?)
3 Actor: Já. Answer
(Yes.)
4 Speaker A: Nei.
(No.)

Ellipsis was used in 79.21% (80/101 instances) of all the cases in the corpus, which makes the most frequently used CR in the collected corpus. It was found in situations in which

Speaker A was trying to achieve grounding and asked the actor a question in order to clarify what has been said (by the actor). This CR was done by omitting some words in the question and thus producing only the 'keywords' that helped him/her to clarify the problem. This type was produced with four types of intonation. The body language in this type is summarised in Table 11 (focus group) and Table 12 (control group) below. The speakers use various types of intonation in the following occasions, but it is not clear what may be affecting the intonation. From research notes of data observation and analysis, it could be suggested that the intonation may vary according to the emphasis of what the native speaker wants to clarify more, or whether the CR Ellipsis has more than one syllable, e.g. the word "Hitt húsið" with rising-falling intonation _/_ (Hitt hú|sið) or rising-falling-rising _/\/ (Hitt hú|sið):

- falling in 58 instances,
- rising-falling in 16 instances,
- neutral in 4 instances,
- rising-falling-rising in 2 instances.

By analysing the body behaviour according to each intonation, no specific pattern was detected. Native speakers, whether men or women, used their body behaviour in a similar manner, depending, however, on the situation, i.e. the actor speaking quietly, or strong background noise from passing vehicles, or work on a construction site, for example. Then the subjects were more likely to slightly frown their forehead or draw their eyebrows slightly together. By analysing the multimodal behaviour for this CR type, it shows that the head is directed at the actor (the one asking a question), forehead is either neutral or slightly frown, eyebrows are either neutral or slightly drawn together, speaker A is directly looking at the actor (direct gaze), the mouth remains slightly open, hands and fingers are not involved in speaking, the body posture and legs are not moving, but the torso only in very few examples leans slightly forward towards the actor as if speaker A wanted to come closer. The body behaviour in the control group is very similar, however with fewer instances of CRs.

Out of 165 video recordings listed in the video corpus, the Ellipsis occurred 70 times in the focus group speaker pairs and 10 times in the control group speaker pairs.

Focus group:

- 11 instances in MNS-MNNS
- 25 instances in MNS-FNNS

- 17 instances in FNS-MNNS
- 17 instances in FNS-FNNS.

Control group:

- 2 instances in MNS-MNS
- 3 instances in MNS-FNS
- 5 instances in FNS-MNS
- 0 instances in FNS-FNS.

A more detailed overview of the multimodal realisation of the Ellipsis CR in both speaker pair groups is in Table 11 and Table 12.

Table 11: Multimodal annotation grid for CR Ellipsis in the focus group.

Focus Group				
Multimodal Annotation Grid for Ellipsis CR (total average duration 0.769 sec.)				
Tier	**MNS-MNNS (11 instances)**	**MNS-FNNS (25 instances)**	**FNS-MNNS (17 instances)**	**FNS-FNNS (17 instances)**
Mean duration	0.713 sec	0.716 sec	0.804 sec	0.843 sec
Icelandic	hitt húsið	hitt/hvaða hús	hitt húsið	hitt húsið
CR	hitt húsið (hitt húsið)	hitt húsið húsið (hitt húsið) /hvaða hús (which house)	hitt húsið (hitt húsið)	hitt húsið (hitt húsið)
Intonation	falling/falling-rising/rising-falling-rising	falling/falling-rising/neutral	falling/falling-rising/neutral	falling/falling-rising/neutral/neutral
Head	directed at actor, a slight toss up	directed at actor, a slight toss up	directed at actor	directed at actor
Forehead	neutral/slightly frown	neutral/slightly frown	neutral/slightly frown	neutral/slightly frown
Eyebrows	neutral/slightly drawn together	neutral/slightly drown together	neutral/slightly drawn together	neutral/slightly drawn together
Eyes	direct gaze	direct gaze	direct gaze	direct gaze
Mouth	left slightly open	left slightly open	left slightly open/ left slightly open with a smile	left slightly open

Handedness	fixed, no movement/arms crossed	fixed, no movement/beside the body	fixed, no movement	fixed, no movement/beside the body
Fingers	neutral/N/A	neutral/N/A	neutral/N/A	neutral/N/A
Body posture	directed at actor/fixed, no movement	directed at actor/fixed, no movement	directed at actor/fixed, no movement	directed at actor/fixed, no movement
Torso	fixed, no movement/slightly leaning forward towards actor	fixed, no movement, directed at actor	fixed, no movement, directed at actor	fixed, no movement/slightly leaning forward towards actor
Legs	fixed, no movement / N/A	fixed, no movement/ N/A	fixed, no movement/ N/A	fixed, no movement / N/A
Distance	close to actor (ca. 1 m)	close to actor (ca. 1 m)	close to actor (ca. 1 m)	close to actor (ca. 1 m)

Table 12: Multimodal annotation grid for CR Ellipsis in the control group.

Control Group				
Multimodal Annotation Grid for Ellipsis CR (total mean duration 1.617 sec.)				
Tier	**MNS-MNS (4 instances)**	**MNS-FNS (3 instances)**	**FNS-MNS (5 instances)**	**FNS-FNS (0 instances)**
Mean duration	1.022 sec	1.475 sec	0.944 sec	N/A
Icelandic	hitt húsið	hitt húsið	hitt húsið	N/A
CR	hitt húsið (hitt húsið)	hitt húsið (hitt húsið)	hitt húsið (hitt húsið)	N/A
Intonation	falling/falling-rising	falling/rising-falling-rising	falling	N/A
Head	directed at actor, slight toss up, looking direction towards hitt húsið	looking away, slight head tilt, chin slightly up	directed at actor	N/A
Forehead	neutral	neutral	neutral/slightly frown	N/A
Eyebrows	slightly drawn together	neutral	slightly drawn together	N/A

Eyes	direct gaze at actor and then looking sideways	looking sideways	direct gaze at actor	N/A
Mouth	left slightly open	left slightly open/ left closed	left slightly open	N/A
Handedness	fixed, no movement	fixed, no movement	fixed, no movement	N/A
Fingers	N/A	N/A	neutral	N/A
Body posture	fixed, no movement, directed at actor	fixed, no movement, directed at actor	fixed, no movement, directed at actor	N/A
Torso	fixed, no movement, directed at actor	fixed, no movement, directed at actor	fixed, no movement, directed at actor	N/A
Legs	fixed, no movement	N/A	fixed, no movement	N/A
Distance	close to actor (ca.1 m)	close to actor (ca. 1 m)	close to actor (ca. 1 m)	N/A

3.3.3.1.2 Full or Partial Explicit QueryCR (Open). Extract 2 is an example of the Full or Partial Explicit Query(open type) of CR that occurred in an interaction between MNS and FNNS.

Extract 2: Video 89_J_MNS_FNNS_MVI_1530

1 Actor: Fyrirgefðu, má ég að spyrja hvar eh Hitt húsið er? Question
(Excuse me, may I ask where eh Hitt húsið is?)
2 Speaker A: Hvað segirðu? CR
(What are you saying?)
3 Actor: Hitt húsið. Repair
(Hitt húsið.)
4 Speaker A: Hitt húsið.
(Hitt húsið.)
5 Actor: Já.
(Yes.)

The Full or Partial Explicit QueryCR type is the second most commonly used CR (Gísladóttir, 2015). However, the number of occurrences is only 7.92% (8/101 instances) in the whole corpus obtained here. The Full or Partial Explicit Querywas found in situations in which speaker A uses full or partially incomplete questions that may or may not include 'wh-' words. Whether or not the native speakers did not understand or did not hear the actor's question, they used this type of CR. This CR was produced with three kinds of intonation. The body language in this type is summarised in Table 13 (focus group) and Table 14 (control group) below. It shows that the speakers randomly use intonation in the following occasions:

- falling 4 instances
- rising-falling 3 instances
- rising 1 instance

By analysing the body behaviour according to each intonation, no specific pattern was detected. Native speakers, whether men or women, used their physical behaviour in a similar manner, however, only 1 to 2 instances for each speaker pair were captured (see Table 13 below). The subjects would keep their head directed at the actor. Their forehead is neutral or slightly frowned. In half of the cases it is neutral, while in the other half it is a slightly frowned forehead. Eyebrows are in most of the cases slightly drawn together, but in one case slightly raised and, in another case, kept neutral. The mouth is kept slightly open. Hands and fingers are not involved in speaking, the body posture and torso are directed at the actor and slightly leaning forward. The legs remain still. The body behaviour in the control group can only be compared to one instance as it was not possible to analyse the video due to a wrong angle of the camera or low quality of the video. The body behaviour in this one instance is very similar to the one in the focus group. Out of 165 video recordings listed in the video corpus, the Full or Partial Explicit QueryCR occurred 6 times in the focus group speaker pairs and 1 time in the control group speaker pairs.

Focus group:

- 1 instance in MNS-MNNS
- 2 instances in MNS-FNNS
- 2 instances in FNS-MNNS
- 1 instance in FNS-FNNS.

Control group:

- 0 instances in MNS-MNS
- 0 instances in MNS-FNS
- 0 instances in FNS-MNS
- 1 instance in FNS-FNS.

For a more detailed overview of the multimodal realization of the full or partial CR in both speaker pair groups, see Table 13 and Table 14.

Table 13: Multimodal annotation grid for CR full/partial query in the focus group.

Focus Group				
Multimodal Annotation Grid for a Full / Partial Query CR (average duration 1,356 sec.)				
Tier	**MNS-MNNS (1 instance)**	**MNS-FNNS (2 instances)**	**FNS-MNNS (2 instances)**	**FNS-FNNS (1 instance)**
Mean duration	0.927 sec	2.081 sec	0.965 sec	1.433 sec
Icelandic	hvað, Hitt húsið	hvað segirðu / Hitt húsið, það er auglýsingastofa, er það ekki	hvað segirðu/veit hvað	fyrirgefðu, hvar er
CR	hvað, Hitt húsið (What, Hitt húsið)	hvað segirðu (What are you saying)/ Hitt húsið, það er auglýsingastofa, er það ekki (Hitt húsið, that is advertisement company, isn´t it)	hvað segirðu (What are you saying)/veit hvað (Know what)	fyrirgefðu, hvar er (Excuse me, where is)
Intonation	rising-falling	rising-falling / falling	falling	rising
Head	directed at actor	directed at actor + a slow nod /directed at actor	directed at actor /directed at actor and getting slightly forward towards the actor	directed at actor, slightly pushed forward, very slightly tilt+turn to the right to put the left ear closer to the actor

Forehead	neutral	neutral	slightly frown	slightly frown
Eyebrows	slightly drawn together	drawn slightly together /neutral	slightly drawn together	slightly raised
Eyes	direct gaze at actor	direct gaze + side-look to the right and back /direct gaze at actor + longer blink	direct gaze at actor	direct gaze at actor
Mouth	left slightly open	left slightly open / left closed	left slightly open	left slightly open
Handedness	no movement, both hands in the pockets of the jacket	no movement, both hands in the pockets of the trousers	no movement	no movement
Fingers	N/A	N/A	N/A	N/A
Body posture	turning towards actor	directed at actor	directed at actor /directed at actor and coming closer	directed at actor
Torso	slightly leaning forward	directed at actor /leaning slightly forward	directed at actor /directed at actor and leaning slightly forward	slightly leaning forward
Legs	stopped walking	no movement	no movement	one step forward to get closer to actor
Distance	close to actor (ca. 1 m)	close to actor (ca. 1 m)	close to actor (ca. 1 m)	close to actor (ca. 1 m)

Table 14: Multimodal annotation grid for CR full/partial query in the control group.

Control Group				
Multimodal Annotation Grid for a Full / Partial Query CR (average duration 1.617 sec)				
Tier	**MNS-MNS (0 instance)**	**MNS-FNS (0 instance)**	**FNS-MNS (0 instance)**	**FNS-FNS (1 instance)**

Average duration	N/A	N/A	N/A	1,617 sec.
Icelandic	N/A	N/A	N/A	Er það ekki fyrir ofan pósthúsið, veistu það?
CR	N/A	N/A	N/A	Er það ekki fyrir ofan pósthúsið, veistu það (Isn't it above the post house, do you know it?)
Intonation	N/A	N/A	N/A	rising-falling
Head	N/A	N/A	N/A	slightly turned away to the from the actor towards the direction of hitt húsið
Forehead	N/A	N/A	N/A	neutral
Eyebrows	N/A	N/A	N/A	slightly drawn together
Eyes	N/A	N/A	N/A	direct gaze at actor
Mouth	N/A	N/A	N/A	left slightly open
Handedness	N/A	N/A	N/A	right hand pointing towards hitt húsið, left hand holding a handle of a pram
Fingers	N/A	N/A	N/A	right palm neutral, right palm holding a handle of a baby carriage
Body posture	N/A	N/A	N/A	directed at actor, no movement
Torso	N/A	N/A	N/A	directed at actor, no movement
Legs	N/A	N/A	N/A	N/A
Distance	N/A	N/A	N/A	close to actor (ca. 1 m)

3.3.3.1.3 Offering a Candidate CR (Restricted). Extract 3 is an example of the Offering a Candidate (restricted type) CR.

Extract 3: Video 61_O_MNS_FNNS_MVI_1480
1 Actor: Góðan daginn.
(Good day.)
2 Speaker A: Góðan daginn.
(Good day.)
3 Actor: Ég, afsakið, eh:::: má ég fá að spyrja hvar er
(Excuse me, eh:::: may I get to ask where)
eh::: pósturinn? Question
(eh::: the post is?)
4 Speaker A: Posturlinn? CR 1
(The post?)
5 Actor: Já, pósturinn. Answer
(Yes, the post.)
6 Speaker A: Postulinn, þú meinar… CR 2
(The apostle, you mean…)
7 Actor: (já, pósturinn, já) Answer
((yes, the post, yes))
8 Speaker A: Pósturinn?
(The post?)
9 Actor: Pósturinn.
(The post.)
10 Speaker A: Pósturinn, það er niðri í bæ.
(The post, that is in downtown.)

Offering a Candidate CR is the third most common CR. However, the number of occurrences is only 4.95% (5/101 instances) in the whole corpus. A participant makes this type of a CR if he/she does not exactly understand what the other participant has just said and offers a candidate, i.e. an alternative word or phrase, by which he/she suggests a different object that may better clarify the previous utterance. In the context of this research, this type of CR was used by speaker A only when they did not understand what kind of place the actor was asking directions to. This is why they offered, or suggested, a different word to try to make clear to speaker A whether that was the word the speaker is referring to. As shown in Table 15, there were also examples of offering more specific or alternative names for another location by speaker A. The body language in this type is summarised in Table 15 (focus group) and Table 16 (control group) below. It shows that the speakers use only one type of intonation, i.e. rising-falling, which was detected in all five instances. By analysing the body behaviour, one specific pattern was detected. Native speakers, whether men or women, used their body behaviour in a similar manner, however, only very few instances for each speaker pair were captured. The subjects keep their head directed at the actor. Their forehead is neutral or slightly frown in two of the cases (compare speaker pairs in Table 15 and Table 16 below). When the forehead is slightly frown, the eyebrows are then slightly raised. Otherwise both foreheads and eyebrows remain neutral. Speaker A is directly looking at the actor and the mouth is kept slightly open. Hands and fingers were used only in one instance when speaker A was pointing at the building when finishing the CR, otherwise they are not involved in speaking. The body posture and torso are directed at the actor, and in one case the torso is slightly leaning forward. The legs are not moving. The body behaviour in the control group can only be compared to one instance as it was not possible to analyse the video due to a wrong angle of camera or low quality of the video. The body behaviour in this one instance is very similar to the one in the focus group, i.e. the one which does not use a hand for pointing towards the place that speaker A was asked for directions to. This occurred 4 times in the focus group speaker pairs and 1 time in the control group speaker pairs.

Focus group:

- 1 instance in MNS-MNNS
- 2 instances in MNS-FNNS
- 0 instances in FNS-MNNS
- 1 instance in FNS-FNNS.

Control group:

- 1 instance in MNS-MNS
- 0 instances in MNS-FNS
- 0 instances in FNS-MNS
- 0 instances in FNS-FNS.

For a more detailed overview of the multimodal realisation of Offering a Candidate CR in both speaker pair groups, see Table 15 and Table 16.

Table 15: Multimodal annotation grid for CR Offering a Candidate in the focus group.

Focus Group				
Multimodal Annotation Grid for an Offering a Candidate CR (total mean duration 1.057 sec)				
Tier	**MNS-MNNS** **(1 instance)**	**MNS-FNNS** **(1 instance)**	**FNS-MNNS** **(0 instance)**	**FNS-FNNS** **(1 instance)**
Mean duration	0.980 sec	1.463 sec	N/A	0.729
Icelandic	ertu að meina kiki	postulinn, þú meinar	N/A	við hringbraut
CR	ertu að meina kiki (do you mean kiki)	postulinn, þú meinar (apostle, you mean)	N/A	við hringbraut (by hringbraut)
Intonation	rising-falling	rising-falling	N/A	rising-falling
Head	directed at actor	turning slightly right and back directed at actor	N/A	directed at actor
Forehead	N/A	slightly frown	N/A	neutral
Eyebrows	N/A	slightly raised	N/A	neutral
Eyes	direct gaze	one longer eye blink, then looking right and back at actor	N/A	direct gaze
Mouth	left slightly open	slight smile, left closed	N/A	left slightly open
Handedness	right hand showing directions to a bar called Kiki	no movement, both hands inside the coat pockets	N/A	no movement, both hands holding a handbag
Fingers	index finger used for pointing	N/A	N/A	palms holding a handbag
Body posture	directed at actor	directed at actor, turning slightly right and back at actor	N/A	directed at actor

Torso	directed at actor	directed at actor, turning slightly right and back at actor	N/A	directed at actor
Legs	N/A	N/A	N/A	no movement
Distance	close to actor (ca. 1 m)	close to actor (ca. 1 m)	N/A	close to actor (ca. 1 m)

Table 16: Multimodal annotation grid for CR Offering a Candidate in the control group.

Control Group				
Multimodal Annotation Grid for an Offering a Candidate CR (average duration 3,002 sec.)				
Tier	**MNS-MNS (1 instance)**	**MNS-FNS (0 instances)**	**FNS-MNS (0 instances)**	**FNS-FNS (0 instances)**
Mean duration	3.002 sec	N/A	N/A	N/A
Icelandic	svona fjölþjóða eitthvað	N/A	N/A	N/A
CR 2	svona fjölþjóða eitthvað (something like the multinational)	N/A	N/A	N/A
Intonation	rising-falling	N/A	N/A	N/A
Head	directed at actor, then turning left towards machine	N/A	N/A	N/A
Forehead	slightly frown	N/A	N/A	N/A
Eyebrows	N/A	N/A	N/A	N/A
Eyes	direct gaze at speaker then turning left towards parking machine	N/A	N/A	N/A
Mouth	left slightly open	N/A	N/A	N/A
Handedness	busy inserting coins into the parking machine	N/A	N/A	N/A

Fingers	busy holding coins	N/A	N/A	N/A
Body posture	directed at the parking machine, turned slightly right towards actor	N/A	N/A	N/A
Torso	directed at the parking machine, turned slightly right towards actor	N/A	N/A	N/A
Legs	N/A	N/A	N/A	N/A
Distance	close to actor, ca. 1.5m	N/A	N/A	N/A

3.3.3.1.4 Fragment / Interjection Strategy CR (Open). Extract 4 is an example of the Fragment or Interjection Strategy (open type) CR that occurred in an interaction between FNS and FNNS.

Extract 4: Video 85_J_FNS_FNNS_MVI_1524
1 Actor: Fyrirgefðu, má ég að spyrja hvar Hitt húsið er? Question
(Excuse me, may I ask where Hitt húsið is?)
2 Speaker A: Ha? CR
(Huh?)
3 Actor: Hvar Hitt húsið er. Repair
(Where Hitt húsið is.)
4 Speaker A: Hitt húsið, það er hérna rau-rauð hús.
(Hitt húsið, that is here re-red house.)
5 Actor: Já.
(Yes.)

The Fragment or Interjection Strategy CR is the fourth most commonly used CR. However, the number of occurrences is only 3.96% (4/101 instances) in the whole corpus. It was found in situations, in which speaker A uses a fragment type of CR or an interjection to ask the actor for a clarification. This type was produced with one kind of intonation, i.e. falling in four instances. The body language in this type is summarised in Table 17 (focus group) and Table 18 (control group). One specific pattern was detected for body behaviour. Native speakers, whether men or women, used their body behaviour in a similar manner, however, only very few instances for some of the speaker pairs were captured, which limits the analysis to this available data. The subjects keep their head directed at the actor. Their forehead is mostly neutral but in one case it is slightly frown. When the forehead is slightly frown, then the eyebrows are slightly raised. Otherwise both the forehead and eyebrows stay neutral. Speaker A is directly looking at the actor. The mouth is kept slightly open. Hands and fingers are not used. The body posture and torso are directed at the actor and in one case the torso is slightly leaning forward. The legs are not moving, except for when in the one case the torso is slightly leaning forward then also the legs move to form one step forward to get closer to the actor. The body behaviour in the control group can only be compared to one instance as it was not possible to analyse other videos due to a wrong angle of camera or low quality of the video. The body behaviour in this one instance is very similar to the one in the focus group, i.e. without any movement of legs or torso leaning forward.

Out of 165 video recordings in this study, it occurred 3 times in the focus group speaker pairs and 1 time in the control group speaker pairs.

Focus group:

- 0 instances in MNS-MNNS
- 0 instances in MNS-FNNS
- 1 instance in FNS-MNNS
- 2 instances in FNS-FNNS.

Control group:

- 0 instances in MNS-MNS
- 1 instance in MNS-FNS
- 0 instances in FNS-MNS
- 0 instances in FNS-FNS.

For an overview of the multimodal realisation of the Fragment and Interjection Strategy CR in both speaker pair groups, see Table 17 and Table 18.

Table 17: Multimodal annotation grid for CR Fragment/Interjection Strategy in the focus group.

Focus Group				
Multimodal Annotation Grid for a Fragment / Interjection Strategy CR (total mean duration 0.440 sec)				
Tier	**MNS-MNNS (0)**	**MNS-FNNS (0)**	**FNS-MNNS (1)**	**FNS-FNNS (2)**
Mean duration	N/A	N/A	0.263 sec	0.528 sec
Icelandic	N/A	N/A	ha	ha
CR	N/A	N/A	ha (huh)	ha (huh)
Intonation	N/A	N/A	falling	falling
Head	N/A	N/A	directed at actor	directed at actor/ directed at actor + slightly pulled forward
Forehead	N/A	N/A	neutral	neutral/slightly frown
Eyebrows	N/A	N/A	neutral	neutral/slightly raised
Eyes	N/A	N/A	direct gaze	direct gaze at actor
Mouth	N/A	N/A	left slightly open, with a both mouth corners slightly up simulating a smile mode	left slightly open
Handedness	N/A	N/A	left hand holding a handle of a baby carriage, right hand taking off earphones	no movement, beside the body
Fingers	N/A	N/A	N/A	N/A
Body posture	N/A	N/A	directed at actor	directed at actor
Torso	N/A	N/A	directed at actor	directed at actor/ directed at actor + leaning slightly forward
Legs	N/A	N/A	N/A	no movement/one

				step towards actor
Distance	N/A	N/A	close to actor (ca. 1 m)	close to actor/ getting closer to actor

Table 18: Multimodal annotation grid for CR Fragment/Interjection Strategy in the control group.

Control Group				
Multimodal Annotation Grid for a Fragment / Interjection Strategy CR (total average duration 0.933 sec)				
Tier	**MNS-MNS (0)**	**MNS-FNS (1)**	**FNS-MNS (0)**	**FNS-FNS (0)**
Mean duration	N/A	0.933 sec	N/A	N/A
Icelandic	N/A	ha	N/A	N/A
CR	N/A	ha (huh)	N/A	N/A
Intonation	N/A	falling	N/A	N/A
Head	N/A	directed at actor	N/A	N/A
Forehead	N/A	neutral	N/A	N/A
Eyebrows	N/A	neutral	N/A	N/A
Eyes	N/A	direct gaze at actor	N/A	N/A
Mouth	N/A	left slightly open	N/A	N/A
Handedness	N/A	no movement, both hands beside the body	N/A	N/A
Fingers	N/A	N/A	N/A	N/A
Body posture	N/A	directed at actor	N/A	N/A
Torso	N/A	directed at actor	N/A	N/A
Legs	N/A	N/A	N/A	N/A
Distance	N/A	close to actor (ca. 1m)	N/A	N/A

3.3.3.1.5 Repetition (Restricted). Extract 5 is an example of the Repetition (restricted type) of CR that occurred in an interaction between MNS and FNNS.

Extract 5: Video 49_J_MNS_FNNS_MVI_1460
1 Actor: Ehm góðan daginn.
(Ehm good day.)
2 Speaker A: Daginn.
(Day.)
3 Actor: Am era að bara að spyrja hvar er Hitt húsið? Question
(Am I am only asking where is Hitt húsið)
4 Speaker A: Hvar er Hitt húsið? CR
(Where is Hitt húsið?)
5 Actor: Já. Repair
(Yes.)
6 Speaker A: Heyrðu, það er bara rauða húsið þarna á horninu.
(Well, that is just the red house over there at the corner.)

The Repetition CR is the fifth most commonly used CR. However, the number of occurrences is only 1.98% (2/101 instances) in the whole corpus. It was found in a situation in which speaker A partially repeats what the actor has just said in order to ask for a clarification. This type was produced with two kinds of intonation, i.e. rising-falling (1 instance) and falling (1 instance). The body language in this type is summarised in Table 19 (focus group). As there were no data available for this type of CR in the control group, there is no data to be compared. By analysing the body behaviour, one specific pattern was detected, with a variation of head movement. Due to the position of the actor, which was in close proximity to the place to which he/she was asking for directions, one speaker turns the head and gaze towards that direction. Otherwise, the body behaviour is very similar, i.e. the subjects keep their head directed at the actor (or turning away), the forehead and eyebrows are neutral, and the mouth is kept slightly open. Hands and fingers are not used. The body posture and torso are directed at the actor, but in one case the torso is also slightly leaning towards the actor. The legs are not moving, except for when speaker A has stopped walking because the actor asked while the speaker was walking towards the actor. Body behaviour cannot be compared to any in the control group because there was no instance available in the control group data corpus.

The repetition CR occurred only 2 times in the focus group speaker pairs and never in the control group speaker pairs. But even this low occurrence in the corpus helps to teach learners unusual or rare responses native speakers use when asking for clarification.

Focus group:

- 0 instances in MNS-MNNS
- 1 instance in MNS-FNNS
- 0 instances in FNS-MNNS
- 1 instance in FNS-FNNS.

Control group:

- 0 instances in MNS-MNS
- 0 instance in MNS-FNS
- 0 instances in FNS-MNS
- 0 instances in FNS-FNS.

A more detailed overview of the multimodal realisation of the Repetition type CR in both speaker pair groups is given in Table 19.

Table 19: Multimodal annotation grid for CR Repetition in the control group.

Focus Group				
Multimodal Annotation Grid for a Repetition CR (total mean duration 1.1 sec)				
Tier	**MNS-MNNS (0)**	**MNS-FNNS (1)**	**FNS-MNNS (0)**	**FNS-FNNS (1)**
Mean duration	N/A	1.1 sec	N/A	0.562 sec
Icelandic	N/A	hvar er Hitt húsið	N/A	Hitt húsið er
CR	N/A	hvar er Hitt húsið (where is Hitt húsið)	N/A	Hitt húsið er (Hitt húsið is)
Intonation	N/A	rising-falling	N/A	falling
Head	N/A	directed at actor + slight turn to right (direction of Hitt húsið) and back at actor	N/A	turning away to right (towards direction of Hitt húsið)
Forehead	N/A	neutral	N/A	N/A
Eyebrows	N/A	neutral	N/A	N/A

Eyes	N/A	direct gaze + looking right and back at actor	N/A	direct gaze at actor then looking to right towards direction of Hitt húsið
Mouth	N/A	from smile to neutral, left slightly open	N/A	left slightly open
Handedness	N/A	no movement, both hands in the pocket of the trousers	N/A	no movement, both hands beside the body
Fingers	N/A	N/A	N/A	N/A
Body posture	N/A	directed at actor	N/A	approaching actor
Torso	N/A	directed at actor + slight movement to the right and back at actor	N/A	slightly leaning forward
Legs	N/A	no movement	N/A	stopped walking
Distance	N/A	close to actor (ca. 1 m)	N/A	close to actor (ca. 1 m)

3.3.3.1.6 Non-Verbal CR (Implicit/Open). Extract 6 is an example of the Non-Verbal (implicit or open type) of CR. It shows how speaker A, a male native speaker of Icelandic, used non-verbal means to ask the female actor, a non-native speaker of Icelandic, for a clarification. Its multimodal realisation is described Table 20 and Table 21.

Extract 6: Video 96_ O_MNS_FNNS_MVI_1540.MOV

1 Actor: Góðan daginn, afsakið.
(Good day, excuse me.)

2 Speaker A: Daginn.
(Day.)

3 Actor: (undetected speech) má ég spyrja hvar er
h'lem,mur? Question
(may I ask where is h'lem,mur?)

4 Speaker A: (no speech) CR(1)

5 Actor: H’lem,mur. Repair(1)
(H’lem,mur.)

6 Speaker A: (no speech) CR(2)

7 Actor: H’lem,mur, h’lem,mur, e:h strætó. Repair(2)
(H’lem,mur, h’lem,mur, e:h bus)

8 Speaker A: Stræt –ah, ‘hlem,mur, ‘hlem,mur, já
‘hlem,mur, ‘hlem,mur er langt hérna.
(Bus –ah, ‘hlem,ur, ‘hlem,mur, yes ‘hlem,mur,
‘hlem,mur is far from here.)

9 Actor: (undetected speech)

10 Speaker A: Já, en það eru strætó bara þarna.
(Yes, but there are some buses over there.)

In the extract above, the actor approaches speaker A, explicitly announces her presence by saying *góðan daginn* (good day) and asks for directions. In this conversation, speaker A executes two non-verbal CRs. The sequence consists of five turns: Question, CR (1), Repair (1), CR (2), and Repair (2). The turn sequence is divided into two parts and the multimodal behaviour associated with each of the two CRs is explained in the examples here below.

Example 1

3 Actor: (undetected speech) má ég spyrja hvar er h’lem,mur? Question
(may I ask where is h’lem,mur?)

4 Speaker A: (no speech) CR(1)

5 Actor: H’lem,mur. Repair(1)
(H’lem,mur.)

The second example of a non-verbal clarification request CR2 shows the multimodal behaviour of speaker A being carried out in the same way as in CR(1), except for one difference in the opening of the mouth: the mouth being closed (lips together in sequence 1) for about 0.320 sec, but then in the second part of the sequence 2 the mouth of the speaker slightly opens up (lower jaw sinks slightly down) and lasts until the end of the turn for about 0.768 sec. The whole CR(2) lasts about 1.088 sec. The turn sequence is as follows: Repair(1), CR(2), and Repair(2).

Example 2

5 Actor: H'lem,mur. Repair(1)
(H'lem,mur.)
6 Speaker A: (no speech) CR(2)
7 Actor: H'lem,mur, h'lem,mur, e::h strætó. Repair(2)
(H'lem,mur, h'lem,mur, e:: bus.)
8 Speaker A: Stræt –ah, 'hlem,mur, 'hlem,mur, já
(Bus –ah, 'hlem,mur, 'hlem,mur, yes)
'hlem,mur, 'hlem,mur er langt hérna.
('hlem,mur, 'hlem,mur is far from here)

The non-verbal CR was found only in one situation and was produced twice by the same speaker when he did not understand what the actor was saying. Proportionally, however, this utterance is only 1.98% (2/101 instances) in the whole corpus. Instead of the native speaker asking the actor for a clarification the usual way, i.e. uttering words, he kept a direct gaze, was mute and used his body cues to signal the actor to clarify herself on the previous utterance. This type was produced without any speech. As there were only two instances of this type produced by the same native speaker interacting with the same actor, the body language for each instance summarised in Table 20 and Table 21. As there were no data available for this type of CR in the control group, there is no data to be compared with. By analysing the body behaviour, one specific pattern in both instances was detected. The head of speaker A is directed at the actor and is leans down. This slight learning can be caused by the actor being of a lower height than speaker A. Gaze is also directed at the actor. The forehead is slightly frown. The eyebrows are slightly drawn together. The mouth is left slightly open at the end of the utterance. There is no movement in hands, fingers, body posture, torso and legs. Except for the torso part is slightly leaning forward, which may have been caused by the height difference between the participants. The body behaviour cannot be compared to any in the control group because there was no instance available in the control group data corpus.

The CR(1) turn sequence shows three different communicative functions, a question, a clarification request CR(1), and a repair Repair(1). In this sequence, asking a question is the communicative function executed by the actor. During this action, speaker A has his mouth slightly open and is looking directly at the actor and is putting his hands inside the pockets of the trousers. Speaker's A body posture, torso, head and gaze had shifted towards the actor when she explicitly announced her presence by saying *góðan daginn* (Engl. good day). Speaker A acknowledged it by responding *daginn* (a short version of good day). Next, the actor approached speaker A, the body posture, torso, and legs of speaker A are unchanged, and he is looking at the actor and listening to her. The actor asks for directions. Speaker A responds by executing a non-verbal CR(1) (Table 20). As a result, the actor executed the Repair1 function and responded by repeating the key word *Hlemmur* that had been previously said but was not understood by speaker A. One possible explanation for speaker A's not understanding the actor may lie in her imperfect pronunciation, because the letter <h> is pronounced silently and the stress is put on the second letter <l> in the word Hlemmur (h'lem,ur). The correct pronunciation is ('hlem,mur).

Table 20: Multimodal annotation grid for CR (1) Non-verbal "Freeze Look".

Tier	**CR(1) (duration 0.257 sec)** Speaker A
Icelandic	-
CR	(no speech)
Intonation	-
Head	directed at the actor, bowed slightly forward
Forehead	slightly frown
Eyebrows	slightly drawn together
Eyes	direct gaze at actor
Mouth	left slightly open
Handedness	no movement, both inside the pockets of trousers
Fingers	N/A
Body posture	no movement, directed at the actor

Torso	no movement, directed at the actor, bowed slightly forward
Legs	no movement, slightly apart
Distance	close to the actor (1m)

The multimodal realisation of CR2 is described in Table 21.

Table 21: Multimodal annotation grid for CR(2) Non-verbal "Freeze Look".

Tier	**CR2 (duration 1.088 sec)** Speaker A
Icelandic	-
CR	(no speech)
Intonation	-
Head	directed at the actor, bowed slightly forward
Forehead	slightly frown
Eyebrows	slightly drawn together
Eyes	direct gaze at actor
Mouth	closed for 0.320 sec and opened again for 0.768 sec
Handedness	fixed, no movement, both inside the pockets of trousers
Fingers	N/A
Body posture	no movement, directed at the actor
Torso	no movement, directed at the actor, bowed slightly forward
Legs	no movement, slightly apart
Distance	close to the actor (1m)

The non-verbal CR occurred only twice in an MNS-FNNS speaker pair, and both times the same speaker A produced it (Table 22).

Focus group:

- 0 instances in MNS-MNNS
- 2 instances in MNS-FNNS
- 0 instances in FNS-MNNS
- 0 instances in FNS-FNNS.

Control group:

- 0 instances in MNS-MNS
- 0 instances in MNS-FNS
- 0 instances in FNS-MNS
- 0 instances in FNS-FNS.

Table 22: Multimodal annotation grid for CR Non-verbal "Freeze Look" in the control group.

Focus Group				
Multimodal Annotation Grid for a Non-Verbal CR (total mean duration 0.673 sec)				
Tier	**MNS-MNNS (0)**	**MNS-FNNS (2)**	**FNS-MNNS (0)**	**FNS-FNNS (0)**
Mean duration	N/A	0.673 sec	N/A	N/A
Icelandic	N/A	(no speech)	N/A	N/A
CR	N/A	(no speech)	N/A	N/A
Intonation	N/A	(no sound)	N/A	N/A
Head	N/A	directed at the actor, bowed slightly forward	N/A	N/A
Forehead	N/A	slightly frown	N/A	N/A
Eyebrows	N/A	slightly drawn together	N/A	N/A
Eyes	N/A	direct gaze at actor	N/A	N/A
Mouth	N/A	slightly open / first closed then opened and left open	N/A	N/A
Handedness	N/A	no movement, both inside the pockets of trousers	N/A	N/A
Fingers	N/A	N/A	N/A	N/A

Body posture	N/A	fixed, no movement, directed at the actor	N/A	N/A
Torso	N/A	no movement, directed at the actor, bowed slightly forward	N/A	N/A
Legs	N/A	no movement, slightly apart	N/A	N/A
Distance	N/A	close to the actor (ca. 1 m)	N/A	N/A

3.4 The Multimodal CR Models

Based on the findings from the study above, six types of CRs were identified in the conversational scenario first encounters asking for directions: Ellipsis, Full or Partial Explicit Query, Offering a Candidate, Fragment/Interjection Strategy, Repetition, and the Non-Verbal CR. The model suggests that each CR type should consist of three stages, the initial stage (T-1) which represents the time when the ECA is listening to the learner´s input, the final stage (T0) which is one that ECA executes its turn in, and the post final stage (T+1) which after the ECA executes the CR. Therefore, each CR type consists of three sequences. For this reason, the proposed model is also divided into three sequences. Each sequence is characterised by a specific multimodal behaviour. These models should give ECAs the ability to interact with learners in an authentic fashion. The models suggest an average time duration, linguistic means, the particular multimodal behaviour associated with each turn, and the proximate distance between the ECA and the learner’s avatar based on interpretation of real-life data. The multimodal features presented in each of the sequences (turns) in the CR models were chosen in accordance with the *Virtual Reykjavik* annotation scheme (Appendix B).

3.4.1 Ellipsis

The ECA can use Ellipsis to ask the learner for a clarification by eliminating all the words in the previously said utterance from the learner and repeating only one key word, or key word sequence, found in that utterance. For instance, the learner asks: How do I get to the Hitt húsið? And the ECA repeats only “Hitt húsið?”. This is shown in the turn sequence order, where the first turn is the learner’s question marked as the CR sequence (T-1); the second turn is the Ellipsis CR said by the ECA which is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). This sequence together with the multimodal behaviour associated with each turn is described in Table 23.

Table 23: The Multimodal CR Ellipsis Sequence Model.

ECA’s Ellipsis CR Sequence					
Sequence no.			**T-1**	**T0**	**T+1**
Function			Question	CR	Answer
Triggering			The user asks for directions to a particular place	The ECA understands or does not understand the question and needs to reconfirm it by mentioning some parts of asking a question for directions	The user acknowledges it by giving a positive/negative answer, or a repair
Duration			Time during which the user asks the ECA a question	0.6 sec	Time during which the user answers the ECA
BML	Verbal behaviour	Linguistic means	-	e.g. place name, “Hitt húsið?”, “Where?”, “Which place?”	-
		Intonation	-	rising-falling	-
	Non-verbal behaviour	Head	directed at user	directed at user	directed at user
		Forehead	neutral	slight frown	neutral
		Eyebrows	neutral	slightly drawn together	neutral
		Eyes	direct gaze at user	direct gaze at user	direct gaze at user
		Mouth	closed	left slightly open	slightly open
		Handedness	no movement	no movement	no movement
		Fingers	N/A	N/A	N/A
		Body posture	adjusting towards user	directed at user	directed at user

		Torso	adjusting towards user	slightly leaning forward towards the user	going back to its initial position
		Legs	adjusting with body position	no movement, slightly apart	no movement
	Distance		within detection radius	close to user (ca. 1 m)	close to user (ca. 1 m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-
	Type		-	clarification-request	-

Figure 13 shows how the multimodal behaviour appears on a native speaker performing the Ellipsis CR; it shows all three sequences, i.e. how the speaker looks before (T-1), while (T0), and after T(+1) performing this CR type. The symbol on the left signifies the position of the actor. All of the three sequences need to be incorporated into ECAs' conversational behaviour in order to achieve a more realistic interaction with human users. Specifications for the multimodal behaviour involved in each sequence are in Table 23.

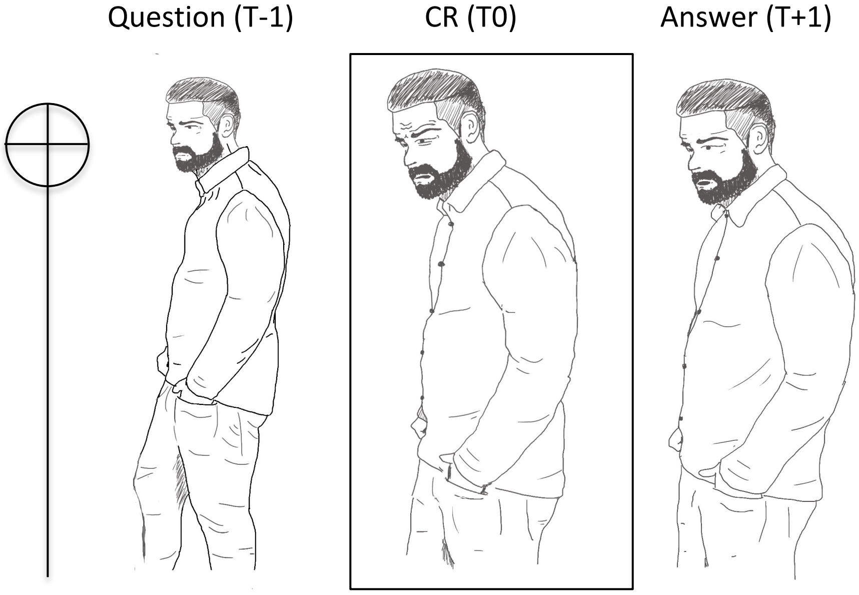


Figure 13: CR Ellipsis sequence. The symbol on the left signifies the position of the actor.

The CR Ellipsis (e.g. Hitt húsið) part is produced in the T0 sequence. In order for the ECA produce an Ellipsis in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The agent's head is directed at the user's avatar. The forehead of the agent is slightly frown and the eyebrows are slightly drawn together. The agent looks directly at the user with the mouth is slightly open after saying the clarification word/phrase with a rising-falling intonation. The hands are beside the body and there is no movement of hands nor fingers. The body posture of the agent is directed at the user. The agent is slightly leaning forward toward the user. The legs are slightly apart but not moving, and the agent is in close proximity to the user, which is about 1 m or less in real life. The model is based on the example above because its realisation is very similar to most of the instances found in the Ellipsis category. Moreover, Figure 13 above is drawn based on an optimal video example in the collected corpus.

3.4.2 Full or Partial Explicit Query

The ECA can use a Full or Partial Explicit Queryto request clarification from the learner by asking a full or partial question about the previously said utterance. For instance, the learner asks: How do I get to the Hitt húsið? And the ECA responds with either a full question, e.g. Hvað segirðu? (What are you saying?), or a Partial Explicit Query, e.g. Hvað, Hitt húsið? (What, the Hitt húsið)? This is shown in the turn sequence order, where the first turn is the learner's question marked as the CR sequence (T-1); the second turn is the Full or Partial Explicit Queryand is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). This sequence together with the multimodal behaviour associated with each turn is described in Table 24.

Table 24: The Multimodal CR Full or Partial Explicit QuerySequence Model.

ECA's Full or Partial Explicit QueryCR Sequence			
Sequence no.	**T-1**	**T0**	**T+1**
Function	Question	CR	Answer
Triggering	The user asks for directions to a particular place	The ECA understands or does not understand the question and needs to reconfirm it by asking a full or partial query	The user acknowledges it by executing a repair

Duration			Time during which the user asks the ECA a question	0.9 sec	Time during which the user answers the ECA
BML	Verbal behaviour	Linguistic means	-	Excuse me, where is? What are you saying? What, the “place name”?	-
		Intonation	-	rising-falling	-
	Non-verbal behaviour	Head	directed at user	directed at user	directed at user
		Forehead	neutral	neutral	neutral
		Eyebrows	neutral	neutral	neutral
		Eyes	direct gaze	direct gaze, 1x longer blink	direct gaze
		Mouth	closed	left closed	open
		Handedness	no movement, hands beside the body (or in trouser pockets)	no movement, both hands in trouser pockets	no movement, both hands in trouser pockets
		Fingers	N/A	N/A	N/A
		Body posture	directed at user	directed at user	directed at user
		Torso	no movement	leaning slightly forward	leaning slightly forward
		Legs	no movement	no movement	no movement
	Distance		within detection radius	close to user (ca. 1 m)	close to actor (ca. 1 m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-
	Type		-	clarification-request	-

Figure 14 shows how the multimodal behaviour looks on a native speaker performing the Full or Partial Explicit QueryCR; it shows all three sequences, i.e. how the speaker looks before (T-1), while (T0), and after T(+1) performing this CR type. The symbol on the right signifies the position of the actor. All three sequences need to be incorporated into an ECA's conversational behaviour in order to achieve a more realistic target-like interaction. Specifications for the multimodal behaviour involved in each sequence are in Table 24.

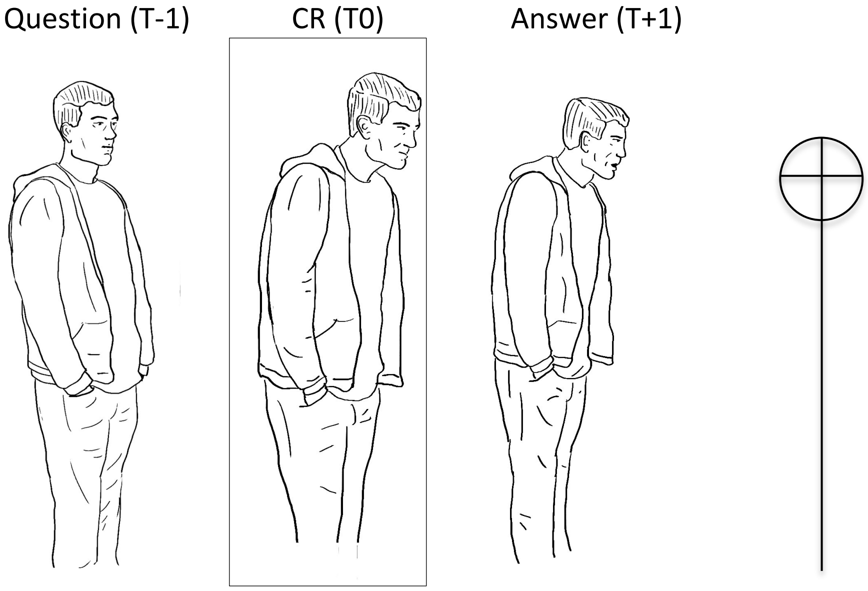


Figure 14: CR Partial Explicit Query sequence. The symbol on the right signifies the position of the actor.

The CR Full or Partial Explicit Query(e.g. Full Explicit Query Hvað segirðu? (What are you saying?), or a Partial Explicit Query, e.g. Hvað, Hitt húsið? (What, the Hitt húsið?)) is produced in the T0 sequence. In order for the ECA to do it in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The head is directed at the user (user's avatar) and the forehead together with eyebrows are neutral. The agent is directly looking at the user and blinks once. The mouth remains closed after saying the clarification word/phrase with a rising-falling intonation. There is no movement of hands nor fingers involved. Body posture is directed at the user. The agent is slightly leaning forward toward the user. There is no movement of legs and the agent is in a close proximity to the user, which is about 1 m or less in real life. The model is based on the example above because its realisation is very similar to most of the instances found in the

Full or Partial Explicit Querycategory. Moreover, the picture in Figure 14 above is drawn based on an optimal video example in the collected corpus.

3.4.3 Offering a Candidate

The ECA can use ‘Offering a Candidate’ to ask the learner for a clarification about the previously said utterance. For instance, the learner asks: How do I get to the Hitt húsið? and the ECA offers an alternative question, or an alternative word/phrase, i.e. offers a candidate. The words or phrases that are alternative to the original question, e.g. the ECA offers a different place name by saying “You mean [the other place name]”, or only saying [the other place name]. This is consequently either confirmed or rejected by the learner usually by “yes” or “no”, or by the learner repeating the question or word/phrase. This is shown in the turn sequence order, where the first turn is the learner’s question marked as the CR sequence (T-1); the second turn is ‘Offering a Candidate’ and is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). This sequence, together with the multimodal behaviour associated with each turn, is described in Table 25.

Table 25: The Multimodal CR Offering a Candidate Sequence Model.

<table>
<tr><th colspan="6">ECA’s Offering a Candidate CR Sequence</th></tr>
<tr><td colspan="3">Sequence no.</td><td><b>T-1</b></td><td><b>T0</b></td><td><b>T+1</b></td></tr>
<tr><td colspan="3">Function</td><td>Question</td><td>CR</td><td>Answer</td></tr>
<tr><td colspan="3">Triggering</td><td>The user asks for directions to a particular place</td><td>The ECA understands or does not understand the question but needs to reconfirm it offering the user an alternative word</td><td>The user acknowledges it by executing a repair</td></tr>
<tr><td colspan="3">Duration</td><td>Time during which the user asks the ECA a question</td><td>1.5 sec</td><td>Time during which the user answers the ECA</td></tr>
<tr><td rowspan="2">BML</td><td rowspan="2">Verbal behaviour</td><td>Linguistic means</td><td>-</td><td>“You mean ‘the other name’?”,
“’The other name’?”</td><td>-</td></tr>
<tr><td>Intonation</td><td>-</td><td>rising-falling</td><td>-</td></tr>
</table>

	Non-verbal behaviour	Head	directed at user	turning slightly right and back	directed at user
		Forehead	neutral	slightly frown	neutral
		Eyebrows	neutral	slightly risen	neutral
		Eyes	direct gaze	one longer eye blink, then looking right and back at user	direct gaze
		Mouth	slightly open	slight smile, left closed	closed
		Handedness	no movement, both hands inside coat pockets	no movement, both hands inside coat pockets	taking both hands out of coat pockets (preparing for showing directions)
		Fingers	N/A	N/A	neutral
		Body posture	directed at user	directed at user, turning slightly right and back at user	directed at user
		Torso	directed at user	directed at user, turning slightly right and back at user	directed at user
		Legs	N/A	N/A	N/A
	Distance		close to user (ca. 1 m)	close to user (ca. 1 m)	close to actor (ca. 1 m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-
	Type		-	clarification-request	-

Figure 15 shows how the multimodal behaviour appears on a native speaker performing Offering a Candidate CR; it shows all three sequences, i.e. how the speaker looks before (T-1), while (T0), and after T(+1) performing this CR type. All of the three sequences need to be incorporated into ECAs' conversational behaviour in order to achieve a more realistic interaction with human users. Specifications for the multimodal behaviour involved in each sequence are in Table 25.

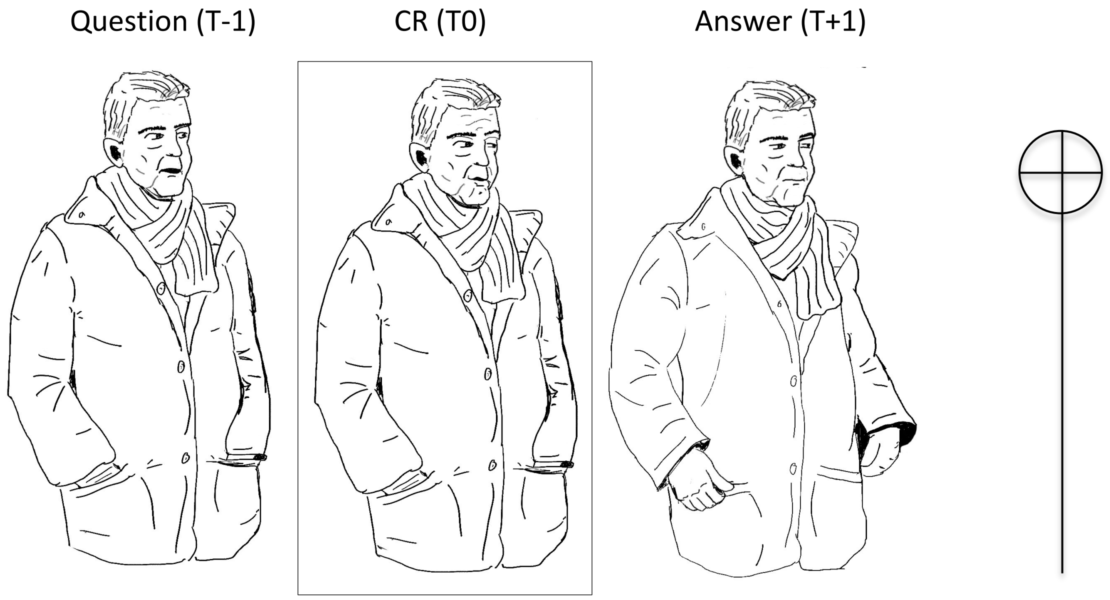


Figure 15: CR Offering a Candidate sequence. The symbol on the right signifies the position of the actor.

The ‘Offering a Candidate’ CR is produced in the T0 sequence. In order for the ECA to do this in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The head is turned slightly right to the opposite direction and then back to the user (user’s avatar). The forehead is slightly frown and the eyebrows are slightly risen. There is one longer eye blink and the agent is looking slightly right and back to the user (synchronised with the head movement). The agent should perform a subtle smile and the mouth is left closed after saying the clarification word/phrase. The hands are beside the body and neither hands nor fingers perform any movement. The body posture is directed at the user (user’s avatar) and the body is turning slightly right and back to the user (synchronised with the head and eye movement). The torso is directed at the user and is turning slightly right and back to the user (synchronised with the body movement). The legs are not moving, and the agent is in close proximity to the user, which is about 1 m or less in real life. The model is based on the example above because its realisation is very similar to most of the instances found in the ‘Offering a Candidate’ category. Moreover, the picture in Figure 15 above is drawn based on an optimal video example in the collected corpus.

3.4.4 Fragment/Interjection Strategy

The ECA can use a 'Fragment or Interjection Strategy' CR to ask the learner for a clarification about the previously said utterance. For instance, the learner asks: How do I get to the Hitt húsið? and the ECA produces a fragment or an interjection, e.g. *Ehm? Huh?* This is shown in the turn sequence order, where the first turn is the learner's question marked as the CR sequence (T-1); the second turn is the fragment or interjection, and is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). This sequence together with the multimodal behaviour associated with each turn is described in Table 26.

Table 26: The Multimodal CR Fragment / Interjection Strategy Sequence Model.

ECA's Fragment/Interjection Strategy CR Sequence					
Sequence no.			**T-1**	**T0**	**T+1**
Function			Question	CR	Answer
Triggering			The user asks for directions to a particular place	The ECA understands or does not understand the question and to reassure the understanding requests the user to clarify his/her utterance by saying "Huh?" or some other interjection that is relative to this meaning	The user acknowledges it by executing a repair
Duration			Time during which the user asks the ECA a question	0.6 sec	Time during which the user answers the ECA
BML	Verbal behaviour	Linguistic means	-	ha	-
		Intonation	-	falling	-
	Non-verbal behaviour	Head	directed at user	directed at user + slightly pushed forward	directed at user
		Forehead	neutral	slightly frown	neutral
		Eyebrows	neutral	slightly risen	neutral

		Eyes	direct gaze at user	direct gaze at user	direct gaze at user
		Mouth	closed	left slightly open	closed
		Handedness	no movement, both hands inside the pockets of the coat	no movement, both hands inside the pockets of the coat	no movement, both hands inside the pockets of the coat
		Fingers	N/A	N/A	N/A
		Body posture	directed at user	directed at user	directed at user
		Torso	directed at user	directed at user + leaning slightly forward	directed at user + slightly leaning forward
		Legs	no movement	one step towards actor	putting legs together to the initial position
	Distance		a little bit further from the user (ca. 1,5 m)	getting closer to the user (ca. 1 m)	close to user (ca. 1 m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-
	Type		-	clarification-request	-

Figure 16 describes the multimodal behaviour of a native speaker performing the Fragment/Interjection CR. It shows all three sequences, i.e. how the speaker looks before (T-1), while (T0), and after T (+1) performing this CR type. All three sequences need to be incorporated into ECAs' conversational behaviour in order to achieve a more realistic interaction with human users. Specifications for the multimodal behaviour involved in each sequence are found in Table 26.

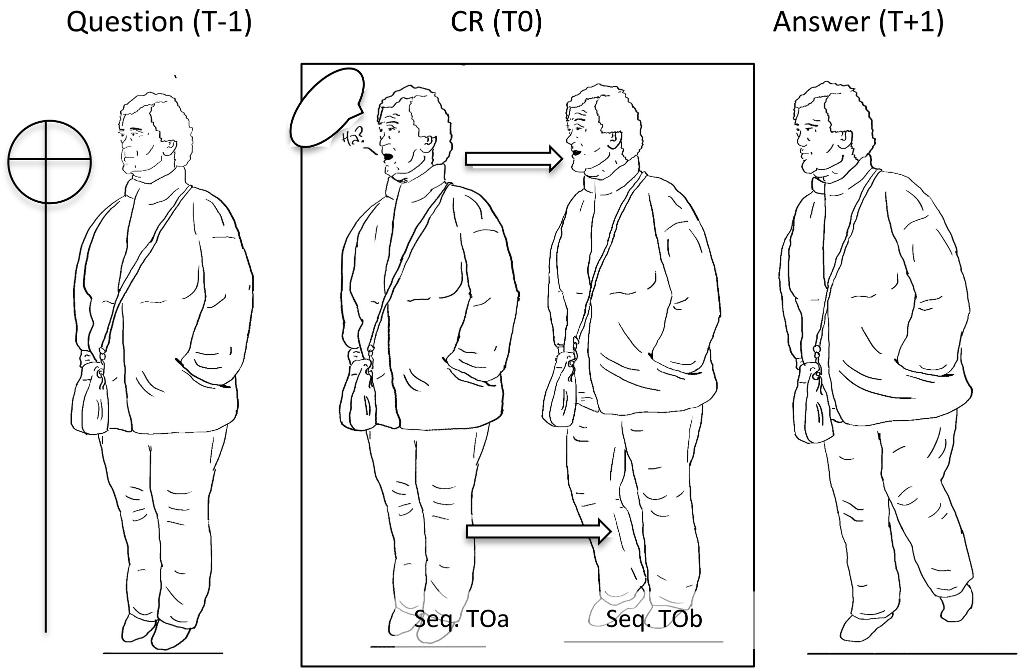


Figure 16: CR Fragment / Interjection Strategy sequence. The symbol on the left signifies the position of the actor.

The CR Fragment/Interjection is produced in the T0 sequence. This sequence is divided into two: T0a and T0b. The ECA pronounces the interjection in T0a and then leaves its mouth slightly open in T0b. In order for the ECA to do it in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The head is directed at the user (user's avatar) and is slightly pushed forward. The forehead is slightly frown and the eyebrows are slightly risen. The agent is looking directly at the user and the users says the CR (interjection) with a falling intonation as shown in T0a, and then leaves the mouth slightly open as shown in T0b. Neither hands nor finger movements are involved. The body posture is directed at the user. The torso is also directed at the user, however, according to the example above it is slightly leaning forward. The agent performs one small step forward to approach the user. The agent is in a close proximity to the user, which is about 1 m or less in real life. The model is based on the example above because its realisation is very similar to most of the instances found in the Fragment/Interjection category. Moreover, the picture in Figure 16 above is drawn based on an optimal video example in the collected corpus.

3.4.5 Repetition

The ECA can use the Repetition CR to ask the learner for a clarification about the previously said utterance. For instance, the learner asks: How do I get to the Hitt húsið? and the ECA produces a repetition of the same question, or part of the question, e.g. “How do I get to the Hitt húsið? or “How do I get to?”. If the learner asks differently, such as “Where is Hitt húsið?”, then the ECA can repeat the same, e.g. “Where is Hitt húsið? or “Where is?”. This is shown in the turn sequence order, where the first turn is the learner’s question marked as the CR sequence (T-1); the second turn is the repetition CR, and is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). Some of the multimodal behaviour in this CR Repetition indicates that the questions asked, e.g., Where is X or Where X is, are not performed as typical questions may have been, because the eyebrows of the speakers stay neutral. In typical questions, the eyebrows would be slightly raised but, in this case, the neutral eyebrow status could be interpreted such that the speaker is more concentrated on getting a clarification on the previous utterance, rather than asking a question out of curiosity. Also, the context and the situation is different: unknown encounters asking for directions to a particular location. The speaker may be more concentrating on the information from the actor and trying to understand what was said, which might affect the eyebrow movement. This sequence together with the multimodal behaviour associated with each turn is described in Table 27.

Table 27: The Multimodal CR Offering a Candidate Sequence Model.

ECA’s Repetition CR Sequence			
Sequence no.	**T-1**	**T0**	**T+1**
Function	Question	CR	Answer
Triggering	The user asks for directions to a particular place	The ECA understands or does not understand the question but needs to reconfirm it repeating a part of or the full phrase/question	The user acknowledge s it by executing a repair
Duration	Time during which the user asks the ECA a question	1.1 sec	Time during which the user answers the ECA

BML	Verbal behaviour	Linguistic means	-	“Where is …?” or “Where … is?” (this example is based on the speaker’s asking for directions)	-
		Intonation	-	rising-falling	-
	Non-verbal behaviour		-	rising-falling	-
		Head	directed at user	directed at user + slight turn to right/left (=direction of the place) and back at user	turning right/left (=direction of the place)
		Forehead	neutral	neutral	N/A
		Eyebrows	neutral	neutral	N/A
		Eyes	direct gaze	direct gaze + looking right and back at actor	N/A
		Mouth	subtle smile	from subtle smile to neutral, left slightly open	N/A
		Handedness	no movement, both hands in coat pockets	no movement, both hands in coat pockets	no movement, both hands in coat pockets
		Fingers	N/A	N/A	N/A
		Body posture	directed at user	directed at user	directed at user
		Torso	directed at user	directed at user + slight movement to the right/left (=direction of the place) and back at user	turned right/left (=direction of the place)
		Legs	no movement	no movement	no movement
	Distance		close to user (ca. 1 m)	close to user (ca. 1 m)	close to user (ca. 1m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-
	Type		-	clarification-request	-

Figure 17 shows how the multimodal behaviour appears on a native speaker performing the Repetition CR. It shows all three sequences, i.e. how the native speaker looks in all three sequences, i.e. in the turn before (T-1), while (T0), and after T (+1) performing this CR type. The symbol on the left signifies the position of the actor. All of the three sequences need to be incorporated into ECAs' conversational behaviour in order to achieve a more realistic interaction with human users. Specifications for the multimodal behaviour involved in each sequence are to be found in Table 27.

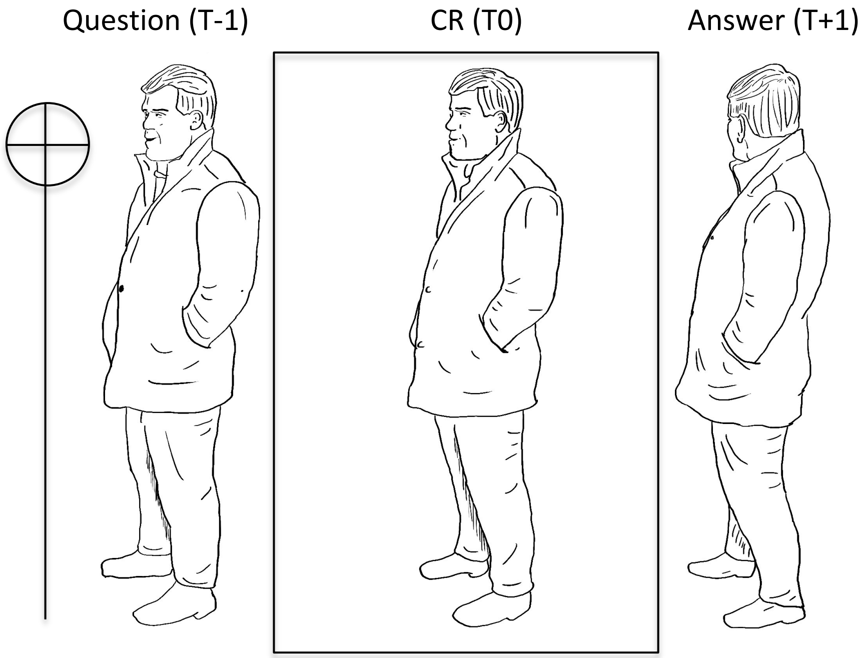


Figure 17: CR Repetition sequence. The symbol on the left signifies the position of the actor.

The CR Repetition is produced in the T0 sequence. In order for the ECA to do it in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The agent is directed at the user (user's avatar) and slightly turned to the left or right depending on the direction to the place the user asks about. Afterwards, the agent is directed back again at the user. The forehead and eyebrows are neutral and the agent is looking directly at the user. The agent then looks briefly left or right, depending on the direction of the place the user asks about. The agent looks back again at the user (synchronised with the head movement). The mouth has a subtle smile and it is left slightly open after saying the clarification word/phrase with a rising-falling intonation. Neither hands nor fingers are involved. The body posture is directed at the user as well as the torso.

The torso is slightly turned to the left or right (synchronised with the body movement). The legs are not moving, and the agent is in a close proximity to the user, which is about 1 m in real life. The model is based on the example above because its realisation is very similar to most of the instances found in the Fragment/Interjection category. Figure 17 above is drawn based on an optimal video example in the collected corpus.

3.4.6 Non-Verbal / "Freeze Look"

The ECA can use the Non-Verbal "Freeze Look" CR about the learner's previously said utterance. For instance, the learner asks: How do I get to the Hitt húsið? and the ECA produces a non-verbal clarification 'freeze look' by remaining silent and applying a specific multimodal behaviour that refers to the 'freeze look' , i.e. by looking directly at the user (user's avatar) and not saying anything. The use of no linguistic means enables the ECA to use only body behaviour to induce the user to repair him/herself. In the beginning of this non-verbal CR, the ECA has its mouth closed for about 0.3 sec (T0 sequence a), but after that it slightly opens the mouth for about 0.7 sec. (T0 sequence b). This triggers the user to repeat the previously said phrase/question. This is shown in the turn sequence order, where the first turn is the learner's question marked as the CR sequence (T-1); the second turn is non-verbal "freeze look" CR and is marked as turn sequence (T0); and the third turn is the answer of the learner marked as turn sequence (T+1). The specific multimodal features associated with this behaviour are described in the TO section in Table 28.

Table 28: The Multimodal CR Nonverbal 'Freeze Look' Sequence Model.

ECA's Non-verbal 'Freeze Look' CR Sequence			
Sequence no.	**T-1**	**T0**	**T+1**
Function	Question	CR	Answer
Triggering	The user asks for directions to a particular place	The ECA does not understand the question and requests the user to clarify it by the use of body behaviour	The user acknowledge s it by executing a repair
Duration	Time during which the user asks the ECA a question	0.7 sec	Time during which the user executes the repair

BML	Verbal behaviour	Linguistic means	-		-
		Intonation	-	[no sound]	-
	Non-verbal behaviour	Head	directed at user	directed at user, leans slightly forward	directed at user, bowed and put slightly forward
		Forehead	neutral	slightly frown	neutral
		Eyebrows	neutral	slightly drawn together	neutral
		Eyes	direct gaze at user	direct gaze at user	direct gaze at user
		Mouth	closed	closed for 0.320 sec and opened again for 0.768 sec	closed
		Handedness	no movement, both inside the pockets of the trousers	no movement, both inside the pockets of the trousers	no movement, both inside the pockets of the trousers
		Fingers	N/A	N/A	N/A
		Body posture	no movement, directed at user	no movement, directed at user	no movement, directed at user
		Torso	no movement, directed at the user	no movement, directed at user, bowed slightly forward	no movement, directed at user, bowed slightly forward
		Legs	no movement	no movement	no movement
	Distance		close to user (ca. 1 m)	close to user (ca. 1 m)	close to user (ca. 1 m)
FML	Track type		-	Interactional	-
	Functional category		-	Grounding	-

	Type	-	clarification-request	-

Figure 18 shows how multimodal behaviour looks on a native speaker performing the Non-Verbal ‘Freeze Look’ CR. It shows all three sequences, i.e. how the native speaker looks in all three sequences, i.e. in the turn before (T-1), while (T0), and after T(+1) performing this CR type. All of the three sequences need to be incorporated into ECAs’ conversational behaviour in order to achieve a more realistic interaction with human users. Specifications for the multimodal behaviour involved in each sequence are in Table 28.

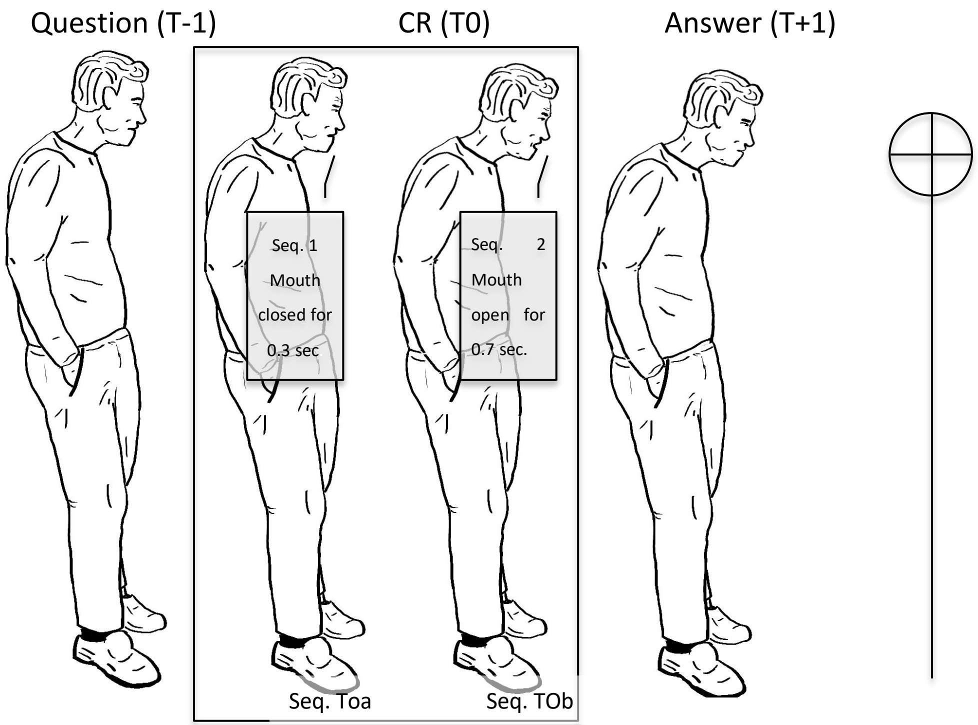

Question (T-1)
CR (T0)
Answer (T+1)
Seq. 1 Mouth closed for 0.3 sec
Seq. 2 Mouth open for 0.7 sec.
Seq. Toa
Seq. TOb

Figure 18: CR Non-verbal ‘Freeze Look’ sequence. The symbol on the right signifies the position of the actor.

The Non-Verbal ‘Freeze Look’ CR is produced in the T0 sequence. This sequence is divided into two: seq. T0a and seq. T0b. In order for the ECA to do it in a realistic manner, the following multimodal behaviour needs to be employed in the T0 sequence. The head is directed at the user (user’s avatar) and leans slightly forward. The forehead is slightly frown and the eyebrows are slightly drawn together. The agent is directly looking at the user. The mouth is closed for 0.320 sec (seq. T0a) and then opened again for 0.768 sec (seq. T0b). No sound comes from the mouth (mute), neither the hands nor fingers are involved, body posture is still, no movement of the torso, close proximity to the user, which is about 1 m or less in real life. The model is based on the example above because its realisation is very similar to the second instance produced by the same native speaker. Figure 18 above is drawn based on the video example in the collected corpus.

3.5 The Pilot Study about User Responses to *Virtual Reykjavik*

This chapter describes the pilot user study, which was conducted to explore the efficacy of the endowed ECA's in interacting with learners. It describes the process of learners playing the game *Virtual Reykjavik* and interacting with ECAs that were endowed with multimodal features in two CR strategies. The pilot study is the second auxiliary study, which is based on the theoretical notion of the IMTEE model (Figure 8). This model suggests that it is important to evaluate experiences of users of artificial systems, in our case *Virtual Reykjavik*, in order to gain a better insight into what the user's general and specific experience with the system.

3.5.1 Stating a Problem

After the identification of the six multimodal CR models described earlier, the ECAs in *Virtual Reykjavik* were endowed with the ability to respond in accordance with selected behaviours identified in two of the six CR strategies, the CR Ellipsis and the CR Fragment/Interjection Strategy. Due to resource constraints, only these two could be implemented prior to the survey being executed. The success of these features in the ECA behaviour was then tested in a follow-up study which is presented in this chapter. During the study, the game was tested for the first time with six adult learners of Icelandic who were first-year students in the Icelandic Practical Diploma Course at the University of Iceland. For the testing part, *Virtual Reykjavik* had only one story/scenario, Týnda hljómsveitin (The lost music band), with one chapter, Hvar er Hitt húsið? (Where is Hitt húsið?). The game environment was populated with ECAs that were endowed with human-like multimodal behaviour in various communicative functions other than the CR (turn-taking, feedback, emphasis, reference, and another new function, the Explicit Announcement of Presence (EAP) (Ólafsson, 2015)), and then the CR, which is the focus of this thesis.

The aim was to determine to what extent learners noticed the multimodal features incorporated in ECAs' conversational behaviour when executing the communicative function of CRs, and whether this behaviour would help them learn something about Icelandic language and culture. The aim was to investigate this in an indirect manner, i.e. not to address the learners with questions about the execution of CRs, but to find out which multimodal behaviour of the ECAs they noticed in the course of learning and playing the game. The answers would help to detect how ECAs executed the multimodal CR models. The following research question helped to guide the study:

6. How do learners perceive playing *Virtual Reykjavik*?
7. How do learners perceive the interaction with ECAs in the game?
8. How do learners perceive the multimodal behaviour of the ECAs, i.e. speech, facial expressions, hand gestures and body posture while engaged in CRs?
9. Does the ECAs multimodal behaviour feel natural?
10. How effective is the game for learning Icelandic language and culture?

3.5.2 Methodology

A mixed-method procedure of Concurrent Triangulation Strategy (Creswell, 2009, p. 213) was used for data analysis. The approach chosen for this pilot study was a qualitative mixed method study that included a questionnaire, personal interview and video recordings of participants playing the game. As Zou (2008) points out, "[f]indings from limited source of data may not be holistic because they cannot prove their results from different methods. Therefore, it is important to use various methods research to explore more sources of data (such as interviews, questionnaire and observation) to carry out an in-depth study in the CALL field" (p. 156). Figure 19 graphically demonstrates how mixing different methods leads to the triangulation of data based on answers from different sources, such as interviews, questionnaires and observation (Flick, 2018, p. 779).

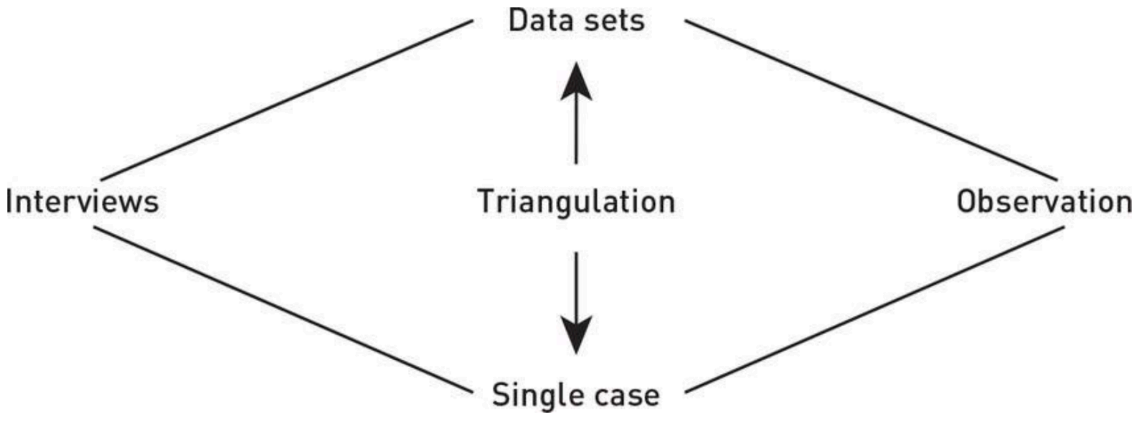


Figure 19: Levels of triangulation when using different methods in qualitative research (Flick, 2018, p. 779).

The questionnaire was not based on any standardised set of questions from previous research. Instead, for the purpose of originality new questions were created in this pilot study. The aim was to tailor-make this questionnaire and create a specific set of questions that would provide answers to the five research questions stated earlier in this chapter. Adopted standardised sets of questions from previous research may use measures that would not be possible to measure in the current state of this game, and could therefore represent a drawback or a limitation to the study (Lane et al., 2013, p. 8). In order to select learners with an equal probability to participate, the Random Sampling Method (Creswell, 2009, p. 148) was selected. A specific group of participants was approached to sample, i.e. first-year students enrolled in the Icelandic Practical Diploma Program at the University of Iceland. During the last week of November 2015, an email was sent to all students in this group asking for participation in the pilot study. In order to motivate them to participate, the announcement included information about a prize, which was an ISK 4000 (USD 35) voucher for the local student bar, that one participant could win. Participants could voluntarily apply by email. The confirmation email remained confidential and was used only for communication purposes between each participant individually to reconfirm the time and place of testing, and for announcing the winning voucher to the winner directly. The aim was to receive ten applicants, which would be approximately a half of those participated in the first study of an online survey (21 participants). Informed consent was obtained from all participants. There were two groups of participants - the instructor and the learners of Icelandic as a second language. The instructor was a volunteer member of the team of investigators from the University of Iceland, who would be seated behind the learner to monitor the testing and advise if necessary. Only six participants applied, four females and two males, aged between 22-31. They represented six different nationalities: Canadian, German, Latvian, Russian, Serbian and Turkish. All six learners were first-year students of the Icelandic Practical Diploma Program at the University of Iceland, five of them were beginners and one intermediate. Three of them were temporary (visiting students) and the other three permanent residents in Iceland. All participants met the criteria of being first-year learners of Icelandic as a foreign or second language, however, at a mixed level of Icelandic, beginner and intermediate, which is typical for these classes. Although six is a small number of participants, in qualitative research, e.g. interviews, the number of participants under twenty can be justified by saying that it "will facilitate the researcher's close association with the respondents, and enhance the validity of fine-grained, in-depth inquiry in naturalistic settings" (Crouch and McKenzie, 2006, p.

483). This small number of participants is not a representation of the population, but it represents a sample of participants in a context of a case study (Firestone, 1993, p. 16). The heterogeneity of the group was represented by their gender, age, and nationality.

The pilot test took place in the Black Hole multimedia laboratory of the Center for Analysis and Design of Intelligent Agents (CADIA) at Reykjavik University during the first two weeks in December 2015 (Figure 20). Due to a technical obstacle of a Unity engine being not compatible with new versions of Google Chrome, the laboratory enabled a repeated download of the old version of Google Chrome onto desktop computers, which was then used for testing. The testing was under the supervision of assoc. prof. Hannes Högni Vilhjálmsson, and in cooperation with the investigative team from the University of Iceland, which was under the supervision of prof. Birna Arnbjörnsdóttir. Altogether, three volunteering staff were present to aid the students. The first volunteer was myself (University of Iceland). I was present during all sessions and stages of the study. I needed to be familiar with the game and had to be always present in the laboratory throughout the whole test period. The second volunteer was another member of the project team (Reykjavik University) who made sure that the computer was set up and ready to be used before each test session took place. And the third volunteer was the project supervisor (Reykjavik University) who made sure that the laboratory was reserved and available for the time of testing.

Figure 20: Supervisor and participant in the multimedia laboratory Black Hole. This photo was taken during the pilot study at the laboratory situated at CADIA centre (Reykjavik University) playing *Virtual Reykjavik*.

All participants and staff were deemed volunteers. Before testing, the learners were informed that the game was a prototype and included only one chapter and one story with a few tasks to be solved, without achieving further levels in the game. In addition to that, they were advised to take as much time as they wished to in order to repeat the same chapter of the game, or the task, to explore its environment, to speak to the agents *via* the external microphone on the desk beside the keyboard of the computer, and to give the agents some more time to 'think', i.e. to process the speech input.

The learners received instructions about which story and chapter to select once they started playing, and that the instructor would only then help them translating some words from Icelandic into English if they did not understand the menu options in the game to proceed further. The testing consisted of three parts: 1) letting the learners play the game, which was recorded on a video camera; 2) filling in a questionnaire, which was given to learners immediately after playing the game; and 3) taking part in an interview, which was taken with each learner individually after the questionnaire had been filled in, and which was recorded on an audio device. The learners' task was to play the game. The first scenario consisted of unknown first encounters asking for directions to Hitt húsið, which is a name of a building in central Reykjavik, and there were three tasks to fulfil:

1. *Náðu athygli einhvers* (Get someone's attention);
2. *Hvar er Hitt húsið*? (Where is Hitt húsið?) to find directions to Hitt húsið; and
3. *Segðu bless* (Say goodbye).

These tasks represented the dialogue structure in the first game scenario.

The learners received the following instructions by demonstration shortly before playing the game. The instructor (myself) sat beside participants and showed them how to navigate in the game. The following instruction was not written down but demonstrated by showing and speaking: *command keys W, A, S, D are used to move back, forth, and sideways; the ESC key is used to activate 'freeze mode' to allow the user to explore the environment of Austurvöllur Square by moving around with the mouse: when the mouse-arrow changes to a magnifying glass, a click on the left mouse button displays cultural information about a particular monument or a building present in the environment; the P key is used to allow the user to go back to the main menu from where they could start the game again; the key M for markmið (goal) is used to enable the user to view the task window, which consisted of three tasks that the user had to fulfil; by clicking on the "i" icon above the main menu it was possible to activate English labels over particular Icelandic words in the menu windows of the game; and the mouse is used to serve for 'first person view' and for direction of walking*. However, the participants were not told beforehand how to operate the 'speaking' button on the mouse because there were clear instructions on the screen, which they could follow while already playing the game. After selecting *Spila* (Play) in the main menu, the user was given the option to either select an already present user, or to create a new one, and then to click on *Áfram* (Continue) (Figure 21).

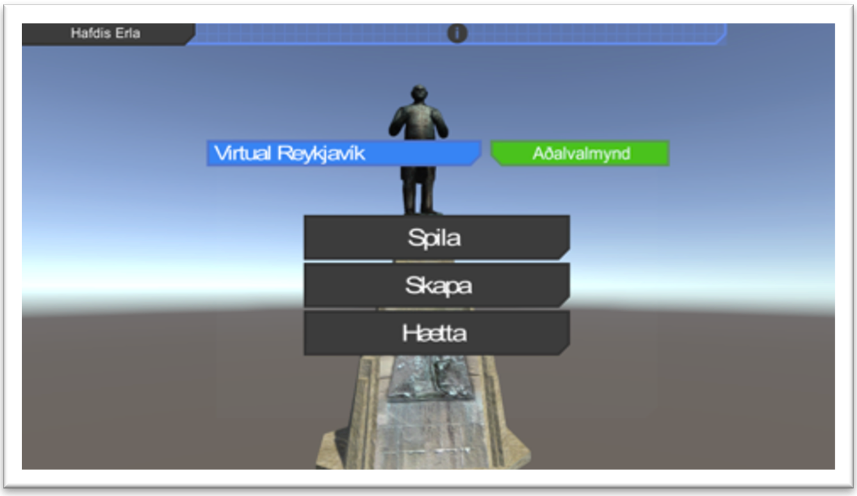


Figure 21: Main menu in *Virtual Reykjavik*.

After that, the user was given the option to select between two stories. The top option was the only active option, *Týnda hljómsveitin* (The lost music band), which also included a short description: *Þú ætlar að skrifa grein um íslenska hljómsveit, en finnur hana hvergi* (You are going to write an article about an Icelandic music band, but the band is nowhere to be found). After selecting it, the user was given a further option to select between two chapters. The first one from the top was called *Hvar er Hitt húsið* (Where is Hitt húsið) and it was the only active one which was designed for testing. It also included a short description: *Kannski er hljómsveitin að æfa í Hinu húsinu. En hvar er það?* (Perhaps the music band is rehearsing in the Hitt húsið. But where is that?). After clicking on it, the user found him-/herself in the virtual Austurvöllur Square and could start playing the game. Figure 22 shows a learner's first-person view in *Virtual Reykjavik* when approaching an agent.

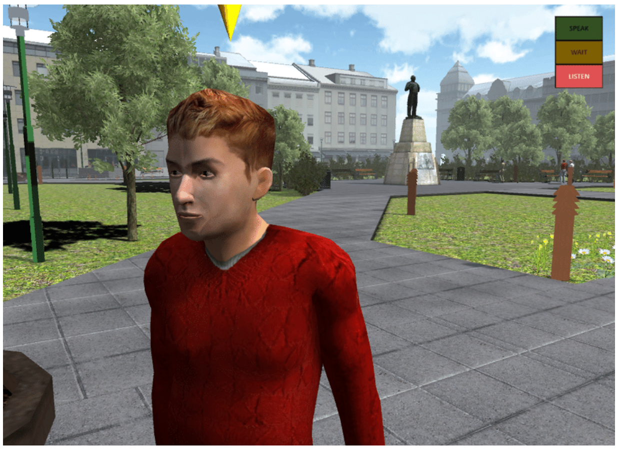


Figure 22: Learner's first-person view when approaching an agent. In this picture, the learner has approached the agent; the yellow arrow above the agent indicates agent's readiness; the red box in the top right corner indicates that the agent is listening.

Participants' privacy was addressed in this study. Despite participation being anonymous and participants signing an informed consent, they provided their email address in order to sign up for the testing part. During the testing, they were recorded on video, audio and provided written answers to the questionnaire. In order to ensure their privacy, all collected materials were stored in a protected file on a coded hard disc, so that only the team of researchers working on this project has the access to it. Names or any other information about participants such as voice, photograph, or video helping to identify or somehow revealing their identity will not be published. The email address was used for drawing one candidate who won the voucher. The candidate was contacted for the winning voucher to be delivered to him/her. The emails were afterwards deleted.

3.5.3 Data Collection and Analysis

The testing proceeded in three parts. During the first part, the learners were recorded on a video camera while playing the game. The camera was of type Canon EOS XS, had an inbuilt microphone and was placed to capture the interaction between the subject and the supervising staff. This served the purpose of noticing how much interference there was needed from the supervisor, who sat next to the participants, and to review the participants' reaction while playing the game. Due to technical problems, only four out of six recordings were made. After the testing, the video recordings were reviewed and notes on how participants were playing the game were taken and compared with the answers from both the questionnaire and interview.

During the second part, the answers from learners about questions related to pilot study were collected *via* a questionnaire. Each learner was left alone to fill in the questionnaire, which asked about their overall experience, the perceived level of ease and difficulty in the game, the learning effect, and the perception of multimodal behaviour of ECAs. The questionnaire included the following questions:

1. What is the main reason you are learning Icelandic?
2. How did you find playing the game? (Choose one per line):
    a. Enjoyable/neither/frustrating
    b. Boring/neither/exciting
    c. Difficult/neither/easy
    d. Educational/neither/pointless
3. Other terms that describe your experience (maximum 3 terms)
4. Did you encounter any difficulties?
5. How easy/difficult was it for you to understand the agents in the game when he/she spoke Icelandic to you? (Choose one): very easy/easy/neither/difficult/very difficult
6. Did playing the game help you learn anything new about Icelandic language?
7. Did playing the game help you learn anything new about Icelandic culture?
8. Any comments or suggestions?
9. How did you perceive the agent's behaviour regarding:
    a. Spoken language: natural/neither/robotic
    b. Facial expressions: natural/neither/robotic
    c. Hand gestures: natural/neither/robotic
    d. Body movement: natural/neither/robotic
10. Did you find the agents were… (Choose any that apply): natural, disturbing, appropriate, robotic, friendly, inappropriate, or other? ...
11. Did you notice any particular expressions in the agent's behaviour, e.g. particular facial expressions, hand gestures, body posture, etc.)?
12. I think the agent did not understand what I said: true/false
13. The interaction with the agent felt natural: true/false
14. I think the agent really listened to me: true/false
15. I think that the agent understood what I said: true/false
16. The interaction with the agent(s) felt satisfying: true/false

17. The agent’s overall behaviour felt: (Choose any that apply): natural, warm, fake, positive, disagreeable, spontaneous, negative, sincere, disinterested, agreeable, cold, interested.

Once the questionnaire has been filled in, a follow-up interview took place, conducted by a member of the supervising staff. Answers were recorded using the Voice Memos application on an iPhone. After the interview, all answers were transcribed into a Word document. The audio recordings were transferred in a coded file on a separate hard disc and deleted from the device. The interview asked about the participant’s general description of the playing and learning experience of the game, how the interaction with ECAs was perceived, and about any notable or disturbing element of the game.

The third part, the interview, which immediately followed the questionnaire (see Appendix C – Questionnaire for the Pilot User Study) and was recorded on an audio device. The supervising staff would sit next to the user and ask the following seven questions:

1. Can you describe how was your general impression from playing the game?
2. What did the game help you learn?
3. How did you perceive the interaction with the agents?
4. Can you describe how did the agents behave when they interacted with you?
5. Was there anything eye-catching in the game?
6. Did you find anything disturbing in the game?
7. Is there anything else you would like to suggest?

The interview was taken in order to solicit views and opinions from the participants on playing the game, their reactions to the multimodal behaviour of ECAs, and the learning effect. The questionnaire was given to the individual to answer questions related to the research objective of playing the game, and the qualitative observation and video-recording of learners playing the game allowed the researcher take notes on the observed behaviour and activities of an individual during the process of playing the game. In this pilot study, the data from all three methods of analysis (observations of learners playing the game, questionnaire and interview) were compared to find out how learners perceived playing the game and the interaction with ECAs. In order to better analyse the answers, it is important to find common themes in both the questionnaire and the interview, and to compare them. Structured interview (Brinkmann, 2018, p. 1001) with questions motivated by the questionnaire was conducted and answers were analysed using thematic analysis (Braun, 2014). According to Brinkmann (2018, pp. 1001), structured interviews are based on a strategic logic of questionnaires and lead to answers that can be compared across participants, and possibly quantified, if the study is large enough. Interviewers read questions with exact wording to each participant, and do not provide any information beyond what is written in the questionnaire. Structured interviews are in other words passive recordings of people's opinions and attitudes and do not take advantage of a dialogue between the interviewee and interviewer. This kind of interview seemed suitable for the purpose of this study. In order to make a better sense of how learners were playing the game, video recordings were used to provide data of real-time interactions with the agents in the game. According to Margolis and Zunjarwad, (2018), videos create a way to study real-time interactions and can draw attention to details and reconstruct complex patterns of actions (p. 1051). Although this method is usually used in Micro Ethnography, it seemed suitable here as well. The results are presented in the following section.

3.5.4 Results from The Pilot Study

The results described in this section are divided into three parts. First, the results from the questionnaire are presented, followed by the results of the interviews and finally the results from the observation of video recordings are presented. Lastly, the summary of results is provided. This section sheds light whether the learners could learn anything while playing and interacting with ECAs in *Virtual Reykjavik*. Moreover, the results provide information how the learners perceived the ECAs' multimodal behaviour including the CRs.

3.5.4.1 Results from the Questionnaire. In order to determine whether participants were learning Icelandic as a foreign or second language, they were asked whether they were temporary or permanent residents. This would also help to compare the results and find out any possible differences for example in terms of motivation based on this factor. Figure 23 shows the proportion of permanent and temporary residence of participants in Iceland.

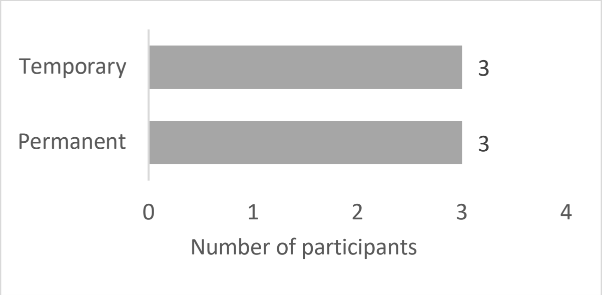


Figure 23: Proportion of permanent and temporary residence of participants in Iceland.

The participants gave the following general reasons for learning Icelandic as a foreign or second language:

a) To become part of Icelandic society (permanent resident);
b) I live in Iceland and study at the University of Iceland (permanent resident);
c) To be able to communicate and understand Icelandic in daily life (permanent resident);
d) Linguistic research, enjoyment in Icelandic literature, culture and music (temporary resident);
e) It was my dream to come here and learn Icelandic because I like the country (temporary resident);
f) Interest in Nordic countries, languages and culture (temporary resident).

The answers above show that permanent residents are second language learners and temporary residents are foreign language learners, because their common denominator is a short-time residence in order to learn Icelandic for various motivational reasons.

The summary of results from the questionnaire about learners' perception of the game is summarised in Figure 24.

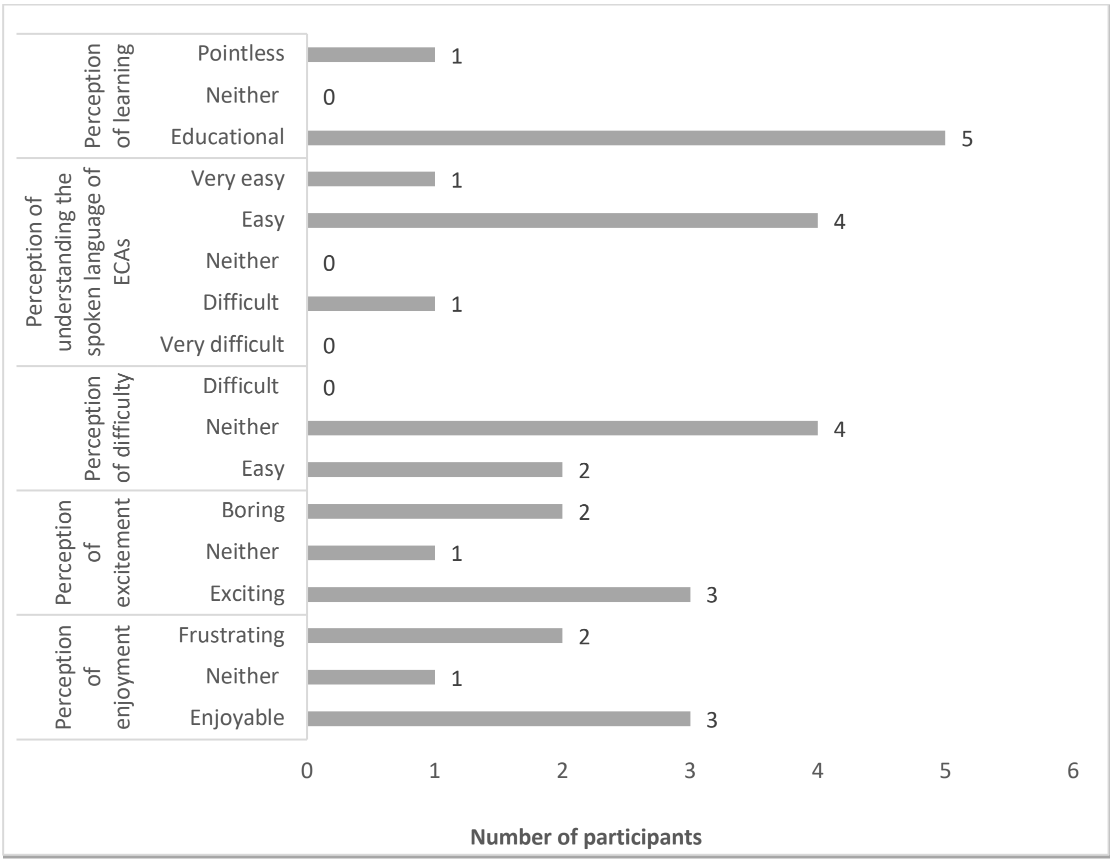


Figure 24: Learners' perception of various parts in the game. The chart shows results from the questionnaire during pilot testing.

Regarding the learning effect, i.e. whether the learners learned anything while playing and interacting with ECAs, and how they perceived the ECAs' multimodal behaviour including CRs, the results were as follows. Only three out of six learners indicated that they learned something: two learners learned about culture and one about the language. The rest of the learners did not indicate anything. Regarding playing the game, five of the learners indicated that the game was educational, three felt it was relatively easy to play and exciting, however, one learner indicated that it was pointless, two frustrating, but none said that it was difficult.

Other terms that learners also used to describe the game were: amusing, funny, interesting, slow, static, it just didn't work. However, five out of six learners encountered the following difficulties while playing the game:

- The computer did not understand me, I was immediately stuck and had to repeat one million times;
- Problem with voice recognition,
- Some problems with the microphone (maybe I wasn't loud or understandable enough),
- Difficulty in producing speech that would be correctly interpreted by the software,
- Difficulty in obtaining a response from the agents,
- Pronunciation,
- Commands,
- Smoothness of movement.

Most of the learners thought that the agents did not understand what they said, even though most of them thought that the agents seemed to listen to them. Results from the questionnaire divided users equally into two groups: one group (three learners) felt the interaction with the agents natural, but the other group (three learners) did not. However, most of them (four learners) thought the agent's behaviour in interaction with them was cold because the agents did not seem to show any emotions or give smile. The perception of agent's overall multimodal behaviour is presented in Figure 25.

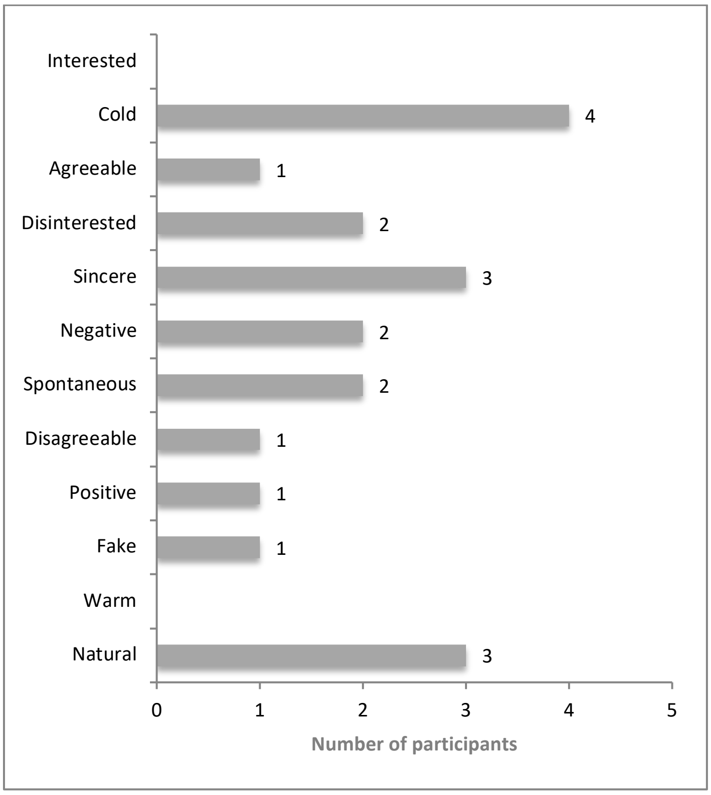


Figure 25: Perception of the agent's overall behaviour.

Four out of six learners noticed the following expressions in the agent's multimodal behaviour:

- A male agent crossing arms on his chest when interacting with the learner,
- A female agent slowly folding her hands,
- A 'pained' expression when the agent misunderstood the learner's speech,
- One learner noticed that the multimodal behaviour was not very natural, and he/she did not feel that they understood him/her.

The perception of a particular agent's specific multimodal behaviour is presented in Figure 26.

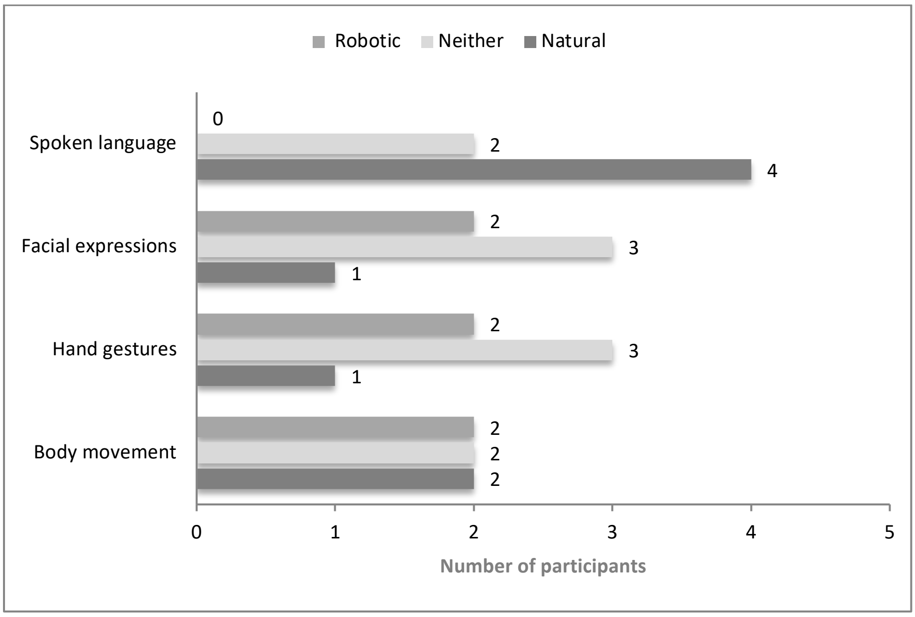


Figure 26: Perception of agent's specific multimodal behaviour.

The above results indicate that learners encountered several difficulties when playing the game. Nonetheless, three of them indicated that they learned something related to culture and language. Despite the learners having perceived the ECAs' multimodal behaviour differently, they did notice a pained expression in the agent's face when executing a multimodal clarification strategy. The spoken language of ECAs included the two CR strategies with a specific multimodal behaviour, but as mentioned above, questions were not formed to investigate the multimodal behaviour of CRs directly.

3.5.4.2 Results from the Interviews. The interviews reconfirmed that most of the learners said that they did not learn anything new from their session. One learner mentioned that the game helped him/her learn to start speaking in Icelandic, and other two learners mentioned that they learned cultural information about two buildings: the parliament building Alþingishúsið and the cathedral Dómkirkjan. However, other four learners mentioned that once the game would be ready, players without previous knowledge of Icelandic can learn about buildings and monuments in Reykjavik, and that it will make it generally easier for the future players to start talking and communicating in

Icelandic. According to these participants, it seems to be easier in silico and with virtual agents, than in reality with real people.

During the interview, learners were asked to describe their general impression from playing the game and whether the game helped them to learn anything new about Icelandic language and culture. Three out of six learners found the game in the following way:

1. Exciting in regards of both being able to talk to virtual agents and seeing what the agents were going to say,
2. Fun when playing,
3. Impressive as to the general visual surroundings in the game, such as design of the buildings; and
4. Helpful in regard of a future reference, i.e. once the game is ready it will enable to practise a conversation in Icelandic.

However, other three learners experienced frustration over the following:

1. Slow pace of the game and no option for running or going faster,
2. Not being recognised by the speech recognition system,
3. Lack of introductory information about the goal of the game and purposes of the tasks before starting to play, and
4. Lack of ability to ask different questions than those presented in the tasks.

Unlike in the questionnaire, in the interview, there was no question about the difficulty level of the game. Some learners indicated that the difficulty part was rather connected to the speech recognition system and not being recognised when speaking *via* the microphone to the agents than to playing the game itself, except for a slow pace of the game. Some learners' speech input was not recognised due to either speaking too softly to the microphone or incorrect pronunciation of the letter <r>.

In the interview, there were two general questions about the agents. The first question was about how the users perceived the interaction with the agents, and the second one was to describe how the agents behaved when they interacted with the user. Spoken language was generally perceived in the following way:

- It was interesting to talk to the virtual agents and see what they were going to say next,
- It was nice when each of the agents said something different to the user and that also the user could say different things to them,

- During the first task when the Explicit Announcement of Presence (EAP) (Ólafsson, 2015) communicative function was used by ECAs, it wasn't just one *góðan daginn* (good day) with the definite article, but you could say to them *góðan dag* (good day) without the definite article as well, so it seemed like you were actually speaking to someone;
- In this sense, the agents were quite realistic because they were open and wanted to talk, but there was one guy (male ECA) who crossed his arms on his chest and said that he knew where *Hitt húsið* was, so that is like real people are;
- Some agents pointed out which direction one was supposed to go to,
- The agents responded in a clear and very natural way,
- But one user expressed frustration over the agents' spoken language due to their restricted response options to the particular task because it was not possible to ask the agents any other questions. This was not the weakness of the agents because they were deliberately modelled to provide specific answers to particular task-related questions.

More specifically, the spoken language of the CR models was perceived in the following way:

- According to some users, the agents were quite rude when they only said: "I don't hear anything", or "Ha?";
- But, on the other hand, the agents were perceived to behave like some Icelanders do, i.e. whatever you say to them, they all the time response with "Ha?".

As mentioned above, the implemented CR model was supported by the study on multimodal CRs and included an Interjection Strategy, which was designed so that the ECAs would say *Ha?* in case of a miscommunication or an improper input from the users. As a result, the spoken language in CRs was perceived as slightly rude, but natural in general. One learner suggested more politeness to be implemented into the agent's' behaviour, which should make agents say: "Sorry, I can't hear you" instead of only "I can't hear you", but this suggestion would not correspond with the natural language use in reality by Icelandic speakers. Although this would make the agents more polite, it would not correspond to the multimodal CR models and to how people really speak and behave. Moreover, another learner was frustrated for not being understood by the agents because they were only saying: "I don't understand you" or "Ha?", and, according to this learner, because of this there was really no conversation going on. There was no mention about the other CR “Hitt húsið?” strategy.

In general, learners indicated that the facial expressions looked natural but also ‘different’ or creepy depending on where one was standing. The only record about facial expressions connected to a CR strategy was by one learner when he/she mentioned that the agents had very “good” facial expressions. This could mean that the agent’s facial expressions were clearly visible to the learner. The learner did not specify more about the facial expressions and mentioned that the only problem was that the agents did not understand him/her. This indication is triangulated from the observations. Hand gestures were perceived only generally as pointing gestures or crossed arms, but none related to the CR strategy, which is in concordance with the two models that did not include any particular hand gestures. None of the users explicitly mentioned the body movement when ECAs were executing CR strategies; they only characterised it from a general point of view as very good and natural (three learners) and robotic (three learners). This observation is in concordance with the multimodal behaviour implemented in the CR models, which includes, if at all, very subtle body movements, such as directing the agent’s body towards the learner’s avatar and a slight movement forward of the agent’s torso. This may not have been noticeable if the learner’s avatar was too close or directly in front of the agent.

3.5.4.3 Results from the Observation of Video Recordings. After the learners filled in the questionnaire and answered questions in an interview, the notes from observations of video-recorded playing sessions were reviewed. Problems in pronunciation were noted on the video camera recordings, e.g. one learner had a difficulty pronouncing sound of <r> and <tt> consonants in Icelandic, which consequently caused the speech recognition software to fail to properly recognise the speech input. As a result, this triggered a frequent execution of a CR function (Interjection Strategy) by the ECAs. This situation caused that the learner to become slightly frustrated. Due to this fact, it became obvious that the learner constantly tried to improve the pronunciation and eventually managed to pronounce the word or phrase so that it was recognised. The effect of this was that the frustration persisted. Despite the frustration, the CR function led as a side effect to improved pronunciation of the above consonants.

The observations of video recordings revealed that the learners needed assistance in many cases. For instance, they were guided by the supervising staff throughout the whole session when playing the game as to where to click, what to do when they wanted to start again. They were also advised by the supervising staff to repeat or speak more slowly when the speech input was repeatedly not understood by the agents, or when the learners were too often or too quickly clicking on the left mouse button to activate the microphone and speak, which resulted in the agents ‘not understanding’ or ‘not hearing’ them. Moreover, it was sometimes needed to remind the learners to give the agents more time to process the input. This was reflected in the learner’s occasional frustration. Nonetheless, the CR function was very effective, especially when the ECAs did not receive the input, which meant that despite the above-mentioned hindrances, the agents tried to keep a continuous flow of a conversation.

The video recordings indicated technical hindrances between the learners’ speaking to the microphone and the system’s ability to recognise the spoken input. The learners had often experienced difficulties with being understood, i.e. either their pronunciation was suboptimal, or the speech recognition system did not work properly. One learner asked the supervising staff to repeat the same phrase to the microphone for him/her to see whether it would be recognised. Neither the supervising staff’s nor the learner’s phrase was recognised, and it had to be repeated for the system to be able to recognise the spoken input. During this process, the CR function was often triggered, which indicated that the ECAs were trying to keep the conversation going in a natural way. No particular results from the observations of the video recordings were gathered in relation to the multimodal behaviours of ECAs when executing the CR request, as the purpose was to observe how learners behaved and responded to tasks during playing the game, and how the CR functioned.

3.5.4.4 Summary of Results from the User Response Study. This section presents a summary of results from the questionnaire, interview and the analysis of video recordings from the pilot study. In general, the combination of results from the questionnaire, the interview and the observations from video recordings suggest that half of the learners perceived playing the game as enjoyable and fun, but also as educational, easy and frustrating mainly due to pronunciation difficulties and the lack of smoothness in movement of learners’ avatars. Based on the questionnaire and the interviews, for most of the learners, the game had a learning effect in both languages as to starting to speak in Icelandic, and culture as to learning about buildings in central Reykjavik. Moreover, the execution of the CR Interjection Strategy by ECAs demonstrated how real Icelanders speak, because they all the time said “Ha?” when they did not understand something. In spite of the game being educational, it was also frustrating especially when the agents did not understand the learner when having a difficulty in pronouncing sounds of consonants <r> (voiceless alveolar trilled [r]) and <tt> (voiceless combination of sounds [xt]) in Icelandic, which triggered a CR function by the agents and made learners repeat their spoken part.

In case of pronunciation, there was a reported learning effect because the learners had to repeat certain words several times until they were understood by the speech recognition. In this view, the CR function was effective because it led to the improvement of pronunciation. The interaction with ECAs was often perceived as difficult because the learners felt the agents did not understand them. This was mostly due to technical problems such as the internet connection lagging behind and consequently the ECAs' not understanding the learners' input, or when learners spoke too softly to the microphone and/or did not pronounce certain words properly, which consequently triggered the communicative function of CR by the ECAs. The multimodal behaviour of ECAs in relation to the CR Interjection Strategy model was perceived as if the agents had a 'pained' expression when they misunderstood the learner. Another learner perceived the facial expression as very good but did not specify exactly what it meant. The learners mainly noticed the ECAs' verbal language. According to some learners, the agents were a bit rude, although genuine, when they only said: "I don't hear anything", or "Ha?". Agents were perceived as life-like because it is what Icelanders often say in real life, i.e. whatever you say to them they all the time response with "Ha?". As such, learners mostly noticed the CR Interjection Strategy "Ha?". Even though the ECAs also used the CR Ellipsis by saying "Hitt húsið" (an exact number of occurrences in the game is not known), the learners did not mention any information regarding this type. The frequency of any of the CRs by ECAs was not measured. The ECAs usually used CR Ellipsis when they received a correct input. The process was as follows: the ASR recognised the learner's words when saying "Hitt húsið" and consequently the ECAs asked a clarifying question whether it was that place. The assumption is that the CR Ellipsis seems more natural in a dialogue even though it is sometimes not necessary. Its function is two-fold. Firstly, to clarify or reconfirm that the learner is really asking directions to Hitt húsið. And secondly, as a rhetorical question that similarly helps to keep a more natural flow of a conversation, because the learner gets a confirmation that the agent has understood him/her. In both of the cases, this CR contributes to reaching a common ground in a dialogue. On the other hand, the learners noticed that the CR Interjection Strategy "Ha?" was used more often than the Ellipsis. The

Interjection Strategy could occur a) more often because the ECA did not understand the learners' spoken input through the ASR; and b) the learners might have noticed it more because this Interjection Strategy is culturally bound. Ha? is pronounced with falling pitch, which in some other European languages, such as Dutch, is produced with a rising pitch (Dingemanse et al., 2014, p. 11). Compared to the frequency of use in real language with real speakers, the previous study on multimodal CRs informed that the Ellipsis was used in 80 instances (79.21%), whereas the other five CR types, including the Interjection Strategy, were only used in the range of 2-8 instances per CR which is 1.98%-7.92%. The agents were not programmed to use Ellipsis more frequently than the Interjection Strategy. The choice of learners' words and other usually technical problems caused that the Interjection Strategy seemed to be more often used. It was probably due to the fact that the learner's input was not properly understood which triggered the Interjection Strategy. Despite the Interjection Strategy being less used in real life, it should nonetheless be considered as a valid instance in *Virtual Reykjavik*, because this belongs to clarification strategies used in real life. Regarding body movement and hand gestures during both CR models, the learners did not notice anything particular, which on the other hand is in concordance with the multimodal behaviour implemented in the models that include subtle body movements of the agent's body towards the learner's avatar and a slight movement forward of the agent's torso. The following section concludes on the three studies presented in this chapter.

3.6 Conclusion

The purpose of the first auxiliary study, which was conducted in a form of an online survey, was to give answer to six different questions that are discussed as follows. There were eighteen female and three male L2 learners of Icelandic from twenty different countries. They expected to practise similar language skills in both the language course and the computer game, however, less speaking in *Virtual Reykjavik* than in the traditional classroom. The learners moreover expect the game to have a good storyline with particular conversational scenarios for speaking practice, a voice recognition in order to be able to communicate with virtual agents that would take on the role of native speakers and interact with them in the game on various topics. The game would contribute to building a confidence by interacting with virtual agents and by enabling them to focus only on Icelandic switching to English. The disadvantage would be that the agents would not be able to give as much feedback as a regular teacher would, thus resulting in a more laborious work with dictionary and individual learning. Since learners feel negative and intimidating when speaking with native speakers of Icelandic in real life, they would expect the game to become a practical tool for practising Icelandic oral language skills before using the language in real life. As indicated before, this survey neither included questions about multimodal behaviour of ECAs, nor about multimodal CRs. Its purpose was to gather general information about learners' expectations from *Virtual Reykjavik* populated with ECAs. The following study will in more detail discuss the particular multimodal behaviour in CRs in Icelandic, which partly contributed to designing a more realistic human-agent interaction with learners in the game.

The main study on multimodal CRs helped to answer the three previously stated research questions in the following way:

1. In a given scenario of asking for directions in central Reykjavik, the common CRs used by native Icelandic speakers (men and women, aged 18-70) in a face-to-face interaction with other native and non-native speakers (actors, men and women, aged 20-40) in order to initiate speech repair are the following six types: (1) Ellipsis, (2) Full/Partial Explicit Query, (3) Fragment/Interjection Strategy, (4) Offering a Candidate, (5) Repetition, and (6) Non-Verbal/'Freeze Look'. Table 29 below presents each type with an example.

Table 29: Types of CRs found in first encounters asking for directions.

CR types		Example
Explicit	Ellipsis (Restricted)	Hitt húsið? (Hitt húsið) Hvar? (Where?)Hvaða hús? (Which house?)
	Full or Partial Explicit Queries (Open)	Fyrirgefðu, hvar er? (Excuse me, where is?) Hvað segirðu? (What are you saying?) Hvað, Hitt húsið? (What, Hitt húsið?)
	Fragment/Interjection Strategy (Open)	Ha? (Huh?)
	Offering a Candidate (Restricted)	Postulinn, þú meinar? (The apostle, you mean?) Þú meinar Kiki? (You mean Kiki?), Við Hringbraut? (By Hringbraut?) Svona fjölþjóða eitthvað (Something like the multinational?)
	Repetitions (Restricted)	Hvar er Hitt húsið? (Where is Hitt húsið?), Hvar Hitt húsið er? (Where Hitt húsið is?) Hitt húsið er? (Hitt húsið is?)

Implicit	Non-Verbal 'Freeze Look' (Open)	(Without speech; interplay of gaze, mouth, head, handedness, torso, body posture, legs)

2. No major differences were found in the multimodal production of CRs between gender and NS and NNS speaker pairs of the focus and control group pairs in this study. This means that neither gender nor the native origin of speakers played a significant role in this. A difference was observed in the frequency of use of CRs. The study showed that CRs are much more frequently produced in the focus group of NS-NNS speaker pairs (69.44%) than in the control group of NS-NS speaker pairs (19.30%). This leads to the conclusion that compared to NS-NS speaker pairs, more misunderstandings or more information for clarifying seem to occur in NS-NNS speaker pairs than in NS-NS speaker pairs. Gender did not seem to play role in this difference. A language barrier that lies between NS-NNS could have been the reason for a more frequent occurrence of certain CRs.
3. Based on the analysis, the following multimodal features (Table 30) were suggested to be included into the multimodal CR model, which can be implemented into the conversational architecture of ECAs in order to simulate natural conversation of Icelandic *in silico*.

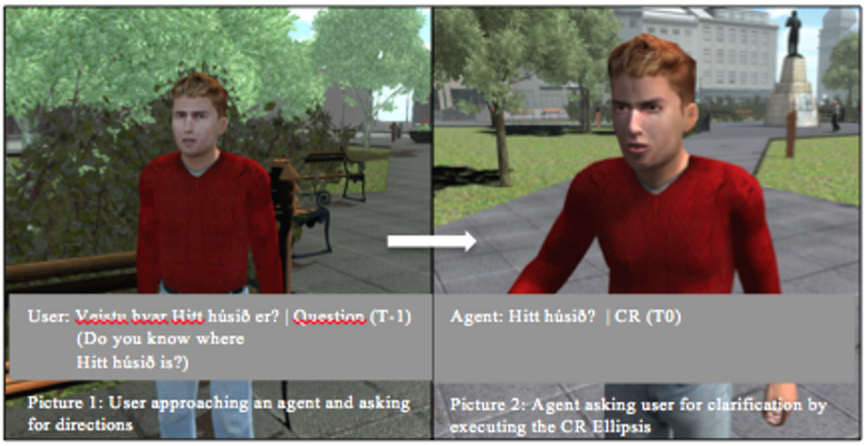
Table 30: Multimodal features for a CR.

Multimodal features	
Time duration of the utterance	
Linguistic form of a CR utterance (spoken word(s) or nonverbal 'Freeze Look'	
Paralinguistic features	Intonation
Body features	Head movement
	Forehead movement
	Eyebrows
	Eyes (Gaze)
	Mouth/Lip movement

	Handedness (hand movements or hand gestures)
	Fingers (movement of visible fingers)
	Body posture
	Torso
	Legs
	Foot
Proximity to user’s avatar	

Some of the types defined by this study correspond in most part with the results of the previous study by Gísladóttir (2015). However, the frequency of different CR types is very different. This may be because the previous study examined CRs in a different context and a conversational scenario, i.e., at a dinner party among friends and family members. The present study, however, examined CRs in unknown first encounters asking for directions between both native and non-native Icelandic speakers. This study moreover presents a new type of CR, the non-verbal ‘Freeze Look’. The multimodal analysis presented in this study showed which verbal and non-verbal features are associated with each CR type, which none of the previous research listed in such detail. Regarding the frequency of CRs, the most frequent type was the Ellipsis 80 (79.21%) instances in the whole corpus. Other CR types were not very frequent, but each of them is very important to note and to describe. They typically occurred unexpectedly in the course of dialogue between unknown speakers and therefore should be considered as a valid instance, which is part of the language that native Icelandic speakers use. These were Full or Partial Explicit Query8 (7.92%) instances, Offering a Candidate 5 (4.95%) instances, Fragment/Interjection Strategy 4 (3.96%) instances, repetition 2 (1.98%) instances, and the non-verbal ‘freeze-look’ with 2 (1.98%) instances. Even though more instances would be needed to compare the multimodal realisation in between speaker pairs, this small number is sufficient to describe their multimodal features. Based on the most frequent CR type Ellipsis, the multimodal realisation is very similar, if not the same in most instances, which makes it more likely that other CR types would be also similar in realisation. The following chapter describes the application of the results presented in this study. It uses six CR types to create six multimodal CR models that can be suggested for implementation into the ECAs conversational behaviour in *Virtual Reykjavik* in order to both keep a smooth flow of a conversation and teach learners how native speakers ask for clarification in Icelandic.

The linguistic means and intonation together with the movement of main body parts from head to foot, including facial features, and proximity (i.e. distance between speakers/agent and learner's avatar) were described based on selected examples from the video corpus. These examples were selected because of their optimal quality, which shows the whole-body posture with relatively clear facial expressions. The proximity between the native speaker and the actor was on average 1 m or less in real life, which was interpreted as a close distance. From observation in the field while collecting data and from the video corpus, one could not detect differences in distance based on native origin or gender. However, further research may focus on this aspect and investigate this problem more closely. Based on the description of multimodal features in CRs, a model for implementation into the ECAs conversational behaviour was suggested. This included two multimodal CRs, the Ellipsis and the Fragment. Figure 27 shows an ECA executing the multimodal CR Fragment/Interjection. Note that compared to the CR sequence demonstrated above, the model below includes only the first turn T-1 (question) and the second turn T0 (CR). The third turn T+1 (answer) was omitted.



User: Veistu hvar Hitt húsið er? | Question (T-1)
(Do you know where
Hitt húsið is?)

Picture 1: User approaching an agent and asking for directions

Agent: Hitt húsið? | CR (T0)

Picture 2: Agent asking user for clarification by executing the CR Ellipsis

Figure 27: An ECA executing the CR Ellipsis.

The suggested models for multimodal CR shed light on the features to implement when modelling the communicative function of CR, in order to achieve a more realistic interaction between agents and human learners in VR. The ECAs were endowed with the multimodal feature of the two CRs so that they correspond with the proposed two models, the CR Ellipsis and the CR Fragment. It was a manual task during which a programmer implemented the verbal and non-verbal features into Unity. Verbal features were relatively easy to implement as they were directly programmed in respect of the choice of words and the intonation. The non-verbal features, however, were more difficult to implement. Different body movements were part of the layout in Unity, which is a cross-platform game engine developed by the company Unity Technologies. However, what made it difficult was to programme the correct angle of movement. For this reason, several sessions took place during which the multimodal behaviour was implemented, then observed on the agent, and fine-tuned. For instance, the opening of the mouth and eyebrow movement was particularly challenging to programme. Detailed and nuanced movements were implemented but, on the screen, were not clearly detectible by eye. After additional adjustments, these became clearer on the screen, but at the same time looked exaggerated. It was difficult to optimise the movement and more sessions needed to be taped to adjust these particular movements.

The pilot user response study described in the previous section indicated learners' perceptions from playing *Virtual Reykjavik*. It moreover informed about how learners perceived ECAs and shed light what learners thought about some of the CRs used in the game, especially the Interjection Strategy. Besides the learners' general perception of the game and the multimodal behaviour of agents, and the learning effect, it indirectly investigated the effect the CR strategies had on learners. The use of the CR function triggered learners repeating what they have just said if the ECAs did not understand their spoken input, consequently leading by some participants to a side-effect of learning the proper pronunciation of particular sounds. This suggests that this function does contribute to maintaining a relatively smooth flow of a conversation between agents and learners (avatars), and also contributes to building a common ground in human-agent interaction. Regarding the use of authentic features in the model, with which the ECAs were endowed, the learners noticed only the Interjection Strategy “Ha?”, especially in the spoken form, and perceived it as life-like, which is according to some of them similar to what Icelanders do frequently. Regarding authentic features of facial expressions in the CR model, some learners perceived a pained expression on the agent's face when executing the Interjection Strategy. There was, however, no notice about the Ellipsis strategy. Regarding the body movement and the hand gestures during these two CRs, the learners did not notice anything particular, which on the other hand is in concordance with the multimodal behaviour implemented in the CR models including, if at all, very nuanced body movements of the agent's body towards the learner's avatar and a slight movement forward of the agent's torso. Nonetheless, the pilot study suggests that further fine tuning is necessary to eliminate the 'pained expression' and make it more realistic. Also, further testing only for the perception of CR models is required to gain more information about the perception of its design. The following chapter discusses the results in the context of the theoretical background.

4 Discussion

4.1 Introduction

This chapter is devoted to the discussion of results, and the theoretical and practical implications of this study. It discusses the link between research in real life and the application of the study findings in the context of modelling realistic human-agent interaction for learning a language and culture in *Virtual Reykjavik*. Here, both the answer to the overarching questions guiding this three-phased project presented in this thesis (section 1.3) and a discussion about theories, concepts and approaches used in the theoretical background (section 2) in the context of this thesis will be provided. The ECAs in the computer game are endowed with multimodal features found in two out of six CRs; they interact with learners and use this communicative function towards maintaining conversational flow, as well as to teach different ways how native Icelandic speakers ask for clarification.

4.2 Realistic Interactions in *Virtual Reykjavik*

This section discusses the efforts towards achieving realistic interactions in *Virtual Reykjavik*. In the process of improving the ICALL effort in Iceland, the general goal of the *Virtual Reykjavik* project was to create a simulation of a real environment in a 3D computer game, in which learners can practise on an individual basis spoken language skills in various communicative exercises, by solving tasks and thus learn Icelandic language and culture (section 1.2). For this reason, *Virtual Reykjavik* is a tool that allows learners to practise spoken language skills and keep their focus on the target language without switching into another language, usually English. By playing *Virtual Reykjavik*, learners can speak to the ECAs endowed with authentic features for natural language and thus contextualise the communicative task. The learners can actively use various sensory channels while playing the game and practise speaking (pronunciation, vocabulary, grammar) and other language skills, such as reading, or writing (typing) that are related to each communicative task (Morton et al., 2012, p. 2). *Virtual Reykjavik* employs a multimodal approach to teaching communicative language skills. This approach involves multimodal cues from gesture, gaze, facial expression, shifting of the body posture, all of which also occur in a face-to-face interaction. These multimodal cues may be specific to a particular social or cultural context and the resources available to people at the time of making meaning, i.e. producing and perceiving language (Jewitt, 2013a, p. 2). *Virtual Reykjavik* is a prototype of an educational tool that supports a mix of images, colours, texture, written and spoken language, through which different senses are employed in the process of learning.

Creating realistic interactions between ECAs and learners is a complex process because it involves planning the dialogue which will unfold in the communicative tasks the learners will engage in when playing the game (section 2.3). Complexity Theory was used to help understand the three complexities that co-occur therein: (1) the process of conveying the information, (2) the process of understanding the information, and (3) the process of those two to find a common ground (Jörg, 2011, p. 208). These suggest that both learners and agents need to have a prescribed role in order to find a mutual ground between two realities - one virtual and the other real. Since speaking involves conveying information, the ECAs were endowed with multimodal features to enable them to convey more precise and comprehensible information and appear more 'natural' to learners. The CR function was used throughout the course of the interaction between the agents and learners and represented the function of requesting clarification each time the agents did not understand the learner's spoken input (section 3.5). In this way, the agents could react realistically to learners' mispronunciation - their speaking too softly to the microphone, or other technical obstacles, such as the internet connection delays (causing an incomplete learners' spoken input), and thus keep a natural flow of a conversation, as is often the case in real life when miscommunication occurs. As it is very difficult due to time demand, data collection and multimodal analysis (transcription of speech, segmenting utterances and tagging with data from a coding scheme) to investigate all utterances in a conversational scenario such as this one (asking for directions), the focus to multimodally analyse an utterance with a practical communicative function was on the CR. Whether the thesis provides answer to the following overarching question "What are learners' general experiences with playing *Virtual Reykjavik*, in which the ECAs use the suggested multimodal CR models in first encounters?" is subject to the micro view provided by results from the preliminary pilot user study presented earlier this thesis (section 3.5.4.4). These results moreover inform about the participants' noticing the two types of CR functions performed by ECAs. The multimodal behaviour of ECAs in relation to the CR Interjection Strategy model ("Ha?") with falling intonation was perceived as if the agents had a 'pained' expression when they misunderstood the learner. The learners perceived it as life-like because it is what Icelanders would often say in real life, i.e. whatever you say to them they all the time response with “Ha?”. The other CR type Ellipsis model (“Hitt húsið”?) was not particularly mentioned by the participants. The following section discusses six multimodal CR strategies that were analysed and described in detail (in section 3.3).

4.3 Six Multimodal CR Strategies

The basic and primary use of language is the one used in a face-to-face conversation (Fillmore, 1981 in Clark, 1996, p. 8). By identifying the features that help humans convey meaning in face-to-face interaction, such features can be used to build a realistic multimodal model of a specific communicative function. In this thesis, the CR function was investigated in a similar situation as learners would encounter in real life, i.e. non-native speakers (Icelandic learners) encountering native Icelandic speakers. Communicative functions can consist of either a sound or a word/words, or even of no words at all, e.g. a listener signals non-verbally to the speaker that something has not been understood and is doing so only by non-verbal cues incorporated into facial expressions, head, torso and body movements. In this view, the CR study shed further light on language as a multimodal phenomenon (O’Connell et al., 1990; Lakoff and Johnson, 1999; Barsalou et al., 2003; Vigliocco et al., 2014; Skipper, 2014; Jacobsen, 2015). The answer to the next overarching question “What multimodal features in CRs are needed in order to design a more realistic human-agent interaction in *Virtual Reykjavik*” is based on results presented in the CR study (section 3.3.3). Briefly, based on natural-language research, six multimodal CR types were found in first encounters. Their detailed description is presented in six multimodal CR models (section 3.4).

Concerning the study on multimodal CRs, only two of the six identified multimodal CR strategies (section 3.4) that locals use in a conversation with others when asking for directions were implemented into the multimodal behaviour of agents in this project. This is due to resource constraints. The results from the pilot study (section 3.5) showed that one CR model, the Interjection Strategy “Ha?”, was particularly noticed by some learners and was described as life-like (section 3.5.4.4). However, from the ‘pained’ facial expression on the agent’s face that the learners reported, and according to their observations, informed the design team that it requires additional fine-tuning. The agent’s natural behaviour nevertheless showed learners how Icelanders would ask for a clarification, and thus demonstrated an authentic cultural feature of Icelandic peoples. As in other computer games that support realistic interaction, learners spoke to the agents by using a speech recognition programme, and thus contextualise a particular communicative task while practising the target language (Morton et al., 2012, p. 2). Despite the realistic execution of the CRs, based on results from the preliminary pilot study some learners suggested instead a different phrase to be used, i.e. in English “Sorry, I don’t hear you”, which may add more politeness into the agents’ behaviour. This phrase would have been artificially invented and not in line with what native speakers typically might say in first encounters asking for directions. The preliminary pilot study was also based upon the theoretical IMTEE model (Alvarez et al., 2004, p. 391) (section 2.4.6) and the results informed about the concept of fidelity and feedback (section 2.3.5) the learners experienced during the pilot testing of playing the game. The results about experiences with social interaction, visual and auditory sensors in the *Virtual Reykjavik* system can help with improving effective training, feeling of presence, and engagement of learners into interacting with ECAs and learning Icelandic in the future (Lane et al., 2013) and new features can contribute a situation of an intensive real-life scenario that learners can enter and practise their language skills in a save interim space (Grant et al., 2010). This finding also contributes to answer the overarching question about general learner experiences with playing the game.

The purpose of the main study in this thesis was to demonstrate how research in CR strategies can provide authentic data from natural language research to help create a more realistic human-agent interaction in *Virtual Reykjavik*. Multimodal features included in different communicative functions that ECAs use in interaction with L2 learners can improve the understanding and development language and culture skills. This can be done by endowing ECAs with specific multimodal behaviour in these communicative functions and letting learners observe these. It may help learners to adopt a similar behaviour and simulate it in real life. The main study here supports the effort to achieve a realistic human-agent interaction by letting agents request a clarification in a similar way as would native speakers of Icelandic in reality and therefore helps again to answer the second of the overarching questions about what multimodal features in CRs are needed in order to design a more realistic human-agent interaction in *Virtual Reykjavik*. The six multimodal models of CRs not only helped to inform which features should be included into the multimodal behaviour of ECAs, but also helped to understand how language fits into a multimodal phenomenon that includes non-verbal ‘utterances’, i.e. instead of spoken words, humans can request a clarification by the sole use of body language. One of the six suggested models is the implicit non-verbal CR strategy, the 'Freeze Look'. Similar to real people, the agent can also look with a 'freeze look' by staring at the learner's avatar (first-person view) with the mouth closed in the beginning and then having it slightly open, head directed at the learner, hands not moving, but the torso leans slightly forward towards the learner's avatar, legs and the rest of the body posture is unchanged. The speaker's opening his/her mouth signals an activity of communicating, though without any sound, and as there is no sound coming out, it then triggers taking a turn by the other speaker (actor) who understands it as a clarification and provides an answer accordingly. This would be a work for further implementation and an additional pilot study. The following section discusses the multimodal communication and multimodal CRs.

4.4 Multimodal Communication - Multimodal Clarification Requests

The study on multimodal CR strategies reconfirmed that the exchange of messages in a face-to-face conversation between participants does not only depend on the spoken word, what it symbolises and how it appeals to the listener, but also on other cues of non-verbal behaviour that consists of various body movements. Additional cues include the emotional and social status of the speakers, their knowledge of the world, and the context in which language is produced. In this view, language is a mutual exchange of multimodal cues that allow the producer of a message to express his/her idea in a multimodal way. The recipient can also interpret it through the use of embodied cognition of various senses (hearing, seeing, touching, etc.) that he/she had previously experienced in their life, and also through his/her own knowledge of the world (customs, society, politics, education, practical experience, etc.). Speakers use various multimodal cues to reach a common ground. These multimodal cues are all constituents of language as a multimodal phenomenon. The study on multimodal CRs revealed six different types of CR strategies, one of which was the Non-Verbal 'Freeze Look' also described in another language, the Argentine Sign Language (LSA) (Manrique, 2016), which has a similar execution of non-verbal features. Regarding the other CR strategies, this study shed light on the frequency of CR use in first (unknown) encounters and provided a multimodal description of each type, which had previously not been done. This provides useful data for further research, as well as information about how to endow ECAs in other applications to build realistic human-agent interactions. There is no significant difference in CR production between gender or whether native Icelandic speakers talk to other native or non-native speakers. In this regard, it is considered to be universal. This view helps to answer the second of the overarching questions about what multimodal features in CRs are needed to design a more realistic human-agent interaction in that it suggests that such multimodal features could be universally applied across languages, with perhaps certain differences regarding intonation, e.g. compare the Interjection Strategy *Ha? (Huh?)* with falling intonation in Icelandic (this thesis; Gísladóttir, 2015) with *Huh?* usually with rising intonation in English (Enfield et al., 2013, p. 12). Due to the theoretical gap in current literature, this model was developed for the use of this thesis but can be expanded by further scholars and used in other research supporting their findings in the area of language and multimodal communication.

Multimodal communication deals with different signals that may be produced at once by different parts of the body (Pelachaud and Poggi, 2002), in which the vast complexity of different modalities (eye gaze, facial expression, hand gestures, body posture, etc.) participate in the production and perception of meaning in a face-to-face conversation (Fägersten et al., 2010; Norris, 2013). In the context of this thesis, multimodal communication represents an approach to understand spoken natural language and to help characterise multimodal features that occur in particular communicative functions. Based on findings in this thesis, specific multimodal cues have been used to model a realistic conversational behaviour of ECAs when executing the CRs. In this way the agents represent an embodied version of a natural language dialogue system (Nijholt and Heylen, 2002) that interacts with the human user (learners of Icelandic). With the inclusion of multimodal features, the system interacts in a more realistic way, thus contributing to the fidelity. The notion of fidelity is used here to understand various kinds of performance in a simulation of a real-life conversation. This notion moreover guides this thesis to understand the basis for experiencing social, visual and auditory sensors needed for creating a simulation of a real-life environment for effective training, feeling of presence, and engagement of learners (Lane et al., 2013, p. 4).

Research in natural language shed light on the use of CRs in multimodal communication. These CR utterances had previously been studied from the point of view of a linguistic form, use and function in human communication (Duncan and Niederehe, 1974; Schegloff, 1992; Traum and Allen, 1992; Saxton et al., 2005; Purver, 2004; Ellis and Yuan, 2005; Witton-Davies, 2010; Gísladóttir, 2015; Lowder and Ferreira, 2016; Mihas, 2017). This thesis, however, used the categorisation of linguistic forms from other bodies of research (Schegloff et al., 1977; Purver, 2004; Cho, 2007; Gísladóttir, 2015; Manrique, 2016), which altogether listed about twenty-six different categories, and developed a more unified categorical overview of six CR types found in the present study. A new multimodal CR category in a spoken language is the implicit Non-Verbal 'Freeze Look' type, which previous research in spoken language communication face-to-face did not define. This type has also been found in Argentine Sign Language and suggests the same realisation of multimodal features. However, these features were mentioned only generally to describe some kind of a 'freeze look' from which the name has been adopted. This finding nonetheless indicates some kind of a general use of this 'freeze look' with the function of a CR across languages. It would be interesting to find out whether and in which context such a category would occur in a different spoken or sign language. Based on a common pattern for each category and no difference in execution due to gender or native origin, the multimodal CR models can be universally used in modelling a realistic human-agent interaction. If only the linguistic part with verbal features and the non-verbal part in CRs is omitted, then an incomplete picture would form the communicative function CR. With only the verbal part, the ECAs would speak in a robotic way, which is not natural for the human user to see. Humans are familiar with both verbal and non-verbal, i.e. multimodal, behaviour in communicative functions that can be found in natural language interactions with other humans. They expected the same or similar to see when listening to virtual agents speaking. In a telephone conversation, one can detect a CR strategy without multimodal behaviour, because this means of communication can transmit audio alone. But once humans are involved in a face-to-face interaction, appropriate multimodal behaviour for each communicative function in a conversation is expected. Moreover, non-verbal features in communicative functions may vary in some cultures. For this reason, it is important to teach learners of Icelandic how the CRs are produced by the native Icelandic speakers in order to prepare them for real life. The following section discusses the multimodal behaviour of ECAs.

4.5 Multimodal Behaviour of Embodied Conversational Agents in CRs

The findings from the study on multimodal CRs were used to endow the ECAs with authentic multimodal behaviour in the communicative function of CR. As Poggi et al. (2005) suggest, the combination of various multimodal communicative signals, such as words, prosody, gesture, face, body posture and movements that are displayed by ECAs, are determined by different aspects, such as (a) contents to communicate, (b) emotions, (c) personality, (d) culture, (e) style, and (f) context. These aspects also determine what virtual characters will say, and how (Poggi et al., 2005, p. 3). Therefore, based on this particular conversational scenario, six multimodal CR models were suggested for implementation, which contributed to the known repertoire of communicative functions, especially for Icelandic. Even though these six models may be bound to the specific conversational context, there might be other CR types in a different conversational scenario. These different types may also have a different frequency of use. For instance, when the study on multimodal CRs in this thesis is compared to the study by Gísladóttir (2015), there is clear difference in frequency of use in CR Interjection Strategy *Ha? (Huh?)*, probably because of a different conversational setting. Similar CRs were found in both studies, the present study and in the study by Gísladóttir (2015), however, the non-verbal CR was found only in the present study. The multimodal CR models can also be used to endow ECAs speaking other languages, especially the non-verbal type, which seems to be similar in the Argentine Sign Language (Manrique, 2016), for instance. The models will be part of a turn-taking process which will take place when agents will for whatever reason do not understand the learners' input. This process may, however, be more difficult in human-agent interaction, because the agent and the learner do not possess the same ability to perceive the other participant's intention to take turn. For instance, the agent must rely on the spoken input or another command performed by the user, whereas the user has more opportunities to detect whether the agent is yielding a turn. However, agents can use various multimodal features to perform a CR function for requesting a turn. Concerning the interjection type, one of the features seems specific for the Icelandic language only, i.e. the falling intonation (Gísladóttir, 2015). A practical demonstration of two types of multimodal CRs, the Ellipsis and Interjection Strategy, were done in the pilot testing of *Virtual Reykjavik*. In this, the agents used these two functions in order to maintain a smooth flow of a conversation when they did not understand the learner's spoken input. The ECAs can now benefit from the suggested multimodal CRs models because they are available for them to be used in further development of *Virtual Reykjavik*. These features contributed to the improvement of a realistic human-agent interaction. In addition to this, playing the game can have an educational potential, which will be discussed in the following section.

4.6 The Learning Effect

Virtual Reykjavik belongs to the category of serious games that, according to Johnson and Wu (2008), represents a platform that helps learners quickly acquire knowledge of foreign language and culture through a combination of narrative lessons that focus on particular skills. It has the ability to provide learners with a specific learning environment that, according to Meyer (2009), can contextualise knowledge and immerse the learner into an environment outside of a formal language course. It can also the immersive properties of a narrative story that lead the learner through the game scenario, and that can promote the perception that the story is real and live, thus "helping to break down barriers between virtual reality and user" (Shin, 2018, p. 69). This thesis used the Flow Concept (Nakamura and Csikszentmihalyi, 2002, p. 89) to understand the nature and conditions of immersion and enjoyment by learners pursuing playing *Virtual Reykjavik* for practising spoken language. As the findings from the preliminary survey on learners' expectations from *Virtual Reykjavik* (section 3.2) and the findings from the pilot user study of *Virtual Reykjavik* (section 3.5.4.2) reveal, the learners expect the game expect the game to have a good storyline with particular conversational scenarios for speaking practice, a voice recognition in order to be able to communicate with virtual agents that would assume the role of native speakers and interact with others in the game. They moreover expect the agents to be funny and entertaining, perhaps be able to tell a joke, and to give feedback on grammar and the learners' choice of vocabulary, and to learn practical information about the city. At the same time, they experience the game as educational, enjoyable and exciting, with a great potential to bridge the gap between traditional Icelandic courses and the use of language in a real environment. Moreover, the ECAs in *Virtual Reykjavik* are part of a multimodal interface and use multimodal behaviour (eye gaze, facial expressions, gesture, body posture) when speaking Icelandic to learners. These findings help partially answer two of the overarching questions about what general expectations do L2 learners of Icelandic have from a 3D game for learning Icelandic with virtual characters, and what are learners' general experiences with playing *Virtual Reykjavik*, in which the ECAs use the suggested multimodal CR models in first encounters. For more consultation, detailed answers to the questions from the preliminary survey are listed in section 3.2.4 and to the pilot user study in section 3.5.4. As the learning focus in *Virtual Reykjavik* is on oral communicative skills, ECAs endowed with authentic multimodal features in (not only) CRs can contribute to a more authentic learning experience. Since multimodal features of the spoken communication system between real speakers are utilised frequently in real life, these multimodal features can be utilised between L2 learners and agents in the game, to achieve a mutual understanding when fulfilling the communicative tasks and to foster an authentic linguistic development (Wild, 2015, p. 50). Such linguistic development can be observed in particular on the use of clarification strategies by ECAs that learners may observe and adapt later in their communicative tasks in the game. This may prepare them to both understand and use similar clarification strategies when speaking to native Icelandic speakers in reality. Through various activities in the game, learners communicate with other agents by which they learn both to use the language adaptively in a game-solving situation (Zheng et al., 2012, p. 358; Lombardi, 2012, p. 49; Mayer, 2019, p. 533), and to facilitate a face-to-face interaction with speakers of a particular culture (Bédi et al., 2016, p. 42; Ayedoun et al., 2019, p. 30).

The effort of creating a realistic interaction in *Virtual Reykjavik* had a positive effect on developing spoken language skills by surveyed learners. Findings from the auxiliary pilot user study reveal that the game encouraged some learners to start talking in Icelandic. It was not only how the game was methodologically constructed, for instance solving communicative tasks and using Icelandic throughout the whole process of playing the game, but also the inclusion of a realistic multimodal behaviour into the CR function that ECAs used in interaction with learners. This had a side-effect of making learners repeat certain words in case of mispronunciation, and in this way, learners could adjust their pronunciation so that the speech recognition system would recognise the word. This process is not always available when using spoken language in real life. In connection with the realistic interaction and the multimodal behaviour of ECAs in the game, some surveyed learners observed in particular the manner of ECAs when executing the CR Interjection Strategy. They reported that this is a culturally bound spoken behaviour that Icelanders often use in real life when they do not understand what has been said. These findings furthermore show that the game had an effect on learning about culture. As the learners could click on the 'ESC' button to go to a freeze mode and explore the environment by reading text about buildings, the learners reported to have learned about certain cultural sites in central Reykjavik. These findings therefore contribute to answer the third of the overarching questions about learners' expectations. The answer to particular questions guiding the pilot user study are presented in detail in section 3.5.4.

The purpose of the game was to teach Icelandic language and culture with agents that know how to use the language. Two CR strategies with associated multimodal behaviour were implemented to serve this purpose in a pilot user study. The reported learning effect from the pilot study demonstrates that the game has an educational potential, especially for teaching speaking and focusing on practising interactions in Icelandic. This is in line with learners' needs and thus helps to answer the overarching question about learners' expectations. Learners expect the game to have an educational purpose, teach them vocabulary, listening, grammar, speaking, cultural understanding, writing, and reading in this preferred order as the auxiliary survey reveals. The pilot user study (section 3.5.4), however, reveals a different order in learning experience: learning about the culture and to start speaking. Only when the game is improved a more rigorous study can be conducted to compare learners' needs with their experiences form learning. The following section discusses personal remarks about the aims and findings in this thesis.

4.7 Concluding Remarks

This thesis expanded the view of teaching Icelandic as L2 in Iceland in that it provided insights into the problems learners face when learning this language and living in Iceland. What the field of Icelandic L2 studies only assumed, the findings in this thesis show that speaking practice is still a problematic area in Iceland. On the one hand, the findings revealed that most learners do feel intimidated and stressed out about speaking Icelandic face-to-face with native speakers, but on the other hand it seems that by being helpful, native speakers often switch into English to make the communication easier. From the point of view of a language learner, this is a cycle of seldom achieving enough exposure to, and practice in, spoken Icelandic. In connection to this, the CALL effort for developing the Icelandic language and culture training application *Virtual Reykjavik* makes it more relevant to focus on training speaking skills and thus provide learners with more opportunities to practise speaking with virtual characters that know how to use the language.

There are several language learning applications yet only few of them are for training speaking skills. Existing language learning applications such as *The Tactical Iraqi Language and Culture Training System for Arabic*, the *Danish Simulator* and *ELSA* enable speaking practice with the support of speech recognition. These games remain limited, however. *Virtual Reykjavik* has scope for development into a fully functional application enabling learners to gain more practice in spoken Icelandic and boost language and cultural skills. Contemporary research and technical advancements will make it easier to build such applications with an in-built speech recognition and text-to-speech programmes for a particular language to practise speaking. These are necessary components in language learning that supports Swain's (2007) 'Output Hypothesis' that language output in a spoken form helps to connect the theoretical input about vocabulary, pronunciation and grammar, and thus plays a very important role in the learning process of a language learner.

However, further research needs to be conducted to measure the impact of the agents' multimodal behaviour on the learning part of *Virtual Reykjavik* players.

Virtual Reykjavik builds on research in natural language in that it uses data from face-to-face interaction between real speakers of Icelandic to help create lessons, game scenarios, conversational situations, and to design the verbal and non-verbal behaviour of conversational agents towards resembling native Icelandic speakers. The second study on multimodal CRs was part of collecting multimodal data by the means of video recordings for the *Virtual Reykjavik* corpus. This created the basis for not only designing a realistic behaviour of ECAs in the CR function, but also in another communicative function, the EAP (Ólafsson et al., 2015), which contributed to the understanding of the dialogue structure and the use of language in first unknown encounters. In the CR communicative function, this study helped to define six types of CRs native Icelandic speaker utilise when talking to others during first encounters to ask for directions. One of the types was the Fragment/Interjection Strategy *Ha? (Huh?)*. Gísladóttir (2015) defined the Fragment/Interjection Strategy as the most frequent clarification strategy used amongst friends and acquaintances in Icelandic, but which the current research ranked as the fourth most frequent CR used amongst unknown encounters. Despite the lower frequency, this Interjection Strategy is a very good representation of a culturally bound use of language. It is often used in any kind of a situation in real life and has a falling intonation, to which to a visitor to Iceland could be perceive as slightly rude. Unlike in other languages, it is the intonation contour in Icelandic, namely falling pitch, which decides for trouble-presenting repetition (Gísladóttir, 2015, p. 321). It was for this reason that it was included into the conversational behaviour of ECAs in *Virtual Reykjavik*. Due to resource constraints, only one other multimodal CR was implemented, the Ellipsis *Hitt húsið?*. This Ellipsis CR was the mostly used strategy among unknown encounters in the current study.

From the perspective of L2 learners of Icelandic, it is very important to become familiar with the way how language is used in reality, because it can minimise misunderstandings coming from the lack of knowledge about the use of language in the target culture. The CR Interjection Strategy *Ha? (Huh?)* is one such case. It is moreover important to increase opportunities of the target language exposure and the practice of language skills, especially connected to speaking and listening, because that seems to be one of the hardest parts of developing communicative language skills not only in Icelandic (Walker, 2014; Ulum, 2015; Leong and Ahmadi, 2017; Lowie et al., 2018). Learners may benefit from an interim space in which they can safely practise language online. 'Safe' in this context means that learners would not need to feel intimidated by speaking to virtual agents, and for this reason could without hesitation practise the target language with mistakes. According to Brown (2007), for L2 learning, a share of classroom-oriented language is context reduced, while a share of language in a face-to-face communication with people of the target language is context embedded (p. 196). It is not only the spoken communicative competence that learners may develop when practising language face-to-face, but also the more general pragmatic competence of language (Bachman, 1995) that gives learners a sensitivity to 'naturalness', dialect or a variety of register, the understanding of cultural references and figures, and other functions of language (p. 87). It is therefore very important for L2 learners of Icelandic, especially for those living in the country who wish to integrate into society, to be exposed to face-to-face communication in the language. *Virtual Reykjavik* could become a tool that L2 learners of Icelandic can use remotely, enter an interim learning space in virtuality and practise spoken Icelandic face-to-monitor with virtual characters that speak. Learners can improve specific cognitive skills if they repeatedly practise them in the game (Mayer, 2014, p. 172). These cognitive skills are those ones that are most likely to be transferred to non-game contexts (Mayer, 2019, p. 541), which in this case is comparing and contrasting with the learners' language and culture, and reflecting on the content based on their personal experience in other areas of studies or personal experience (Yoshida, 2010, p. 1). It is therefore imperative to do more research in natural language in order to design the communicative behaviour of agents in the game so as to resemble native speakers. This would be expected to contribute to the development of learners' pragmatic competence of language and thus make the use of spoken Icelandic more available and accessible online at individual basis. Learners can train themselves in (not only) speaking while providing a safe place for language learning, in which they use their cognitive skills to transfer new information into their linguistic and cultural repertoire. As Mayer (2014) furthermore suggests, the improvement of cognitive skills is mostly associated with first-person 'shooter games' that according to that study, can improve perceptual attention by learners (p. 217). Further research needs to be conducted to test this on *Virtual Reykjavik*.

The thesis showed that the effort to achieve a realistic interaction between agents and learners in *Virtual Reykjavik* was in part successful. The findings from the pilot study reveal that implemented multimodal behaviour in CR functions, especially the language part in the Interjection Strategy, showed that learners found it believable, or life-like, when agents executed it. Despite the fact that there is no evidence about the Ellipsis CR, one may in this case consider no specific evidence as evidence in that it aided in promoting the smooth flow of a conversation between agents and learners. Further investigation is needed however to validate this. Similarly, further implementation of the other four CR models into the conversational architecture of ECAs is needed, as well as additional pilot testing including all CR models. Even though some technical issues occurred and the ECAs required fine-tuning to some of their multimodal behaviour, especially in the CR functions, which consequently by some learners caused frustration and boredom and made the agents appear cold and robotic, most of the surveyed learners perceived the game as enjoyable and fun to play, but also educational and straightforward. For most of the learners, the game was useful because they said that they learned something, i.e. it helped them to start speaking in Icelandic and they learned about iconic buildings in central Reykjavik. This corresponds with the aim of *Virtual Reykjavik*. Full development may result in a very successful Icelandic language and culture teaching tool. But no matter how functional and advanced the technology and believability of ECAs' design will be in the future, one may apply Shine's (2018) view that "the key is to focus on the story, not the technology itself or any special 3D effects. The real challenge is not so much that things can look too real or not real enough; instead, it involves the feel of the piece, as perceived by the users of VR stories" (p. 72). Further development and user studies need to be conducted in order bring *Virtual Reykjavik* to an optimal state for learning Icelandic language and culture.

4.8 Conclusion

This chapter demonstrated how research in real language can contribute to the improvement of human-agent interaction and provide an interim learning space that can be used to develop L2 skills, especially speaking and pronunciation which have proved difficult for online language learning. The partial building blocks towards achieving a realistic interaction in *Virtual Reykjavik* were discussed, i.e. data from natural language that was used for designing multimodal behaviour and speaking input of ECAs, the metaphorical understanding of life as a stage and humans as actors in it, which brought about the theoretical foundation for another dimension of a dialogue in which scenarios and the roles of both virtual characters and learners are combined in two realities, the real and the virtual. The learning effect of playing *Virtual Reykjavik* was discussed in connection with realistic interaction with the agents and their multimodal behaviour. The following chapter concludes this thesis.

5 Conclusion

5.1 Introduction

This interdisciplinary thesis describes how verbal and non-verbal features in CRs in real-life interactions were collected and analysed in order to create theoretical models of multimodal CRs. This thesis is part of a larger project, whose team is developing a 3D computer game for learning Icelandic language and culture *Virtual Reykjavik*. The general goal of the larger project is to create a serious computer game in a 3D virtual learning environment populated with virtual characters who can speak and act like real people and help learners of Icelandic practise spoken language skills. The specific goal of this thesis was to conduct a rigorous study on multimodal CRs in real-life interaction and, based on the results, deliver six theoretical models for multimodal CRs that can be implemented into the conversational behaviour of ECAs when performing the CR function. Apart from this main study, also two auxiliary studies were conducted. These two shorter studies contributed to the general goal of the larger project and helped to situate this study into the CALL effort for Icelandic *via* the application *Virtual Reykjavik*. A serious game aimed to promote Icelandic language and culture learning in a context-specific virtual learning environment, to allow learners to practise spoken language skills and interact with conversational agents that use the language.

A short survey addressed the importance of practising various language skills including vocabulary and grammar, as well as the importance of practising spoken Icelandic without switching into English.

The main focused on research in natural language, which helped to endow ECAs with multimodal features for executing the communicative function of a CR. This is important towards keeping a natural flow of conversation and contributes to a realistic human-agent interaction in *Virtual Reykjavik*. The material collected in this study helped to improve the multimodal behaviour of ECAs and the learning content of the game. The transcribed dialogues created the basis for designing dialogues in each conversational scenario, which resulted in a teaching lesson on a specific topic, which was in this case first encounters asking for directions. Instead of using material from language textbooks, the ‘real language’ from the *Virtual Reykjavik* corpus was used to design conversational scenarios, which allowed creation of a database of possible phrases the agents would use in a conversation with learners, and to predict possible language input from learners based on a script from real life.

The third study revealed how learners perceived the interaction with ECAs and what effect it had on their perceptions of learning, both in terms of language and culture. It resulted in necessary improvement, i.e. fine-tuning, of the agents' multimodal behaviour of CRs. As this application represents a serious game using a communicative language teaching approach, game-based and task-based learning, the pilot study showed that it had an educational potential. *Virtual Reykjavik* supported speaking practice, which is a problem according to the findings from the initial survey. This problem is two-fold. On the one hand, Icelanders tend to switch into English whenever speaking to foreign learners, and on the other hand, learners of Icelandic often feel insecure and nervous when speaking Icelandic with native speakers in real life.

5.2 Contributions

This interdisciplinary thesis made a contribution to both practical and theoretical areas of ICALL and CRs in Icelandic in the context of design of embodied conversational agents, and human-agent interaction.

5.2.1 Intelligent Computer Assisted Language Learning (ICALL)

In the field of ICALL, this thesis demonstrated how a language and culture training application can support learning by practising spoken language skills with ECAs endowed with authentic multimodal features modelled on real-life language use. In this context, the definition given by Morie et al., (2012) has been adopted which considered ECAs as intelligent programmes that are delivered through a three–dimensional graphic entity, which has the ability to interact with a human user by both text or speech, either on a web or stand-alone computer (p. 2). The dialogues in *Virtual Reykjavik* were designed according to a scenario captured on a video camera from real-life situations in central Reykjavik. Agents were equipped with language data as found in the transcription and analysis of these video recordings. The language data also contributed to the development of learning materials and creating lessons in *Virtual Reykjavik*. These materials and lessons are, however, part of future work and therefore not included in this thesis. In general, it would be practical for learners to become familiar with how real language is constructed by native speakers in reality, and learn about those expressions with particular multimodal behaviour that is specific for the culture. Some expressions, such as the CR

Fragment/Interjection Strategy *Ha? (Huh?)* might seem rude to other cultures, but in Iceland it is considered as a common way to seek clarification. Regarding the spoken output, some learners reported that the game enabled them to start speaking Icelandic, which they previously lacked even in real life in Iceland. They also reported learning about places in central Reykjavik, which means that they additionally practised reading. The perception of learners for the game's future is the same purpose, i.e. to hold a realistic interaction with agents who resemble native speakers. By doing so the learners will learn how language is used in reality and become familiar with real-life situation in a safe virtual environment, where they can re-take lessons each time they require. This puts *Virtual Reykjavik* into the category of 3D VLEs that simulate real-world situations and support interaction with virtual characters by spoken language.

5.2.2 Clarification Requests in Icelandic

This thesis also contributed to the field of pragmatics in connection with CRs in Icelandic, in that it conducted another research on CR strategies. The previous research was conducted by Gísladóttir (2015). From a linguistic point of view, pragmatics deals with "intentional acts of speakers at times and places, typically involving language" (Korta and Perry, 2015, p. 2). According to Korta and Perry, it is the role of pragmatics, especially 'far-side pragmatics', to explain the information participants in a conversation convey and the actions they perform in or by saying something (ibid, p. 3). As pragmatics was also chosen as the main philosophy behind the study on CRs, the contribution is therefore relevant in this field. Based on Gísladóttir (2015) and the study on multimodal CRs presented in this thesis, the linguistic types of CR categories were compared, which revealed an interesting outcome. Both studies list six types of CR, however, each study was conducted in a different conversational setting. Gísladóttir (2015) investigated repair initiators in an everyday interaction involving three people or more, with an age range between ten to late eighties, whereas the current study investigated the repair strategies in first encounters between native and non-native speakers of Icelandic. The difference in the conversational setting and the choice of participants may explain the difference in frequency of some CR types. For instance, Gísladóttir (2015) lists the Interjection Strategy as the most frequent one (34.7%), whereas in the present study it is much less (3.9%). The falling pitch of intonation is the same in both cases. On the other hand, the Ellipsis is the most frequently used CR in the present study (79.2%) whereas Gísladóttir (2015) reported reduced usage at 22.4%. In the present study, the category of request (asking for specification) was put into the same category as Ellipsis due to their same form. Similarly, the question word strategy and formulaic were put into the category of partial or full explicit queries. The request type was put into the same category as repetitions because it represents the same form. On the other hand, Gísladóttir did not find any examples of non-verbal repair strategies, in contrast to the present study. The present study did not find any alternative question strategies, as was the case in Gísladóttir (2015). Rather, due to their nature and content reference, these were considered as Offering a Candidate in the present study.

The findings from the present study expanded the view of pragmatics in connection to the use of CR utterances in Icelandic. Comparing, the two studies found differences that are conditioned to the context in which a conversation occurs. This means that conversational scenarios in various other contexts may influence participants to produce other forms of CRs and with different level of frequency. In this view, it would be interesting for pragmatics to extend research on CR strategies involving other situational contexts.

5.3 Limitations

The summary of limitations to this thesis is presented in this section. As it included three separate studies, the limitations from these will be discussed in separate sections.

5.3.1 Theoretical Limitations

The lack of literature on multimodal CRs in Icelandic creates a major theoretical limitation of this thesis. Regarding the literature on modelling agent's multimodal behaviour for CRs, there was only a limited number of known studies (Bickmore and Cassell, 2005; Louwerse et al., 2007; Healey et al., 2015) that discuss this problem from a very brief, descriptive manner. Those studies do not offer any parameters nor specific models that could be adopted. As this thesis suggests theoretical models of six multimodal CRs in Icelandic, it is questionable whether other research might adopt the same structure of the proposed models. It would be, nonetheless, revealing to see another version of the multimodal CR model for other languages as well.

5.3.2 Limitations from the Studies

Each study has encountered several limitations. The initial study in section 3.2, which was based on a survey investigating learners' expectations from *Virtual Reykjavik*, had a limitation in the representative sample and the sampling method. Despite selecting the Random Sampling Method (Creswell, 2009, p. 148), an email with a link to a questionnaire on Google Docs was sent to a university email including groups of learners of Icelandic at the University of Iceland. The limitation here represents a limited sampling pool, which the sampling addressed. Besides classes of Icelandic at the University of Iceland, there are other institutions and language schools that offer classes of Icelandic to foreigners living in Iceland. The sampling could have included additional institutions, which may have increased the responses to the survey and provided a greater variety of answers. Another limitation is the selection of Google Docs as a tool for collecting data. Even though it very easy to operate and open source, it has a bias regarding ethical issues of using data for Google's research or marketing purposes, which could have affected others from using it and thus may have limited the response rate. Another limitation is the restriction only to use of an online form, which could have prevented others from participating who prefer questionnaires on paper. An additional limitation to the survey were its questions that included open answers, which could have prevented others from writing extensive explanations or providing reasoning. Moreover, the questions were originally created only for this study, which could cause low validity. There was no standardised questionnaire from a previous research used to examine the learners' needs and expectations. Instead, questions were tailor-made, which contributed to the limitations of the survey. Tailor-making a questionnaire is more common in SLA studies to help assess learners' needs in a specific learning environment but using a standardised set of questions from previous research in a combination with originally developed questions for this study may undoubtedly increase the validity and reliability of research.

The study on multimodal CRs has several limitations common to data collection in the field. As the study was not carried out in a laboratory setting, not all factors could be controlled for. The data obtained does not include all age categories, because it is difficult to address children without parental permission for video-recordings and use the material for the purpose of this research. Individuals of 70 years old do not frequent central Reykjavik, and therefore it is difficult to obtain a sufficient number of videos from this demographic. The most common age encountered in central Reykjavik is between 20-50 years old, which represents another limitation to this study. The results from this study are partly based on interpretation (Heath et al., 2010; Knoblauch and Tuma, 2011; Veer, 2013), and therefore partly subjective to the researcher. Only those recordings have been used that received permission from the participants after the recording had been made. All other recordings have been deleted leading to fewer recordings used in the research. The last known limitation of this study is the narrow scope of the CR in the presented conversational scenario. Perhaps further recordings would have revealed more types of CRs that did not appear in this study, but this can be included within the section on future work. The inclusion of other conversational scenarios would expand the scope of CRs for different types to be used by ECAs.

There are several limitations to the pilot study. The first one is the low number of participants (n = 6). This could have been caused by the sampling method, which was again the Random Sampling method (Creswell, 2009, p. 148) but addressing only one group of participants: first-year students of Icelandic Practical Diploma studying at the University of Iceland. The second limitation is the lack of clear instructions about the game, how to navigate in it, what the purpose of the tasks is, and how many levels (which game scenarios) can the learners achieve by playing the game. However, one of the main limitations to this study was the speech recognition for Icelandic input. The version used at the time of testing was not optimised. Often it could not recognise the speech input from learners and/or the internet connection was weak. For those reasons, it was often not fully synchronised with the game. In addition, the learning material from the game was missing. Learners did not have any point of reference as to which vocabulary and grammar they were going to exercise in the game. Lessons in the form of learning materials were missing. *Virtual Reykjavik* represents a prototype, but above all, an unfinished version of the game, which include a single scene with one scenario and main features for playing. For instance, the main menu looked unfinished, which could have caused lower initial impressions of users for the game and which could have biased their general perception of it. Moreover, learners encountered some technical difficulties, such as freezing and restarting of the game, that could have affected their responses in the study.

5.3.3 Technical and Practical Limitations

One of the main practical limitations is the inclusion of only two out of six multimodal CR models into the ECA's conversational architecture. Due to time and budget constraints, it was possible to implement only two of those. In addition to that, both models required fine-tuning, especially the Interjection Strategy. Fine-tuning could have only been detected by a trial-and-error basis, i.e. to let learners play the game, report on the multimodal behaviour implemented in the CR functions that ECAs execute, and then fine-tune it accordingly, while keeping it in line with the observations in the video recordings and the proposed CR model.

Another technical limitation is the lack of implemented emotions in the ECA's multimodal behaviour, which was designed to correspond with the proposed model, but nonetheless it lacked emotions in other parts. General findings from the pilot study reveal that agents did not smile, which could have affected the learners' overall perception of the agent's multimodal behaviour including CRs. When reviewing the multimodal *Virtual Reykjavik* corpus, smiles can be found in some videos by some speakers, but they are usually present at the end of the conversation and during a different communicative function. Nonetheless, having not included any obvious signs of friendliness such as a smile, could have affected the learners' perception of the agents' multimodal behaviour and therefore represents a limitation to this study.

There are also some problems associated with realistic interaction in *Virtual Reykjavik*. The nature of spoken natural language by learners is often imperfect with respect to pronunciation or speaking too softly into the microphone. This represents a technical limitation to the study because learners were limited by it. As the findings from the pilot study reveal, such situations may cause frustration for users. Another problem is that the nature of a realistic interaction is often associated with speaking the way one thinks. In human-agent interaction, restrictions apply on the side of agents. Their capability of replying appropriately to any possible phrase spoken by the human user is limited and restricted only to the pre-programmed database of phrases used in each conversational scenario.

5.4 Future work

This thesis supported the mutual interest of both parties, the learners of Icelandic and the designers of this game, to develop a simulation of reality where learners can meet virtual characters that speak Icelandic. From the point of view of learners, such characters would simulate a real-life conversation in virtuality and thus increase the exposure to and practice of spoken language in an authentic virtual interim learning space. From the point of view of the game designers, the characters would represent ECAs with realistic multimodal behaviour, and thus support the development of a realistic human-agent interaction in a serious computer game for learning a language and culture. As it was mentioned in the previous chapter, however, there are several limitations of this thesis that prompt ideas for future work. This chapter, therefore, focuses on further research that can help the field of CALL and the development of a realistic human-agent interaction benefit from new findings. Suggestions for future work are discussed in separate sections here below.

5.4.1 Survey on Learners' Expectations

Based on the findings and limitations of the initial survey which shed light on both the learners' expectations from *Virtual Reykjavik via* an online survey and the problems they encounter with practising Icelandic in the target language environment, future work would be to expand the survey to additional questions about playing computer games in order to find out preferences of different age and gender groups. Furthermore, it would be also interesting to include questions about the learners' previous experience in playing computer games and also about other specific features they expect from *Virtual Reykjavik* regarding gamification, feedback, interaction with virtual characters and learning. In addition, the survey could be distributed in both online and paper form in order to increase the likelihood of participation. Inclusion of other institutions and language schools where Icelandic is taught as a foreign and/or second language both in and outside of Iceland would allow comparison of data between different learner groups. As stated earlier, a standardised set of questions from previous research could be used and adapted for a future needs study. This would contribute to both the validity of questions in the study and the reliability of research.

In order to do a further comparative study, it would be also interesting to do a similar survey in the Greek part of Cyprus on Greek or Cyprian language. Even though the country's languages are Greek and Cyprian, the English language is widely spoken there. Due to prevalent switching into English, learners may encounter similar difficulties in practising the local language as learners in Iceland do.

5.4.2 Multimodal Clarification Requests

Future work could include other conversational scenarios in order to find out whether there are other types of CRs and how they are multimodally executed by native Icelandic speakers. As van Dijk (2008) suggests, it is the proposition of time, place, shared knowledge (common ground) of the participants that should be observed and analysed (p. 11). For instance, the gestures, postures, faces, bodies, movements, place where other speakers are standing in a particular environment (Blommaert and Rampton, 2011, p. 6; Jewitt, 2013b, p. 142) shape the way people talk to each other and behave. Other conversational settings could contribute to finding new types of CRs that are perhaps more relevant to those settings. When doing another research on multimodal CRs, it would also be interesting to hire an actor to let him/her perform various kind of CR strategies, record the performance and compare with findings from a field study. The actor's performance could be tracked by a motion capture system capable of detecting and tracking facial expressions as well as body movements. Such data would provide more precise values for modelling multimodal behaviour in CR functions. However, in order to identify the most crucial non-verbal cues in CRs, it would be very practical to use motion capture, 3D video tracking system and speech activity detector to record all verbal and non-verbal cues when participants speak in different conversational scenarios. This would require a laboratory setting and hiring actors to speak and act in prescribed conversational setting. Even though the authenticity of data would not be as high as from field research, the quality of collected multimodal data would on the other hand be expected to be very high. It would hasten the process of data collection and enable a more accurate and fast data analysis. Instead of the collection of non-verbal cues, the instruments will detect the most crucial ones for modelling multimodal behaviours of ECAs in various communicative functions. These would be eye, eyebrow and mouth movement, forehead shape, head movement, body posture, and movements of hands.

In addition to the above suggestions, it would nonetheless be interesting to conduct the same study as presented in this thesis but in another language and compare the results to see whether there are some similarities in the multimodal production. In case there is a similar pattern across languages, it would inform about the universality of language. However, if there are differences it would inform about the culturally bound differences in speaking a language in different cultures. Since two CRs were implemented into the ECAs conversational behaviour in *Virtual Reykjavik*, future work could include the implementation of the remaining four known CRs. This attempt would contribute towards the full development of conversational behaviour of agents. It would be also interesting to measure the learning effect of how multimodal CR models executed by ECAs helped learners to develop their language and cultural skills, and whether a learning transfer took place.

There is also Icelandic Sign Language (ISL) which could benefit from executing similar research, since there are currently no known studies describing multimodal communicative functions. This would enormously help developers of language learning applications to design ECAs with multimodal behaviour for teaching ISL. It would be a continuation of the current study, which could be based on the Swiss application *Menusigne*[34], which is a serious game for teaching sign language grammar. The whole Icelandic society could benefit from such a game, because it will help teaching ISL *via* CALL, and thus bring the deaf and hearing-impaired community in the area of language education online.

5.4.3 Improving the User Response Study for **Virtual Reykjavik**

In case of testing *Virtual Reykjavik* in a user response study, there are several proposals for future work. As has already been proposed in the previous section on limitations, one of the ideas for future work is to include more participants in the testing of *Virtual Reykjavik* and to adapt standardised sets of questions from similar research. In order to conduct a better User Response Study, the game should be developed to a full product, a finished and fully-functioning version of an Icelandic speech recognition system should be implemented, and the game instructions as well as scaffolding of learning materials for learners should be created. Apart from the general perception of the game, it would also be worthwhile to focus on the perception of multimodal behaviour of ECAs in CRs. This will provide results for further fine-tuning of the agents' multimodal behaviour. Then it would be good to re-do the whole process until an optimal realisation of the agents' behaviour has been achieved. For the process of video recording participants while playing the game, a second camera could be installed closer to the computer to capture images of participants expressions/emotions and the way they operate the instruments (keyboard, mouse). It would be also interesting to include an eye tracking system to find out whether and how much time they spend exploring different parts of the screen (e.g. reading transcription of dialogues, checking their 'speaking button' - red microphone - signalling when they can speak), which would contribute to solving possible issues connected to focussing on 'distractive parts' of the game instead of speaking to the ECAs. The ease or difficulty of access and operating these buttons could have an effect on the learners' overall perception of the game, which eventually may influence their perception of the agent's multimodal behaviour.

At the same time, it would be also interesting to measure the learning effect after the participants had played the game. Combining standardised sets of questions measuring training performance, cognitive learning of new language skills, and comparing them with

[34] http://speech2sign.unige.ch/en/applications/menusigne/

a post-training questionnaire would help shed light on the effectiveness of the game for learning the language and culture. Since participants can use the game in different instructional contexts, e.g. individual study or part of a distant course training, approaches from the Second Language Acquisition (SLA) would help guide the constructing of tests for measuring learning of different language skills in different learning situations. Moreover, it would be very interesting to conduct a longitudinal study, i.e. testing the same participants after a longer period of time, but for this reason the game would need to be fully functioning and offering tasks for different language levels.

5.4.4 Multimodal Behaviour of Embodied Conversational Agents

In regard to the perception of agents' multimodal behaviour, further studies would be to validate the CR behaviour of ECAs as to how close do the agents appear to the native speakers in the videos. The researcher would receive an empty grid to fill in the multimodal features observed on native speakers and then compare them with the features in the suggested CR models. The next idea for future work would be to implement a smile into the agents' multimodal behaviour when finishing a conversation. Based on findings from Cafaro's (2014) PhD thesis, smiles are a very important feature in human conversation which should similarly be implemented into the conversational architecture of ECAs. A first impression of a friendly attitude is important and should also be implemented at the end of the conversation. The findings from the present study on CRs revealed that smile is not used by native speakers upon initial contact, but rather during the conversation or as part of the CR sequence. Learners' answers in the User Response Study indicated that ECAs would have been perceived as friendlier if they smile. In the suggested models for multimodal CRs, the data showed when native speakers performed a smile. It was usually a very subtle smile. This should be implemented next time into the ECA's multimodal behaviour and will indicate signs of friendliness. Endowing agents with a large variety of expressions, including emotions, may contribute to more naturalness in the agent's behaviour (Pelachaud, 2009, p. 3546). A smile at the end of a conversation was observed by some participants in the video recordings and even though this particular feature was not part of this thesis. Future work could include this additional feature and compare the findings with the current study.

5.4.5 *Improving* **Virtual Reykjavik**

In the area of CALL development in Iceland, *Virtual Reykjavik* represents a new stage of intelligent CALL, by not only including authentic ECAs with realistic multimodal behaviour but also the process of storing and using learner data for feedback purposes and their language development. In today's globalised and technically advancing world, another area for future work would be to expand *Virtual Reykjavik* into Mobile-Assisted Language Learning (MALL) so as to adjust the platform to be used in mobile devices, such as smartphones and tablets. Moreover, with ever-more accessible virtual reality (VR) technology, this application could include a VR headset adjusted so as to use in the game to enable learners to immerse themselves in a 3D virtual world, e.g. *MondlyVR*[35]. Continual technological improvements and software innovations have already enabled the use of new devices for language learning, such as mobile phones, personal digital assistants (PDAs) and iPods that lead to a more mobile way of learning, i.e. Mobile Assisted Language Learning (MALL) (Chinnery, 2006, p. 9). The use of smartphones has led to autonomy in language learning in both inside and outside of the language classroom (Leis et al., 2015, p. 75). Learners become 'mobile' by using devices that are "small, autonomous and unobtrusive enough to accompany us in every moment" (Trifonova and Bonchetti, 2003, p. 3) and use them anywhere and anytime for the purpose of learning, whether formal or informal (Kukulska-Hulme and Shield, 2007, p. 3). Such possibilities create new interim learning spaces with making the exposure to L2 and its culture even more realistic due to their mobility and applicability.

In general, there is a need for further research in face-to-face interactions in real life among native and non-native speakers investigating other communicative functions, which would contribute to the area of multimodal communication and multimodal behaviour of ECAs, and thus help create a more realistic human-agent interaction. This would enable learners a greater exposure to an authentic language and culture at distance.

5.5 Conclusion

This chapter concluded the aim of this thesis by summarising the areas in which it contributed, but also those in which it was limited to. Consequently, it suggested ideas for future work in order to continue and expand the work undertaken so far. Therefore, in

[35] MondlyVR belongs to one of the most advanced ways to learn language in a virtual reality by interacting with virtual characters: https://www.oculus.com/experiences/gear-vr/1272636489423125/

conclusion, this thesis informed about the latest development of ICALL in Iceland, by which it introduced the body of work done in the Icelandic language and culture training application *Virtual Reykjavik* project. This aimed to endow ECAs with authentic multimodal behaviour in the CR function to achieve a realistic human-agent interaction, helping learners to develop their communication skills. The ECAs play a define social role with learners who also play a defined role using specific communicative skills to achieve goals. As Lane et al. (2013) note, "[t]he technology challenge is to simulate social encounters in realistic ways and in authentic contexts. The pedagogical challenge is to design scenarios in ways that achieve the learning goals, maintain a high level of real-world fidelity, and stay within an ideal window of challenge (whatever that may be)" (p. 1). The resemblance of real world, or how true to the real world the situation in a VLE is (fidelity), is represented in this thesis by the natural behaviour of ECAs using multimodal CRs when speaking. By selecting one communicative function, it demonstrated that ECAs can be endowed with authentic multimodal features enabling them to speak and act like local speakers. This puts *Virtual Reykjavik* to a category of a 3D interim learning space for bridging the gap between the learning and using the language in real life. In addition to this, the material gathered from the field research in the form of video recordings helped to create lessons and dialogues for particular scenarios with the aim to simulate a realistic interaction and thus keep the learners in a flow, i.e. to motivate learners to remain and re-enter the game for the purpose of learning authentic language skills in Icelandic. Instead of using ready-made dialogues from language textbooks in similar situations, the language material gathered from transcriptions of the video recordings was used to establish connection with the real life. This thesis presented the *Virtual Reykjavik* as an alternative safe virtual space where learners can focus on practising spoken language skills without switching into English, as is often the case in real life.

Bibliography

Abraham, W. and Leiss, E. (2012). *Modality and Theory of Mind Elements across Languages*. De Gruyter Mouton.

Abrilian, S., Devillers, L., Buisine, S., Martin, J. C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *Proceedings of Human-Computer Interaction International*, pp. 1-10. Las Vegas. https://pdfs.semanticscholar.org/fa51/e8714cda19db7208b152e9b9c62fa3a7941b.pdf?_ga=2.236071623.714746202.1583964798-1841246059.1583705776.

Abuczki, A. and Ghazaleh, E. B. (2013). An overview of multimodal corpora, annotation tools and schemes. *Argumentum 9*, pp. 86-98. Debreceni Egyetemi Kiadó. http://argumentum.unideb.hu/2013-anyagok/kulonszam/01_abuczkia_esfandiari_baiat.pdf.

Agee, J. (2009). Developing qualitative research questions: a reflective process. In *International Journal of Qualitative Studies in Education 22*(4), pp. 431-447. Routledge. https://doi.org/10.1080/09518390902736512.

Almeida, L. C. (2012). The Effect of an Educational Computer Game for the Achievement of Factual and Simple Conceptual Knowledge Acquisition. *Education Research International (2012)*, pp. 1-5. Hindawi Publishing Corporation. http://dx.doi.org/10.1155/2012/961279.

Alvarez, K., Salas, E., and Garofano, C. M. (2004). An Integrated Model of Training Evaluation and Effectiveness. In *Human Resource Development Review 3*(4), pp. 385–416. https://doi.org/10.1177/1534484304270820

Alwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C., and Paggio, P. (2005). The MUMIN multimodal coding scheme. *NorFa yearbook 2005*, pp. 129-157. http://sskkii.gu.se/jens/publications/bfiles/B70.pdf.

Amoia, M. (2011). I-FLEG: A 3D Game for Learning French. *Proceedings of the 9th International Conference on Education and Information Systems, Technologies and Applications – EISTA 2011*, pp. 1-6. Orlando, US. https://members.loria.fr/CGardent/publis/eista11.pdf.

Amoia, M., Bretaudiere, T., Denis, A., Gardent, C., and Perez-Beltrachini, L. (2012). A Serious Game for Second Language Acquisition in a Virtual Environment. *Journal of Systemics, Cybernetics and Informatics (JSCI) 10*(1), pp. 24-34. HAL archives-ouvertes.fr. http://www.iiisci.org/journal/CV$/sci/pdfs/HEA308SP.pdf.

Arbuthnott, K. D. and Krätzig, G. P. (2015). Effective teaching: Sensory learning styles versus general memory processes. *Innovative Teaching 4*(2). Ammons Scientific. http://journals.sagepub.com/doi/full/10.2466/06.IT.4.2.

Arnbjörnsdóttir, B. (2004). Teaching morphologically complex languages online: Theoretical questions and practical answers. In P. J. Henrichsen (Ed.), *Call for the Nordic Languages. Tools and Methods for Computer Assisted Language Learning* (pp. 79-94). Copenhagen Studies in Language 30. Copenhagen. Samfundsliteratur.

Arnbjörnsdóttir, B. (2008). Covcell and Less Commonly Taught Languages: View, Experiences and Opportunities. In Arnbjörnsdóttir, B. and Whelpton, M. (Eds), *Open Source in Education and Language Learning Online. Article collections and proceedings* (pp. 47-71). Reykjavik: University of Iceland Press.

Arnbjörnsdóttir, B. (2011). Exposure to English in Iceland: A Quantitative and Qualitative Study. *Netla - Menntakvika 2011*. School of Education. University of Iceland. http://netla.hi.is/menntakvika2011/004.pdf.

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. *International Journal of Artificial Intelligence Education 29*, pp. 29–57. https://doi.org/10.1007/s40593-018-0171-6.

Bachen, C. M., Hernández-Ramos, P. F., Raphael, C., and Waldron, A. (2016). How do presence, flow, and character identification affect players' empathy and interest in learning from a serious computer game? In *Computers in Human Behavior 64*, pp. 77-87. https://scholarcommons.scu.edu/comm/38/

Bachman, L. (1995). *Fundamental Considerations in Language Testing*. Third impression. New York: Oxford University Press, USA.

Bado, N. and Franklin, T. (2014). Cooperative Game-based Learning in the English as a Foreign Language Classroom. *Issues and Trends in Educational Technology 2*(2). The University of Arizona. https://journals.uair.arizona.edu/index.php/itet/article/view/18190/18064.

Barkand, J. & Kush, J. (2009). GEARS a 3D Virtual Learning Environment and Virtual Social and Educational World Used in Online Secondary Schools. *Electronic Journal of e-Learning* 7(3), pp. 215 - 224. http://www.ejel.org/issue/download.html?idArticle=100.

Barrett, K. A. & Johnson, W. L. (2010). Developing Serious Games for Learning Language-in-Culture. In R. Van Eck (Ed.), *Gaming and Cognition: Theories and Practice from the Learning Sciences* (pp. 281-311). Information Science Reference. Hersey-New York, USA. https://psycnet.apa.org/doi/10.4018/978-1-61520-717-6.ch013.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences 7*(2), pp. 84-91. Elsevier. http://decisionneurosciencelab.org/pdfs/Barsalou%20et%20al.%20(2003a).pdf.

Baszanger, I. and Dodier, N. (2004). Ethnography: Relating the part to the whole. In D. Silverman, (Ed.), *Qualitative Research: Theory, Method, and Practice* (pp. 9-34). 2nd Edition. London: Sage.

Bateman, J. A. (2012). Multimodal Corpus-Based Approaches. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-4). Blackwell Publishing. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0824.

Bavelas, J. B., Hutchinson, S., Kenwood, C. & Matheson, H. D. (1997). Using Face-to-face Dialogue as a Standard for Other Communication Systems. *Canadian Journal of Communication 22*(1), pp. 1-7. http://cjc-online.ca/index.php/journal/article/view/973/879.

Beinborn, L., Botschen, T., and Gurevych, I. (2018). Multimodal Grounding for Language Processing. *Proceedings of the 27the International Conference on Computational Linguistics*, pp. 2325-2339. Santa-Fe: COLING. https://www.aclweb.org/anthology/C18-1197.pdf.

Benoit, Ch. (1999). The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces. M. M. Taylor, F. Néel, and D. G. Bouwhuis (Eds.), *The Structure of Multimodal Dialogue II*. Amsterdam: John Benjamins.

Bernstein, J., Najmi, A., & Ehsani, F, (1999). Subarashii: Encounters in Japanese Spoken Language Education. *CALICO Journal 16*(3), pp. 361-384. http://www.tassopartners.com/wp1/wp-content/uploads/2014/02/SUBARASHI-CalicoJ-1999.pdf.

Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., and Pelachaud, C. (2010). GRETA: Towards an Interactive Conversational Virtual Companion. *Artificial Companions in Society: perspectives on the Present and Future*, pp. 143-156. https://www.researchgate.net/profile/Ken_Prepin/publication/225027976_GRETA__Towards_an_Interactive_Conversational_Virtual_Companion/links/0fcfd50b4be876f6ff000000.pdf.

Bédi, B., Arnbjörnsdóttir, B., Vilhjálmsson, H. H., Helgadóttir, H., Ólafsson, S., Björgvinsson, E. (2016). S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016*, pp. 37-43. Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.535.

Bédi, B., Arnbjörnsdóttir, B., Vilhjálmsson H. H. (2017). Learners' Expectations and Experiences in Virtual Reykjavik. J. Colpaert, A. Aerts, R. Kern and M. Kaiser (Eds.) *Proceedings of CALL in Context 2017*, pp. 75-82. University of California, Berkeley. http://call2017.language.berkeley.edu/wp-content/uploads/2017/07/CALL2017_proceedings.pdf.

Bickmore, T. and Cassell, J., (2005). Social Dialogue with Embodied Conversational Agents. In J. van Kuppevel, L. Dybkjaer, and N. Bernsen (Eds.), *Advances in Natural Multimodal Dialogue Systems* (pp. 23-54). New York: Kluwer Academic. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.4544&rep=rep1&type=pdf.

Bjarnadóttir, K. Ý. (2009). *Bein ræða í samtölum*. BA Thesis. University of Iceland. https://skemman.is/bitstream/1946/2168/1/Kolbrun_Yrr_Bjarnadottir_fixed.pdf.

Blommaert, J. and Rampton, B. (2011). Language and Superdiversity. *Diversities 13*(2), pp. 1-21. UNESCO. http://www.unesco.org/shs/diversities/vol13/issue2/art1.

Blöndal, Th. (2008). Turn-final eða ('or') in spoken Icelandic. In J. Lindstöm (Ed.), *Språk og Interaktion 1* (pp. 151-168). Helsingfors Universitet. http://hdl.handle.net/10138/28494.

Blyth, C. (2018). Immersive technologies and language learning. *Foreign Language Annals 51*(1), pp. 225-232. American Council on the Teaching of Foreign Languages. https://doi.org/10.1111/flan.12327.

Bossomaier, T. R. J. (2012). *Introduction to the Senses: From Biology to Computer Science*. Cambridge University Press.

Braun, V., Clarke, V. and Rance, N. (2014) How to use thematic analysis with interview data. In Vossler, A. and Moller, N. (Eds.), *The Counselling & Psychotherapy Research Handbook*, pp. 183-197. London: Sage http://www.uk.sagepub.com/textbooks/Book239261.

Brinkmann, S. (2018). Visual Research. In K. D. Norman and Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research,* pp. 997-1038. Fifth Edition. SAGE.

Brown, D. H. (2007). *Principles of Language Learning and Teaching*. Fifth Edition. Pearson Education. USA.

Brown, P. and Levinson, S. C. (1987). Politeness: Some universals in language usage. *Studies in International Sociolinguistics 4*. Cambridge University Press.

Burgos, D., van Nimwegen, C., van Oostendorp, H., and Koper, R. (2007). Game-based learning and the role of feedback: a case study. In *Advanced Technology for Learning 4(*4), pp. 188–193. https://dl.acm.org/doi/10.5555/1722214.1722217.

Burns, H. L. and Capps, C. G. (1988). Foundations of Intelligent Tutoring Systems: An Introduction. In M. C. Polson and J. J. Richardson (Eds.), *Foundations of Intelligent Tutoring Systems* (pp. 1-19). Lawrence Erlbaum Associates Publishers.

Bytheway, J. A. (2011). *Vocabulary Learning Strategies. Massively Multiplayer Online Role-Playing Games*. MA-Thesis. Victoria University of Wellington. https://core.ac.uk/download/pdf/41337041.pdf.

Cafaro, A. (2014). *First Impressions in Human-Agent Virtual Encounters*. Doctoral Dissertation. Reykjavik University. http://en.ru.is/media/td/Angelo-Cafaro-PhDThesis.pdf.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., and Pelachaud, C. (2014). Representing Communicative Functions in SAIBA with a Unified Function Markup Language Intelligent Virtual Agents. In T. Bickmore, S. Marsella, and C. Sidner

(Eds.), *Proceedings of IVA 2014*, pp. 81-94. Springer. http://www.ru.is/faculty/hannes/publications/IVA2014.pdf.

Caglayan, O., Sanabria, R., Palaskar, S., Barrault, L., and Metze, F. (2019). Multimodal Grounding for Sequence-to-Sequence Redognition. *Proceedings of ICASSP2019*, pp. 1-5. IEEE. https://arxiv.org/pdf/1811.03865.pdf.

Carli, L. L. (1989). Gender Differences in Interaction Style and Influence. *Journal of Personality and Social Psychology 56*(4), pp. 565-576. http://www.communicationcache.com/uploads/1/0/8/8/10887248/gender_differences_in_interaction_style_and_influence.pdf.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L, Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in Conversational Interfaces: Rea. *The CHI´99 Conference Proceedings*, pp. 520-527. http://www.media.mit.edu/gnl/publications//CHI99.pdf.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., and Yan, H. (2000). Human Conversation as a System Framework: Designing Embodied Conversational Agents. In J. Cassell, J. Sullivan, S. Prevost, and Churchill, E. (Eds.), *Embodied Conversational Agents* (pp. 29-63). Cambridge MA: MIT Press. http://www.media.mit.edu/gnl/publications//ECA_GNL.chapter.to_handout.pdf.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). BEAT: the Behavior Expression Animation Toolkit. *Poceedings of SIGGRAPH 2001*, pp. 477-486. Los Angeles. http://www.ru.is/kennarar/hannes/publications/siggraph2001.pdf.

Chapelle, C. A. (2001). *Computer Applications in Second Language Acquisition. Foundations for Teaching, Testing and Research*. M. H. Long and J. C. Richards (Eds.). Cambridge Applied Linguistics. Cambridge University Press.

Chapelle, C. A. (2008). Computer Assisted Language Learning. In B. Spolsky and F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 585-595). Blackwell Publishing.

Chee, Y. S. (2016). *Games-To-Teach or Games-To-Learn: Unlocking the Power of Digital Game-Based Learning Through Performance*. Springer.

Chinnery, G. M. (2006). Emerging Technologies. Going to the MALL: Mobile Assisted Language Learning. *Language Learning and Technology 10*(1), pp. 9-16. http://llt.msu.edu/vol10num1/pdf/emerging.pdf.

Cho, E-a. (2007). Next-Turn Repair Initiators in English Conversation between Korean Speakers. *Linguistics 15* (3), pp. 141-160. https://www.scribd.com/document/310651176/English-Conversation-between-Korean-Speakers.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences 3*(9), pp. 345-351. Elsevier. https://doi.org/10.1016/S1364-6613(99)01361-3.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition 22*(1986), pp. 1-39. Elsevier. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.4427&rep=rep1&type=pdf.

Clark, H. H. and Brennan, S. E. (1991). Grounding in Communication. In L. B. Resnick, J. M. Levine, and S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Second Edition. Washington: American Psychological Association.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clark, J. M. and Paivio, A. (1991). Dual Coding Theory and Education. *Educational Psychology Review 3*(3), pp. 149-210. Premium Publishing Corporation. http://www.csuchico.edu/~nschwartz/Clark%20%26%20Paivio.pdf.

Colman, M. and Healey, P. G. T. (2011). The distribution of repair in dialogue. *Proceedings of CogSci,* pp. 1563-1568. http://mindmodeling.org/cogsci2011/papers/0353/paper0353.pdf.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., and Boyle, J. M. (2012). A Systematic Literature Review of Empirical Evidence on Computer Games and Serious Games. *Computers and Education 59*, pp. 661-686. Elsevier. http://dx.doi.org/10.1016/j.compedu.2012.03.004.

Cornillie, F., Clarebout, G., and Desmet, P. (2012). The role of feedback in foreign language learning through digital role playing games. In *Procedia – Social and Behavioural Sciences 34*, pp. 49-53. Languages, Cultures and Virtual Communities. Elsevier. https://doi.org/10.1016/j.sbspro.2012.02.011.

Covaci, A., Ghinea, G., Lin, C-H., Huang, S-H., and Shih, J-L. (2018). Multisensory game-gased learning-lessons learnt from olfactory enhancement of a digital board game. *Multimedia Tools and Applications*, pp. 1-19. Springer. https://link.springer.com/content/pdf/10.1007%2Fs11042-017-5459-2.pdf.

Coyne, I. T. (1997). Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries? *Journal of Advanced Nursing 26*(3), pp. 623-630. https://pdfs.semanticscholar.org/875d/5c4533edfeffa0b1b8291b82f9a3c13342b3.pdf.

Creswell, J. W. (2009). *Research Design. Qualitative, Quantitative and Mixed Methods*

Crouch, M. and McKenzie, H. (2006). The logic of small samples in interview-based qualitative research. In *Social Science Information 45*(4), pp. 483-499. SAGE. https://doi.org/10.1177%2F0539018406069584.

Crystal, D. (2005). *Speaking of Writing and Writing of Speaking*. Pearson Longman. http://www.pearsonlongman.com/dictionaries/pdfs/speaking-writing-crystal.pdf.

Dale, E. (1969). *Audiovisual Methods in Teaching*. Third Edition. Dryden Press: New York.

Davies, G. (2000). *CALL (Computer assisted language learning). Routledge encyclopedia of language teaching and learning* (pp. 90-93). London: Routledge.

De Paepe, L. (2018). Student performance in online and face-to-face second language courses: Dutch L2 in adult education. In *Journal of Educational Sciences 1*(37), pp. 66-76. https://rse.uvt.ro/pdf/2018/NR1/6.pdf.

Dillenbourg, P. (2000). Virtual Learning Environmnets. *Proceedings of the EUN Conference 2000: Learniong in the New Millenium: Building New Education Strategies for Schools. Workshop on Virtual Learning Environmnets*. University of Geneva. https://tecfa.unige.ch/tecfa/publicat/dil-papers-2/Dil.7.5.18.pdf.

Dillenbourg, P., Schneider, D. K., and Synteta, P. (2002). Virtual Learning Environmnets. A. Dimitracopolou (Ed.), *Proceedins of the 3rd Hellenic Conference Information and Communication Technologies in Education*, pp. 3-18. Kastaniotis Editions, Greece. https://telearn.archives-ouvertes.fr/hal-00190701/document.

Ding, T. I. N. G. (2012). The comparative effectiveness of recasts and prompts in second language classrooms. *Journal of Cambridge Studies 7* (2), pp. 83-97. Association of Cambridge Studies. https://core.ac.uk/reader/35281980.

Dingemanse, M., Blythe, J., Dirksmeyer, T. (2014). Formats for other-initiation repair across languages. An exercise in pragmatic typology. *Studies in Language 38*(1), pp. 5-43. John Benjamins Publishing Company. https://www.jbe-platform.com/content/journals/10.1075/sl.38.1.01din.

Dingemanse, M., and Enfield, N. J. (2015). Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics 1*, pp. 96-118. De Gruyter https://doi.org/10.2478/opli-2014-0007.

Dörnyei, Z. (2003). Attitudes, Orientations, and Motivations in Language Learning: Advances in Theory, Research and Applications. *Language Learning: A Journal of Research in Language Studies 53*(S1), pp. 3-32. University of Michigan. https://doi.org/10.1111/1467-9922.53222.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology 23*(2), pp. 283-292. http://web.media.mit.edu/~geek/class/Duncan.pdf.

Duncan, S. Jr. and Niederehe, G. (1974). On Signalling That It's Your Turn to Speak. *Journal of experimental social psychology 10(*3), pp. 234-247. http://www.sciencedirect.com/science/article/pii/0022103174900705.

Early, M., Kendrick, M., and Potts, D. (2015). Multimodality: Out from the Margins of English Language Teaching. *TESOL Quarterly 49*(3), pp. 447-460. https://onlinelibrary.wiley.com/doi/pdf/10.1002/tesq.246.

Edwards, R. (2004). System Does Matter. *The Internet Home for Independent Role-Playing Games*. The Forge. Retrieved from http://www.indie-rpgs.com/_articles/system_does_matter.html.

Ellis, N. C., Lowes, A. L., Matheny, W. G., and Norman, D. A. (1968). Pilot Performance Transfer of Training and Degree of Simulation: III. Performance of Non-Jet Experienced Pilots Versus Simulation Fidelity. Technical Report NAVTRADEVCEN 67-C-0034-1. Naval Training Device Center, Orlando, Florida. Life Sciences. https://apps.dtic.mil/sti/pdfs/AD0675825.pdf.

Ellis, R. & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 167-192) John Benjamins Publishing Company.

Ellis, N. C. (2011). The Emergence of Language as a Complex Adaptive System. In J. Simpson (Ed.), *Routledge Handbook of Applied Linguistics* (pp. 1-30). Routledge/Taylor Francis. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.3740&rep=rep1&type=pdf.

Enfield, N. J., Dingemanse, M., Baranova, J., Blythe, J., Brown, P., Dirksmeyer, T., Drew, P., Floyd, S., Gipper, S., Gisladottir, R. S., Hoymann, G., Kendrick, K., Levinson, S. C., Magyari, L., Manrique, El, Rossi, G., San Roque, L., and Torreira F. (2013). Huh? What? – A First Survey in 21 Languages. In M. Hyashi, G. Raymond, and J. Sidnell (Eds.), *Conversational Repair and Human Understanding* (pp. 343-380). New York: Cambridge University Press. http://pubman.mpdl.mpg.de/pubman/item/escidoc:1211597:29/component/escidoc:2060688/Enfield%20et%20al.%20-%202013%20-%20Huh%20What%20%20A%20first%20survey%20in%2021%20languages.

Escudero, I., León, J. A., Perry, D., Olmos, R., and Jorge-Botana, G. (2013). Collaborative Versus Individual Learning Experiences in Virtual Education: The Effects of a Time Variable. *Procedia – Social and Behavioral Sciences 83*(4), pp. 367-370. Elsevier. https://doi.org/10.1016/j.sbspro.2013.06.072.

Farías, M., Obilinovic, K., Orrego, R. (2007). Implications of Multimodal Learning Models for foreign language teaching and learning. *Colombian Applied Linguistics Journal 9*, pp. 174-199. https://revistas.udistrital.edu.co/ojs/index.php/calj/article/view/3150/4531.

Farrington, B. (1989). 'Grandeur' or 'Servitude'?. In K. Cameron (Ed.), *Computer Assisted Language Learning: Program Structure and Principles* (pp. 67-80). Ablex Publishing Corporation.

Fägersten, K. B., Holmsten, E. and Cunningham, U. (2010). Multimodal Communication and Meta-Modal Discourse. In R. Taiwo (Ed.), *Handbook of Research on Discourse Behaviour and Digital Communication: Language Structures and Social Interaction, vol. 1*. IGI Global. http://www.irma-international.org/viewtitle/42777/.

Felicia, P. and Pitt, I. (2009). Profiling Users in Educational Games. In Connolly et al. (Eds.), *Games-Based Learning Advancements for Multi-Sensory Human-Computer Interfaces: Techniques and Effective Practices* (pp. 131-156). UK/USA: Information Science Reference.

Ferreira, A., Moore, J. D., Mellish, C. (2007). A Study of Feedback Strategies in Foreign Language Classrooms and Tutorials with Implications for Intelligent Computer-Assisted Language Learning Systems. In International Journal of Artificial Intelligence 17, pp. 382-422. IOS Press. http://hdl.handle.net/1842/4137.

Filipović, J. (2015). *Transdisciplinary Approach to Language Study: The Complexity Theory Perspective*. First Edition. Palgrave: Macmillan.

Fillmore, C. J. (1981): In H. H. Clark (1996), *Using Language*. Cambridge University Press.

Finkbeiner, C., Knierim, M. (2008). Developing L2 Strategic Competence Online. In F. Zhang. and B. Barber (Eds.), *Handbook of Research on Computer Enhanced Language Acquisition and Learning* (pp. 377-402). Information Science Reference. Yurchak Printing.

Firestone, W. A. (1993). Alternative Arguments for Generalizing from Data as Applied to Qualitative Research. In Educational Researcher 22(4), pp. 16-23. American Educational Research Association. http://www.jstor.org/stable/1177100.

Flick, U. (2018). Triangulation. In K. D. Norman and Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research*. Fifth Edition. SAGE.

Forteza, F. R. & Pastor, M. L. C. (2014). Virtual language learning environments: the standardization of evaluation. *Multidisciplinary Journal of Education, Social and Technological Sciences 1*(1), pp. 135-152. Universitat Politecnica de Valencia. https://doi.org/10.4995/muse.2014.2199.

Francoisi, S. J. (2011). A Comparison of Computer Game and Language-Learning Task Design Using Flow Theory. *CALL-EJ 12*(1), pp. 11-25. https://www.researchgate.net/publication/266172528_A_Comparison_of_Computer_Game_and_Language-Learning_Task_Design_Using_Flow_Theory.

Gamper, J. and Knapp, J. (2002). A Review of Intelligent CALL Systems. *Computer Assisted Language Learning 15*(4), pp. 329-342. https://doi.org/10.1076/call.15.4.329.8270.

Gibson, J. J. (1979): Wilson, R. A. and Folgia, L. (2011). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Fall 2011 Edition. Retrieved from: http://plato.stanford.edu/entries/embodied-cognition/.

Gísladóttir, R. S. (2015). Other-initiated repair in Icelandic. *Open Linguistics 2015 1*, pp. 309-328. De Gruyter. http://pubman.mpdl.mpg.de/pubman/item/escidoc:2095161:6/component/escidoc:2129479/opli-2015-0004.pdf.

Godhe, A.-L. and Magnusson, P. (2017). Multimodality in Language Education – Exploring the Boundaries of Digital Texts. In W. Chen et al. (Eds.), *Proceedings of the 25th International Conference on Computers in Education*, pp. 845-854. New

Zealand: Asia-Pacific Society for Computers in Education. http://hkr.diva-portal.org/smash/get/diva2:1163921/FULLTEXT02.pdf.

Goldman, A. I. (2006). *Simulating Minds. The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford University Press.

Goodwin, Ch. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics 32*, pp. 1489-1522. Elsevier. https://www.ida.liu.se/~729G12/mtrl/action_and_embodiment.pdf.

Halliday, M. A. K. (1989). *Spoken and Written Language*. Second Edition. Oxford University Press.

Hampel, L. & Hauck, R. (2006). Computer-mediated language learning: Making meaning in multimodal virtual learning spaces. *The JALT CALL Journal 2*(2), pp. 3-18. http://journal.jaltcall.org/articles/2_2_Hampel.pdf.

Hansen, T. K. (2016). The Danish Simulator: Successfully Exploring the Cost-Cutting Potential of Computer Games in Language Learning. In T. Connolly and L. Boyle (Eds.), *Proceedings of European Conference of Games Based Learning*, pp. 273-277. Reading: Academic Conferences and Publishing International Limited.

Hansen, K. T. (2012). *"The Hunt for Harald" – 3D immersive language and culture learning through gaming and speech recognition*. New Perspectives in Science Education. http://media.wix.com/ugd/d5dc38_05cc160b303c4fa8af3a416906f3daba.pdf.

Hansen, T. and Petersen, Ch. A. (2012). "The Hunt for Harald" – Learning Language and Culture through Gaming. In P. Felicia (Ed.), *Proceedings of the 6th European Conference on Game-Based Learning* (pp. 184-193). Reading: Academic Publishing International Limited.

Harless, W. G., Zier, M. A., Duncan, R. C., (1999). Virtual Dialogues with Native Speakers: The Evaluation of an Interactive Multimedia Method. *CALICO Journal 16*(3), pp. 313-337. http://www.sfu.ca/~heift/Ling480/studentprojects/charlotte.pdf.

Hasegawa, D., Cassell, J., and Araki, K. (2010). The Role of Embodiment and Perspective in Direction-Giving Systems. *Proceedings of AAAI Fall Workshop on Dialog with Robots*, pp. 26-31. http://www.justinecassell.com/publications/HasegawaAAAI2010.pdf.

Hautopp, H. (2014). *Læringsspil i andetsprogsundervisningen: En undersøgelse af Dansksimulatorens læringspotentialer i dansk som andetsprog for vosksne*. Cand.it. thesis. Aalborg Univeristy. http://docplayer.dk/461403-Laeringsspil-i-andetsprogsundervisningen.html.

Hays R. T., Singer M. J. (1989). Simulation Fidelity as an Organizing Concept. In: Hays R. T., Singer, M. J. (Eds.) *Simulation Fidelity in Training System Design. Recent Research in Psychology*. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-3564-4_3

Healey, P. G. T., Plant, N., Howes, C., and Lavelle, M. (2015). When Words Fail: Collaborative Gestures During Clarification Dialogues. *2015 AAAI Spring Symposium Series: Turn-Taking and Coordination in Human-Machine Coordination*, pp. 23-29. http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10309/10103.

Heath, Ch., Hindmarch, J., Luff, P. (2010). *Video in Qualitative Research. Analysing social interaction in everyday life*. Sage.

Heathcote, W. (2012). *Acquiring English Through the Game World of Warcraft*. Examination Thesis. Malmö University. https://pdfs.semanticscholar.org/b0a7/9bfe6a6c7cfec4d5190483ed702e5e9a96cd.pdf .

Hee, E., Artstein, R., Lei, S., Cepeda, C., and Traum, D. (2017). Assessing Differences in Multimodal Grounding with Embodied and Disembodied Agents. *Proceedings of MMSYM 2017*, pp. 1-3. Germany. http://mmsym.org/wp-content/uploads/2017/10/MMSYM2017_paper_7_HeeEtAl.pdf.

Heylen, D., Kopp, S. Marsella, S. Pelachaud, C., and Vilhjálmsson, H. (2008). The Next Step Towards a Function Markup Language. In H. Prendinger et al. (Eds.), *Proceedings of the 8th International Conference on Intelligent Virtual Agents, Lecture Notes in Artificial Intelligence, 5208*, pp. 270-280. Springer. https://link.springer.com/chapter/10.1007/978-3-540-85483-8_28.

Hilmisdóttir, H. (2007). *A Sequential Analysis of 'Nú' and 'Núna' in Icelandic Conversation*. PhD dissertation. University of Helsinki. https://helda.helsinki.fi/bitstream/handle/10138/19644/asequent.pdf.

Hilmisdóttir, H. (2010). The Present Moment as an Interactional Resource: The Case of Nú and Núna in Icelandic Conversation. *Nordic Journal of Linguistics 33*(03), pp. 269–98. Cambridge University Press. https://doi.org/10.1017/S0332586510000211.

Hilmisdóttir, H. (2011). Giving a Tone of Determination: The Interactional Functions of Nú as a Tone Particle in Icelandic Conversation. *Journal of Pragmatics 43*(1), pp. 261–87. Elsevier. https://doi.org/10.1016/j.pragma.2010.07.020.

Hilmisdóttir, H. (2016). Responding to informings in Icelandic talk-in-interaction: A comparison of nú an er það. *Journal of Pragmatics 104*, pp. 133-147. Elsevier. https://doi.org/10.1016/j.pragma.2016.05.002.

Honeycutt, J. M. (2009). Dialogue Theory and Imagined Interactions. In J. M. Honeycutt (Ed.), *Imagine That: Studies in Imagined Interactions* (pp. 195-206). https://www.academia.edu/902883/DIALOGUE_THEORY_AND_IMAGINED_INTERACTIONS.

Hulme, C. and Snowling, M. J. (2013). The Interface between spoken and written language: developmental disorders. *The Philosophical Translation of the Royal Society B(369)*, pp. 1-8. Royal Society Publishing. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866425/pdf/rstb20120395.pdf.

Hult, F. M. (2003). English on the Streets of Sweeden: An Ecolinguistic View of Two Cities and a Language Policy. *Working Papers in Educational Linguistics 19*(1), pp. 43-63. Lund University. http://portal.research.lu.se/ws/files/5288126/624364.pdf.

Hymes, D. H. (1972). On Communicative Competence. In J.B. Pride and J. Holmes (Eds.), *Sociolinguistics. Selected Readings* (pp. 269-293). Part 2. Harmondsworth: Penguin.

Hymes, D. (1992). The concept of communicative competence revisited. In M. Pütz (Ed.), *Thirty Years of Linguistic Revolution. Studies in Honour of René Dirven on the Occasion of his Sixtieth Birthday* (pp. 31-58). John Benjamins Publishing.

Jacobsen, M. H. (2015). Goffman's Sociology of Everyday Life Interaction. In M. H. Jacobsen, and S. Kristiansen (Eds.), *The Social Thought of Erving Goffman* (pp. 67-84). http://www.sagepub.com/upm-data/64381_Jacobsen_Chapter_4.pdf.

Jensen, O. (2014). *EU-MIA Research Report. Dansksimulatoren. Danish Simulator*. ITCILO. http://www.eu-mia.eu/cases/dansk-simulator-danish-simulator.

Jewitt, C. (2012). *An introduction to using video for research*. National Centre for Research Methods. Working Paper 03/12, pp. 1-11. http://eprints.ncrm.ac.uk/2259/4/NCRM_workingpaper_0312.pdf.

Jewitt, C. (2013a). *Learning and communication in digital multimodal landscapes*. London: Institue of Education Press.

Jewitt, C. (2013b). Multimodality and digital technologies in the classroom. In I. de Saint-Georges and J.-J. Weber (Eds.), *Multilingualism and Multimodality: Current Challenges for Educational Studies* (pp. 141-152). Sense Publishers. https://link.springer.com/content/pdf/10.1007%2F978-94-6209-266-2.pdf.

Johnson, L. W., Marsella, S. Vilhjalmsson, H. (2004). The DARWARS Tactical Language Training System. *Proceedings of Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2004*, pp. 1-11. https://www.ru.is/faculty/hannes/publications/IITSEC2004.pdf.

Johnson, L. W., Vilhjalmsson, H., Marsella, S. (2005). Serious Games for Language Learning: How Much Game, How Much AI? In C. K. Looi et al. (Eds.), *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology* (pp. 306-313). IOS Press. https://dl.acm.org/doi/10.5555/1562524.1562569.

Johnson, W. L., Wang, N., and Wu, S. (2007). Experience with Serious Games for Learning Foreign Languages and Culture. *Proceedings of the SimTecT Conference 2007*. Brisbane, Australia. https://www.researchgate.net/publication/228938508_Experience_with_serious_games_for_learning_foreign_languages_and_cultures.

Johnson, W. L., and Wu, S. (2008). Assessing Aptitude for Learning with Serious Games for Foreign Language and Culture. *Proceedings for the 9th International Conference* *on Intelligent Tutoring Systems (ITS) 2008*, pp. 520-529. Montreal, Canada. https://www.alelo.com/wp-content/uploads/2014/06/ITS-Johnson-Wu.pdf.

Johnson, L. W. (2010). Serious Use of a Serious Game for Language Learning. *International Journal of Artificial Intelligence in Education 20(2010)*, pp. 175-195. AIED. https://doi.org/10.3233/JAI-2010-0006.

Jokinen, K. & McTear, M. (2010). Spoken Dialogue Systems. In G. Hirst (Ed.), *Series of Synthesis Lectures on Human Language Technologies*. *Lecture #5*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00204ED1V01Y200910HLT005.

Jones, E. R., Hennessy, R. T., and Deutsch, S. (Eds.). *Human Factors Aspects of Simulation*. National Academy Press. https://doi.org/10.17226/19273

Jonsson, K. Th. (2019). *Language Attitudes: English in Iceland. A qualitative study on language attitudes towards English in Iceland*. MA-thesis. University of Helsinki. https://helda.helsinki.fi/bitstream/handle/10138/300558/Jonsson_Katrin_pro_gradu_2019.pdf?sequence=2.

Jörg, T. (2011). *New Thinking in Complexity for the Social Sciences and Humanities: A Generative Transdisciplinary Approach*. Springer. Netherlands.

Kallunki, K.-P. (2016). *Learning English in World of Warcraft: Perspectives from the players*. MA-thesis. University of Oulu. http://jultika.oulu.fi/files/nbnfioulu-201603111310.pdf.

Kao, P-L. and Windeatt, S. (2014). Long-Achieving Language Learners in Self-Directed Multimedia Environmnents: Transforming Understanding. In J.-B. Son (Ed.), *Computer-Assisted Language Learning: Learners, Teachers and Tools* (pp. 1-20). Cambridge Scholars Publishing.

Kenning, M. (1990). Computer Assisted Language Learning. *Language Teaching 23*(2), pp. 67-76. https://www.cambridge.org/core/journals/language-teaching/article/computer-assisted-language-learning/71A2E204734BC1E565C1D9C0C32C65FC.

Kessler, G. (2018). Technology and the future of language teaching. *Foreign Language Annals 51*(1), pp. 205-2018. WILEY. American Council on the Teaching of Foreign Languages. https://doi.org/10.1111/flan.12318.

Kissmann, U. T. (Ed.) (2009). *Video Interaction Analysis. Methods and Methodology*. Peter Lang Internationaler Verlag der Wissenschaften.

Knoblauch, H., and Tuma, R. (2011). Videography: An interpretative approach to video-recorded micro-social interaction. In E. Margolis and L. Pauwels (Eds.), *The Sage Handbook of Visual Research Methods* (pp. 414-430). Sage. https://methods.sagepub.com/book/sage-hdbk-visual-research-methods/n22.xml.

Knoblauch, H. (2012). Introduction to the special issue of Qualitative Research: video-analysis and videography. *Qualitative Research 12*(3), pp. 251-254. Sage Publications. http://qrj.sagepub.com/content/12/3/251.full.pdf+html.

Koester, A. J. (2002). The performance of speech acts in workplace conversations and the teaching of communicative functions. *System 30*(2), pp. 167-184. Elsevier. http://dx.doi.org/10.1016/S0346-251X(02)00003-9.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., and Vilhjálmsson, H. (2006). Towards a Common Framework for Multimodal Generation: The Behaviour Markup Language. In J. Gratch et al. (Eds.), *IVA 2006, LNAI 4133* (pp. 205-217). http://www.techfak.uni-bielefeld.de/~skopp/download/BML.pdf.

Kormos, J. (2013). Sentence Production in a Second Language. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-4). Blackwell Publishing.

Korta, K. and Perry, J. (2015). Pragmatics. In E. N. Zalta (Ed.), T*he Stanford Encyclopedia of Philosophy*. Winter 2015 Edition. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=pragmatics.

Krashen, S. D. (1982). Principles and Practice in Second Language Acquisition. First Internet Edition. Pergamon Press.

Kukulska-Hulme, A. and Shield, L. (2007). An Overview of Mobile Assisted Language Learning: Can mobile devices support collaborative practice in speaking and listening? *Proceedings of EuroCALL 2007, Conference Virtual Strand, September 2007*, pp. 1-20. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.1398&rep=rep1&type=pdf.

Lakoff, G. and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenges to Western Thought*. Basic Books, New York.

Lane, H. Ch., Hays, M. J., Core, M. G., and Auerbach, D. (2013). Learning Intercultural Communication Skills With Virtual Humans: Feedback and Fidelity. In *Journal of Educational Psychology 105*(4), pp. 1026-1035. APA. https://doi.apa.org/doiLanding?doi=10.1037%2Fa0031506.

Larsen-Freeman, D. (2013). Chaos/Complexity Theory of Second Language Acquisition. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-7). Blackwell Publishing.

Leis, A., Tohei, A., Cooke, S. D. (2015). Smartphone Assisted Language Learning. *International Journal of Computer Assisted Language Learning and Teaching 5*(3), pp. 75-88. IGI Publishing Hershey. http://dl.acm.org/citation.cfm?id=2819643.

Leong, L.-M. and Ahmadi, S. M. (2017). An analysis of factors influencing learners' English-speaking skill. *International Journal of Research in English Education 2*(1), pp. 34-41. ijreeonline.com. http://ijreeonline.com/article-1-38-en.html.

Levy, M. (1997). *Computer-Assisted Language Learning: Context and Conceptualisation*. Oxford: Oxford University Press.

Lightbown, P. M., and Spada, N. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *SSLA 12*, pp. 429-448. Cambridge University Press. https://doi.org/10.1017/S0272263100009517.

Lin, T-J. and Lan, Y-J. (2015). Language Learning in Virtual Reality Environments: Past, Present, and Future. *Educational Technology & Society 18*(4), pp. 486-497. http://tell.tcsl.ntnu.edu.tw/upload/member/files/Language%20learning%20in%20virtual%20realit%20environment%20past%20present%20and%20future.pdf.

Lindaman, D. and Nolan, D. (2015). Mobile-Assisted Language Learning: Application Development Projects Within Reach for Language Teachers. *IALLT Journal 45*(1), pp. 1-22. https://doi.org/10.17161/iallt.v45i1.8547.

Littlewood, W. (1981). *Communicative Language Teaching: An Introduction*. Cambridge University Press. Paberback.

Lombardi, I. (2012). Computer games as a tool for language education. *GAME the Italian Journal of Game Studies 1*, pp. 43-52. https://www.gamejournal.it/computer-games-as-a-tool-for-language-education/.

Louwerse, M. M., Benesh, N., Hoque, M. E., Jeuniaux, P., Lewis ,G., Wu, J., Zirnstein, M. (2007). Multimodal Communication in Face-to-Face Computer-Mediated Conversations. In R. Sun and N. Miyake (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1235-1240). Mahwah, NJ: Erlbaum. http://csjarchive.cogsci.rpi.edu/proceedings/2007/docs/p1235.pdf.

Louwerse, M. M., Benesh, N., Watanabe, S., Zhang, B., Jeuniaux, P., Vargheese, D. (2009). The Multimodal Nature of Embodied Conversational Agents. In N. Taargen and H. van Rijn (Eds.), *The Annual Meeting of the Cognitive Science Society COGSCI 2009* (pp. 1459-1464). http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/320/paper320.pdf.

Lowder, M. W. and Ferreira, F. (2016). Prediction in the Processing of Repair Disfluencies. *Lang Cogn Neurosci 31*(1), pp. 73-79. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4749026/pdf/nihms680146.pdf.

Lowie, W., Verspoor, M., and van Dijk, M. (2018). The acquisition of L2 speaking: A dynamic perspective. In R. Alonso Alonso (Ed.), *Speaking in a Second Lanugage* (pp. 105-126). Amsterda: John Benjamins Publishing Company. https://benjamins.com/catalog/aals.17.05low.

Lücking, A., Ptock, S., Bergmann, K. (2011). Assessing Agreement on Segmentations by means of Staccato, the Segmentation Agreement Calculator according to Thomann. *Proceedings of the 9th international conference on Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, pp. 129-138. Springer Verlag. http://access.uoa.gr/gw2011/proceedingsFiles/GW2011_13.pdf.

Lyster, R. and Ranta, L. (1997). Corrective Feedback and Learner Uptake. *SSLA 20,* pp. 37-66. https://doi.org/10.1017/S0272263197001034.

Malone, T. W. (1980). What Makes Things Fun to Learn? A Study of Intrinsically Motivating Computer Games. *Cognitive and Instructional Sciences Series CIS-7 (SSL-80-11)*. Palo Alto Research Center. https://www.hcs64.com/files/tm%20study%20144.pdf.

Mancuso, D. S. et al. (2010). A Study of Adult Learning in a Virtual World. *Advances in Developing Human Resources 12*(6), pp. 681-699. http://adh.sagepub.com/content/12/6/681.full.pdf+html.

Manrique, E. and Enfield, N. J. (2015). Suspending the next turn as a form of repair initiation: evidence from Argentine Sign Language. *Frontiers in Psychology 6*(1326), pp. 1-21. https://doi.org/10.3389/fpsyg.2015.01326.

Manrique, E. (2016). Other-Initiated Repair in Argentine Sign Language. *Open Linguistics 2*, pp. 1-34. http://pubman.mpdl.mpg.de/pubman/item/escidoc:2263023:3/component/escidoc:2263022/Manrique_2016.pdf.

Mapson, R. (2015). Paths to politeness: Exploring how professional interpreters develop an understanding of politeness norms in British Sign Language and English. In B. Pizziconi and M. A. Locher (Eds.), *Teaching and Learning (Im)Politeness*. 22 Volume. DeGruyter Mouton.

Margolis, E and Zunjarwad, R. (2018). Visual Research. In K. D. Norman and Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research,* pp. 1039-1089. Fifth Edition. SAGE.

Massoth, C., Röder, H., Ohlenburg, H., Hessler, M., Zarbock, A., Pöpping, D. M., and Wenk, M. (2019). High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. In BMC Med Educ 19 (29), pp. 1-8. https://doi.org/10.1186/s12909-019-1464-7.

Maybury, M. and Wahlster, W. (1998). Intelligent User Interfaces: An Introduction. In M. Kaufmann (1998), (Ed.), *Readings in Intelligent User Interfaces* (pp. 1-13). http://www.wolfgang-wahlster.de/wordpress/wp-content/uploads/Introduction_to_intelligent_User_Interfaces.pdf.

Maxwell, J. A. (2012). *Qualitative Research Design. An Interactive Approach*. Third Edition. L. Bickman and D. J. Rog (Eds.). 41 Applied Social Research Methods Series. SAGE.

Mayer, R. E. (2014). *Computer Games for Learning: An Evidence-Based Approach*. MA: MIT Press, USA.

Mayer, R. E. (2019). Computer Games in Education. *Annual Review of Psychology 70*, pp. 531-549. http://www.tut.fi/gamelab/hype/Computer%20Games%20in%20Education.pdf.

Medler, B. (2009). Generations of Game Analytics, Achievements and High Scores. *Eludamos. Journal for Computer Game Culture 3*(2), pp. 177-194. http://www.eludamos.org/index.php/eludamos/article/view/vol3no2-4/127.

Meunier, F. (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths. In *RBLA 11*(2), pp. 459-477. Belo Horizonte. https://www.scielo.br/pdf/rbla/v11n2/a08v11n2.pdf.

Meyer, B. (2009). Designing serious games for foreign language education in a global perspective. In A. Méndez-Villas, J. Mesa González, J. A. Mesa González, and A. Solano Martín (Eds.), *Research, Reflections, and Innovations in Integrating ICT in Education 2* (pp. 715-719). https://vbn.aau.dk/en/publications/designing-serious-games-for-foreign-language-education-in-a-globa.

Meyer, B. (2013). Game-Based Language Learning for Pre-School Children: A Design Perspective. *The Electronic Journal of e-Learning 11*(1), pp. 39-48. Academic Publishing International. https://files.eric.ed.gov/fulltext/EJ1012870.pdf.

Mihas, E. (2017). Conversational Structures of Alto Perené (Arawak) of Peru. *Studies in Language Companion Series 181*. Johns Benjamins Publishing.

Minogue, J. & Jones, M. G. (2006). Haptics in Education: Exploring an Untapped Sensory Modality. *Review of Educational Research 76*(3), pp. 317-348. http://journals.sagepub.com/doi/pdf/10.3102/00346543076003317.

Mitchener, G. W. (2016). Spontaneous Language. In T. K. Shackelford, V. A. Weekes-Shackelford (Eds.), *Encyclopedia of Evolutionary Psychological Science*. Springer, Cham. https://doi.org/10.1007/978-3-319-16999-6.

Morie, J. F., Chance, E., Haynes, K., and Rajpurohit, D. (2012). Embodied Conversational Agent Avatars in Virtual Worlds: Making Today's Immersive Environments More Responsive to Participants. In P. Hingston (Ed.), *Believable Bots* (pp. 99-118). Berlin, Heiledlberg: Springer. https://link.springer.com/chapter/10.1007/978-3-642-32323-2_4.

Morton, H., Gunson, N., & Jack, M. (2012). Interactive Language Learning through Speech-Enabled Virtual Scenarios. *Advances in Human-Computer Interaction 2012*. Hidawi Publishing Corporation http://dx.doi.org/10.1155/2012/389523.

Müller, C. (2013). Introduction. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf (Eds.), *Body, Language, Communication: An International Handbook on Multimodality in Human Interaction*. Volume 1. De Gruyter Mouton. Germany.

Nah, F. F-H., Eschenbrenner, B., Zeng, Q., Telaprolu, R., and Sepehr, S. (2014). Flow in gaming: literature synthesis and framework development. In *International Journal of Information Systems and Management 1*(1/2), pp. 83-124. https://doi.org/10.1504/IJISAM.2014.062288

Nakahama, Y., Tyler, A., van Lier, L. (2001). Negotiation of Meaning in Conversational and Information Gap Activities: A Comparative Discourse Analysis. *TESOL Quarterly 35*(3), pp. 377-405. Teachers of English to Speakers of Other Languges, Inc. (TESOL). https://www.jstor.org/stable/3588028.

Nakamura, J. and Csikszentmihalyi, M. (2002). The Concept of Flow. In C. R. Synder and S. J. Lopez (Eds.), *Handbook of Positive Psychology,* pp. 89-105. Oxford University Press.

Nakamura, J. and Csikszentmihaly, M. (2014). The Concept of Flow. In *Flow and the Foundations of Positive Psychology. The Collected Works of Mihaly Csikszentmihaly*, pp. 239-263. Springer.

Nakano, Y., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a Model of Face-to-Face Grounding. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics – Volume 1*, pp. 553-561. Association for Computational Linguistics. from http://dx.doi.org/10.3115/1075096.1075166.

Nelson, J. R., Balass, M., Perfetti, C. (2005). Differences between written and spoken input in learning new words. *Written Language & Literacy 8*(2), pp. 25-44. John Benjamins Publishing Company. http://www.pitt.edu/~perfetti/PDF/Nelson%20WL&L.pdf.

Newgarden, K. (2015). *Skilled Linguistic Action as a Second Language Learners' Play of World of Warcraft (WoW): A Distributed View*. Doctoral Dissertation. University of Connecticut. https://opencommons.uconn.edu/cgi/viewcontent.cgi?article=7142&context=dissertations.

Newgarden, K. and Zheng, D. (2016). Recurrent languaging activities in World of Warcraft: Skilled linguistic action meets the Common European Framework of Reference. *ReCall 28*(3), pp. 274-304. https://doi.org/10.1017/S0958344016000112.

Nijholt, A. and Heylen, D. (2002). Multimodal Communication in Inhabited Virtual Environments. *International Journal of Speech Technology 5*, pp. 343-354. Kluwer Academic Publishers. https://link.springer.com/content/pdf/10.1023/A:1020913109256.pdf.

Norris, S. (2004). *Analysing Multimodal Interaction. A methodological framework*. Routledge.

Norris, S. (2013). Multimodal Communication: Overview. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-4). Blackwell Publishing.

Nunan, D. (2001). *Designing Tasks for the Communicative Classroom*. Fifteenth Printing. Cambridge University Press.

Nunan, D. (2005). *Research Methods in Language Learning*. Cambridge University Press.

O'Connell, D. C., Kowal, S. and Kaltenbacher, E. (1990). Turn-Taking: A Critical Analysis of the Research Tradition. *Journal of Psycholinguistic Research 19*(6), pp. 345-373. Springer. https://psycnet.apa.org/doi/10.1007/BF01068884.

Ogino, M. (2008). *Modified Output in Response to Clarification Requests and Second Language Learning*. Doctoral dissertation. http://researchcommons.waikato.ac.nz/bitstream/handle/10289/2657/thesis.pdf?sequence=1&isAllowed=y.

Oreström, B. (1983). Turn-taking in English conversation. In C. Schaar and J. Svartvik (Eds.), *Lund Studies in English 66*. LiberFörlag Lund.

Ólafsson, S. (2015). *When Strangers Meet. Collective Construction of Procedural Conversation in Embodied Conversational Agents*. MA-Thesis. University of Iceland. http://skemman.is/stream/get/1946/21400/49468/1/StefanOlafsson_MAthesis.pdf.

Ólafsson, S., Bédi, B., Helgadóttir, H., Vilhjálmsson, H. and Arnbjörnsdóttir, B. (2015). Starting a Conversation with Strangers: Explicit Announcement of Presence. *Proceedings of the 3rd European Symposium on Multimodal Communication*, pp. 62-68. Dublin, Ireland. http://www.ep.liu.se/ecp/105/011/ecp16105011.pdf.

Palomeque, C. and Pujolà, J.-T. (2018). Managing multimodal data in virtual world research for language learning. *ReCALL 30*(2), pp. 177-195. https://doi.org/10.1017/S0958344017000374.

Panova, I. and Lyster, R. (2002). Patterns of Corrective Feedback and Uptake in an Adult ESL Classroom. *TESOL Quarterly 36* (4), pp. 573-595. Teachers of English to Speakers of Other Languages, Inc. (TESOL). https://www.jstor.org/stable/3588241?seq=1.

Papadakis, S. (2018). The use of computer games in classroom environment. *International Journal of Teaching and Case Studies 9*(1), pp. 1-25. https://dx.doi.org/10.1504/IJTCS.2018.10011113.

Parry, R. (2010). Video-based Conversation Analysis. In I. Bourgeault, R. Dingwall, R. de Vries (Eds.), *The SAGE Handbook of Qualitative Methods in Health Research* (pp. 373-395). Sage.

Patton, M. Q. (1990). Qualitative evaluation and research methods. Second Edition. London: Sage Publications.

Pekarova, I. (2010). Sensory modalities model (vakog) application in classes at the Department of foreign languages of the Technical university of Liberec. *Journal of Social Science and Economics 2*(B), Technical University of Liberec. https://dspace.tul.cz/bitstream/handle/15240/21335/ACC_2010_2_14.pdf?sequence=1&isAllowed=y.

Pelachaud, C. and Poggi, I. (2002). Multimodal Embodied Agents. *The Knowledge Engineering Review 17*(2), pp. 181-196.

http://journals.cambridge.org/download.php?file=%2FKER%2FKER17_02%2FS0269888902000218a.pdf&code=6de5c87d5830df780e7fa9d615b280fd.

Pelachaud, C. (2009). Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B (2009) 364*, pp. 3539-3548. The Royal Society. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2781894/pdf/rstb20090186.pdf.

Pellettieri, J. (2011). Measuring Language-related Outcomes of Community-based Learning in Intermediate Spanish Courses. In *Hispania 94*(2), pp. 285-302. American Association of Teachers of Spanish and Portuguese. http://cfflc.pbworks.com/w/file/fetch/58731121/Pellettieri_2011.pdf.

Pereira, A. and Prada, R. and Paiva, A. (2014). Improving Social Presence in Human-Agent Interaction. *Proceedings of CHI'2014 - 32nd annual ACM conference on Human Factors in Computing Systems*, pp. 1449-1458. https://dl.acm.org/doi/10.1145/2556288.2557180.

Perez, R. S. (2007). Summary and Discussion. In H. O'Neil and R. S. Perez (Eds.), *Computer Games and Team and Individual Learning* (pp. 287-306). First Edition. Elsevier.

Perfect, P., White, M., Padfield, G., & Gubbels, A. (2013). Rotorcraft simulation fidelity: New methods for quantification and assessment. In *The Aeronautical Journal (1968)*, 117(1189), pp, 235-282. Cambridge University Press. doi:10.1017/S0001924000007983.

Peterson, M. (2010a). Computerized Games and Simulations in Computer-Assisted Language Learning: A Meta Research. *Simulation and Gaming 41*(1), pp. 72-93. Sage Publications. http://journals.sagepub.com/doi/pdf/10.1177/1046878109355684.

Peterson, M. (2010b). Massively multiplayer online role-playing games as arenas for second language learning. *Computer Assisted Language Learning 23*(5), pp. 429-439. Taylor and Francis Online. https://doi.org/10.1080/09588221.2010.520673.

Peterson, M. (2013). *Computer Games and Language Learning*. First edition. Palgrave Macmillan.

Poggi, I., Pelachaud, C., De Rosis, F., Carofiglio, V. and De Carolis, B. (2005). Greta. A believable embodied conversational agent. In O. Stock and M. Zancanaro (Eds.), *Multimodal Intelligent Information Presentation* (pp. 3-25). https://link.springer.com/chapter/10.1007/1-4020-3051-7_1.

Pourabdollahian, B., Taisch, M., and Kerga, E. (2012). Serious Games in Manufacturing Education: Evaluation of Learners' Engagement. *Procedia Computer Science 15 (2012)*, pp. 256-265. Elsevier. https://doi.org/10.1016/j.procs.2012.10.077.

Prada, R. and Paiva, A. (2014). Human-Agent Interaction: Challenges for Bringing Humans and Agents Together. *HAIDM – 3rd International Workshop on Human-Agent Interaction Design and Models held at AAMAS'2014 - 13th International Conference* *on Autonomous Agents and Multi-Agent Systems* https://pdfs.semanticscholar.org/ebe7/774c91eaa3faa4009a58eb3087e930c7cdd5.pdf?_ga=2.265862223.1913843449.1584061108-1841246059.1583705776.

Prensky, M. (2001). *Digital Game-Based Learning*. McGraw-Hill. New York.

Przybylski, A. K., Rigby, C. S., Ryan, R. M. (2010). A Motivational Model of Video Game Engagement. *Review of General Psychology 14*(2), pp. 154-166. American Psychological Association. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.9080&rep=rep1&type=pdf.

Purver, M., Ginzburg, J, and Healey, P. (2003). On the Means for Clarification in Dialogue. In R. Smith and J. Van Kuppevelt (Eds.), *Current and New Directions in Discourse & Dialogue* (pp. 235-255). Kluwer Academic Publishers. http://www.aclweb.org/anthology/W01-1616.

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. Doctoral Thesis. King's College University of London. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.3455&rep=rep1&type=pdf.

Quek, F., McNeill, D., Bryll, R. Duncan, S., Ma, X-F., Kirbas, C., McCullough, K. E., and Ansari, R. (2002). Multimodal Human Discourse: Gesture and Speech. *ACM Transactions on Computer-Human Interaction 9*(3), pp. 171–193. http://web.media.mit.edu/~cynthiab/Readings/Quek-p171.pdf.

Rama, P. S., Black, R. W., Van Es, E., Warschauer, M. (2012). Affordances for second language learning in World of Warcraft. *ReCall 24*(3), pp. 322-338. European Association for Computer Assisted Language Learning. https://pdfs.semanticscholar.org/bd51/a47ddb836e595349ded9ce318753747637b2.pdf.

Ramirez, U. B. (2011). The sensory modality used for learning affects grades. *Advances in Physiological Education 35*(3), pp. 270-274. American Phisiological Society https://www.ncbi.nlm.nih.gov/pubmed/21908836.

Rankin, Y., Gold, R., and Gooch, B. (2006). 3D Role-Playing Games as Language Learning Tools. In E. Gröller and L. Szirmay-Kalos (Eds.), *Eurographics 25*(3), pp. 1-6. Blackwell Publishing. https://pdfs.semanticscholar.org/273d/1b6216069f85091da1b69f9ce4d02b65c0c8.pdf.

Richards, J. C. (2006). *Communicative Language Teaching Today*. Cambridge University Press. https://www.professorjackrichards.com/wp-content/uploads/Richards-Communicative-Language.pdf.

Ridgeway, C. L. and Smith-Lovin, L. (1999). The Gender System and Interaction. *Annual Review of Sociology 25*, pp. 191-216. http://www.jstor.org/stable/pdf/223503.pdf.

Robinson, P. (2011). Task-Based Language Learning: A Review of Issues. In Peter Robinson (Ed.), *Task-Based Language Learning* (pp. 1-36). Blackwell Publishing.

Romero, F. and Carrió, M. L. (2014). Virtual language learning environments: the standardization of evaluation. Multidisciplinary Journal for Education, Social and Technological Sciences 1(1), pp. 135-152. https://doi.org/10.4995/muse.2014.2199.

Royle, K. and Colfer, S. (2010). *Computer Games and Learning – where next? The Breadth and Scope of the use of Computer Games in Education*. UK: CeDare, University of Wolverhampton.

Rudis, D. and Postic, S. (2017). Influence of video games on the acquisition of the English language. *Verbum 8*, pp. 112-128. http://www.journals.vu.lt/verbum/article/view/11354/9818.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language 50*(4), pp. 696-735. http://www.cs.columbia.edu/~julia/cs4706/Sacks_et_al_1974.pdf.

Saldana, J. (2009). *The Coding Manual for Qualitative Researchers*. Sage Publications.

Salen, K. and Zimmerman, E. (2004). *Rules of Play – Game Design Fundamentals*. Cambridge, Massachusetts London, England: MIT Press. https://gamifique.files.wordpress.com/2011/11/1-rules-of-play-game-design-fundamentals.pdf.

Saxton, M., Houston-Price, C., and Dawson, N. (2005). The prompt hypothesis: Clarification requests as corrective input for grammatical errors. The Journal of Applied Psycholinguistics 26, pp. 393-414. Retrieved from http://journals.cambridge.org/download.php?file=%2FAPS%2FAPS26_03%2FS0142716405050228a.pdf&code=95aa47648c42d97a14e1b68b258da7cd.

Schatzman, L., and Strauss, A. L. (1973). *Field research: Strategies for a natural sociology*. Englewood Cliffs, N. J.: Prentice-Hall.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. Language 53(2), pp. 361-382. Retrieved from http://www.jstor.org/stable/pdf/413107.pdf.

Schegloff, E. A. (1992). Repair after Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *AJS 97*(5), pp. 1295-1345. http://www.jstor.org/stable/pdf/2781417.pdf.

Schulze, M, & Heift, T. (2013). Intelligent CALL. In M. Thomas, H. Reinders, and M. Warschauer (Eds.) Contemporary Computer-Assisted Language Learning (pp. 247-266). Bloomsbury Publishing.

Schwienhorst, K. (2008). *Learner Autonomy and CALL environments*. Routledge.

Seropian, M. A. (2003). General concepts in full scale simulation: Getting started. Anesthesia & Analgesia, 97, 1695-1705. https://doi.org/10.1213/01.ane.0000090152.91261.d9.

Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extend can virtual reality stimulate empathy and embodied experience? In *Computers in Human Behaviour 78*, pp. 64-73. Elsevier. https://doi.org/10.1016/j.chb.2017.09.012.

Shudayfat, E., Moldoveanu, F., Moldoveanu A., Dragos, B. (2012). A 3D Virtual Learning Environment for Teaching Chemistry in High School. *Annals of DAAAM for 2012 & Proceedings of the 23rd International DAAAM Symposium 23*(1), pp. 423-428. https://www.daaam.info/Downloads/Pdfs/proceedings/proceedings_2012/0423_Shudayfatatal.pdf.

Silva, da R. L. (2014). Video games as opportunity for infromal English language learning: Theoretical considerations. *The ESPecialist 35*(2), pp. 155-169. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.878.7424&rep=rep1&type=pdf.

Skehan, P. (1996). A Framework for the Implementation of Task-Based Instruction. *Applied Linguistics 17*(1), pp. 38-62. Oxford University Press. https://academic.oup.com/applij/article-abstract/17/1/38/159436?redirectedFrom=PDF.

Skinner, B. F. (1948). *Verbal Behaviour. William James Lectures*. Harvard University Press. https://www.behavior.org/resources/595.pdf.

Skinner, J., Garg, S., Sünderhauf, N., Corke, P., Upcroft, B. and Milford, M. (2016). High-fidelity simulation for evaluating robotic vision performance. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2737-2744. IEEE. doi: 10.1109/IROS.2016.7759425.

Skipper, J. I. (2014). Echoes of the spoken past: how auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society B 369*, pp. 1-19. Royal Society Publishing. http://dx.doi.org/10.1098/rstb.2013.0297.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pp. 816-820. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf.

Stadler, S. (2013). Cross-Cultural Pragmatics. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, pp. 1-4. Blackwell Publishing.

Swain, M. (2007). Frálagstilgátan. Kenningar og rannsóknir. In A. Hauksdóttir and B. Arnbjörnsdóttir (Eds.), *Mál Málanna* (pp. 117-136). Reykjavik: Stofnun Vigdísar Finnbogadóttur í erlendum tungumálum.

Sykes, J. M., Oskoz, A., Thorne, S. L. (2008). Web 2.0, Synthetic Immersive Environmnets, and Mobile Resources for Language Education. *CALICO Journal 25*(3), pp. 528-546. https://journals.equinoxpub.com/CALICO/article/viewFile/23094/19100.

Sylvén, L. K. and Sundqvist, P. (2012). Gaming as extramural English L2 learning and L2 proficiency among young learners. ReCALL 24(03), pp. 302-321. Cambridge University Press. https://doi.org/10.1017/S095834401200016X.

Tennyson. R. D. & Breuer, K. (2002). Improving problem solving and creativity through use of complex-dynamic simulations. *Computers in Human Behaviour 18 (2002)*, pp. 650-668. Elsevier. https://doi.org/10.1016/S0747-5632(02)00022-5.

Tennyson, R. D. & Jorczak, R. (2008). A Conceptual Framework for the Empirical Study of Instructional Games. In H. O'Neil and R. S. Perez (Eds.), *Computer Games and Team and Individual Learning*. First Edition. Elsevier.

Theodórsdóttir, G. (2011). Second Language Interaction for Business and Learning. In J. K. Hall, J. Hellermann, and S. Pekarek Doehler (Eds.), *L2 Interactional Competence and Development*. SLA. Multilingual Matters.

Theodórsdóttir, G. & Eskildsen, S. W. (2011). The use of English in everyday Icelandic as a second language: establishing intersubjectivity and doing learning. *Nordisk tidsskrift for andrespråksforsking 6(*2), pp. 59-85. Fagbokforlaget.

Theodórsdóttir, G. (2018). L2 Teaching in the Wild: A Closer Look at Correction and Explanation Practices in Everyday L2 Interaction. The Modern Language Journal 102, pp. 30-45. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1111/modl.12457.

Thorne, S. L., Black, R. W., Sykes, J. M. (2009). Second Language Use, Socialization, and Learning in Internet Interest Communities and Online Gaming. *The Modern Language Journal 93 (Focus Issue)*, pp. 802-821. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4781.2009.00974.x.

Thórisson, K. R. and Jonsdottir, G. R. (2008). A granular Architecture for Dynamic Realtime Dialogue. *Intelligent Virtual Science. Lecture Notes in Computer Science 5208*, pp. 131-138. http://cadia.ru.is/wiki/_media/public%3Apublications%3Adynamic-dialog-architecture-8.pdf.

Tobias, S., Fletcher, J. D., Wind, A. P. (2014). Game-Based Learning. In J. Michael Spector., M. David Merill, Jan Elen, and M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology*, pp. 485-503. Fourth Edition. Springer.

Toth, P. D. (2013). Output-Based Instructional Approaches. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*, pp. 1-4. Blackwell Publishing.

Traum, D. R. and Allen, J. F. (1992). A “Speech Acts” Approach to Grounding in Conversation. *Proceedings of the International Conference on Spoken Language Processing 92*, pp. 137-140. http://people.ict.usc.edu/~traum/Papers/92.traum-allen.ICSLP92.pdf.

Trifonova, A. & Bonchetti, M. (2003). Where is Mobile Learning Going? In A. Rossett (Ed.), *Proceedings of E-Learn 2003-World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pp. 1794-1801. Phoenix, Arizona, USA: Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/p/12226/.

Tschichold, C. (2006). Intelligent CALL: The magnitude of the task. *Proceedings of TALN 2006*, pp. 806-814. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=D0B8B5C0B0BDE409D6743572B401CE85?doi=10.1.1.589.9639&rep=rep1&type=pdf.

Ulum, Ö. G. (2015). Listening: the ignored skill in EFL context. *International Journal of Humanities, Social Sciences and Education (IJHSSE) 2*(5), pp. 72-80. Arcjournals.org. https://files.eric.ed.gov/fulltext/ED577306.pdf.

Vallance, M., & Martin, S. (2012). Assessment and Learning in the Virtual World: Tasks, Taxonomies, and Teaching for Real. *Journal of Virtual Worlds Research 5*(2). Asian Perspectives. https://journals.tdl.org/jvwr/index.php/jvwr/article/download/6283/6037.

Vandergriff, I. (2006). Negotiating common ground in computer-mediated versus face-to-face discussions. In Language Learning & Technology 10(1), pp. 110-138. University of Hawaii at Manoa. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/44049/1/10_01_vandergriff.pdf.

Van Dijk, T. A. (2008). *Discourse and Context: Sociocognitive Approach*. Cambridge University Press.

Veer, E. (2013). Ethnographic videography and filmmaking for consumer research. In E. Bell, S. Warren and J. Schroeder (Eds.), *The Routledge Companion to Visual Organization*. Routledge.

Vigliocco, G., Perniss, P., Vinson, D. (2014). Language as a multimodal phenomenon: implications of language learning, processing and evolution. *Philosophical Transaction of the Royal Society B Biological Sciences 369*(1651), 20130292, pp. 1-7. Royal Society Publishing. https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2013.0292.

Vilhjálmsson, H., Cantelmo, N., Cassel, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., van Welberger, H., and van der Werf, R. J. (2007). The Behaviour Markup Language: Recent Developments and Challenges. In C. Pelachaud et al. (Eds.), *Intelligent Virtual Agents 2007. Lecture Notes in Artificial Intelligence 4722*, pp. 99-111. http://www.ru.is/kennarar/hannes/publications/IVA2007.pdf.

Vilhjálmsson, H. H. (2009). Representing communicative Function and Behavior in Multimodal Communication. In A. Esposito et al. (Eds.), *Multimodal Signals: Cognitive and Algorithmic Issues Lecture notes in Artificial Intelligence 5398*, pp. 47-59. http://www.ru.is/~hannes/publications/LNAICOST2009.pdf.

Vilhjálmsson, Hannes H. (2011). *Icelandic Language and Culture Training in Virtual Reykjavik*. The Icelandic Research Fund 2012. Project Grant – New proposal. Appendix B (Detailed project description), pp. 1-13.

Waldrop, M. (1992). *Complexity: The emerging science at the edge of order and chaos*. New York, NY: Simon & Schuster

Walker, N. (2014). Listening: the most difficult skill to teach. *Encuentro 23*, pp. 167-175. Encuentro. https://pdfs.semanticscholar.org/bc94/91fdea00b44d49867d4831376208b6d8dbdf.pdf.

Warschauer, M. and Healey, D. (1998). Computers and language learning: an overview. *Language Teaching 31*. Cambridge University Press. http://hstrik.ruhosting.nl/wordpress/wp-content/uploads/2013/03/Warschauer-Healey-1998.pdf.

Wattana, S. (2013). *Talking while playing: The effects of computer games on interaction and willingness to communicate in English*. Doctoral Dissertation. University of Canterbury. https://pdfs.semanticscholar.org/9068/f63cd28c93a7a68a367cf826fe10e12146ab.pdf.

Weibel, D. and Wissmath, B. (2011). Immersion in computer games: the role of spatial presence and flow. In *International Journal of Computer Games Technology 1*(6), pp. https://doi.org/10.1155/2011/282345.

Weinstein, C. E., Zimmerman, S. A. & Palmer, D. R. (1988). Assessing Learning Strategies: The Design and Development of the LASSI. In C. E. Weinstein, E. T. Goetz, P. A. Alexander (Eds.), *Learning and Study Strategies. Issues in Assessment, Instruction and Evaluation*. Academic Press: Harcourt Brace Jovanovich Publishers.

Wild, J. L. (2015). *Second language learner multimodality and linguistic development in naturalistic settings: A study of L2 learners in the Chinese street market*. MA-thesis. The University of Nottingham. https://www.teachingenglish.org.uk/sites/teacheng/files/dissertation_for_publication_university_of_nottingham.pdf.

Wilson, R. A. and Foglia, L. (2011). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Fall 2011 Edition. http://plato.stanford.edu/entries/embodied-cognition/.

Witton-Davies, G. (2010). The role of repair in oral fluency and disfluency. *Selected Papers from the 19th International Symposium on English Teaching*, pp. 119-129. Taipei: Crane Publishing.

Xu, (Linger) T., Zhang, H., and Yu, Ch. (2016). See You See Me: The Role of Eye Contact in Multimodal Human-Robot Interaction. *Trans Interact Intell Syst. 2016 May 6*(1), pp. 1-30. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5618804/pdf/nihms-776023.pdf.

Yoshida, I. (2010). Teacher vs. Student: The perspectives on cognitive skills in foreign language learning. *Proceedings of the 24th JLTANE 2010*, pp. 1-9. http://sites.williams.edu/jltane/files/2010/09/Yoshida_JLTANE2010.pdf.

Zhang, B. (2013). An Analysis of Spoken Language and Written Language and How They Affect English Language Learning and Teaching. *Journal of Language Teaching and Research* 4(4), pp. 834-838. Academy Publisher. http://www.academypublication.com/issues/past/jltr/vol04/04/24.pdf.

Zhang, Y., Song, H., Liu, X., Tang, D., Chen, Y.-e, & Zhang, X. (2017). Language learning enhanced by Massive Multiple Online Role-Playing Games (MMORPGs) and the underlying behavioral and neural mechanisms. *Frontiers in Human Neuroscience, 11, Article 95*, pp. 1-7. Frontiers in Human Neurosciece. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5332359/pdf/fnhum-11-00095.pdf.

Zheng, D., Newgarden, K., and Youg, M. F. (2012). Multimodal analysis of language learning in World of Warcraft play: Languaging as Values-realising. *ReCALL (Digital Games for Language Learning: Challenges and Opportunities) 24*(3), pp. 339-360. European Association for Computer Assisted Language Learning. https://doi.org/10.1017/S0958344012000183.

Zhonggen, Y. (2019). A Meta-Analysis of Use of Serious Games in Education over a Decade. In M. J. Katchabaw (Ed.), *International Journal of Computer Games Technology 2019*. Hindawi. https://doi.org/10.1155/2019/4797032.

Zou, B. (2008). Research design in a computer-assisted language learning study. In *International Journal of Education and Development using Information and Communication Technology (IJEDICT*) 4(3), pp. 155-165. http://ijedict.dec.uwi.edu/include/getdoc.php?id=4583&article=534&mode=pdf.

Zouari, L. and Chollet, G. (2008). Speech Transcription for Embodied Conversational Agent Animation. *Proceedings of Acoustics 08*, pp. 10521-10525. https://pdfs.semanticscholar.org/a6a1/5b281e8eaa9f8d1b975219b0877232f39e47.pdf.

Zwitserlood, I., Özyürek, A., Perniss, P. (2008). Annotation of Sign and Gesture Cross-Linguistically. *3rd Workshop on the Representation and Processing of Sign Languages*. LREC. http://ingezwitserlood.ruhosting.nl/PDF_files/publications/LREC2008_ZwitserloodOzyurekPerniss.pdf.

Internet sources:

ESOMAR. (2009). Esomar World Research Codes & Guidelines. Retrieved from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR_Codes-and-Guidelines_Passive_Data_Collection-Observation-and-Recording.pdf.

ISO/DIS 24617-2(E). (2010). Language Resource Management – Semantic annotation network – Part 2: Dialogue acts. Retrieved form http://semantic-annotation.uvt.nl/DIS24617-2.pdf.

Virtual Reality Society. (2017). Retrieved from https://www.vrs.org.uk/virtual-reality/what-is-virtual-reality.html.

Appendices

Appendix A – Questions for Learner's Expectations from *Virtual Reykjavik*

1. Which of the below areas do you want to practice the most in the Icelandic language class?
    a. Grammar, v
    b. Vocabulary
    c. Reading
    d. Writing
    e. Speaking
    f. Listening
    g. Cultural understanding
    h. Other:____________

2. In a 3D game Virtual Reykjavik, you can interact with other characters that appear there. Which of the below areas do you want to practice your Icelandic in the most while playing the game?
    a. Grammar, v
    b. Vocabulary
    c. Reading
    d. Writing
    e. Speaking
    f. Listening
    g. Cultural understanding
    h. Other:____________

3. What do you think a 3D game should contain, or have, in order to make it fun for you to learn and practice Icelandic?;

4. In your opinion, what pros and cons there are when you can learn Icelandic in a 3D game, in which you can interact with other characters only in Icelandic?;

5. How do you feel when using Icelandic in a face-to-face interaction with its native speakers?

Appendix B – The Multimodal Annotation Scheme for *Virtual Reykjavik*

Multimodal Annotation Scheme for *Virtual Reykjavik*				
Level of annotation	**Multimodal elements**	**Annotation tags**		**Description and comments**
Linguistic	Language	Icelandic		Orthographic transcription of speech
	Speech	Filled pause, fast, slow, normal, high-pitch, low-pitch, stammer, prosody, intonation		Verbal and paraverbal features including the words to be spoken (for example by a speech synthesizer), prosody information and special paralinguistic behaviours (for example filled pauses)
BML	Head	Nod		Moving the head in a single nod (down)/repeated nod (down), at a given frequency slow/fast
		Jerk		A movement of head backwards up either single or repeated
		Tilt		A movement of head (sideways) single or repeated
		Shake		A repeated of head, a waggle
		Toss		Movement of the head independent of eyes and other facial expressions
		Head-turn		Movement of the head turning at a given direction
		Leaning		Movement of the head leaning at a given angle
	Face	Emotional expressive facial tags	Happy	Different categories according to the observable speaker's behaviour; all include different interplay of facial muscles and gaze; used for describing the speaker's behaviour that is visible through various facial expressions
			Sad	
			Tense	
			Recall	
			Surprise	
			Neutral	
			Anger	
			Joy	
			Gratification	
			Irritation	
			Helplessness	
			Pondering	
			Smile	
			Laughter	
			Scowl	
	Gaze	Directed towards interlocutor		Gaze direction

		Reconnecting with a different interlocutor		
		Up		
		Down		
		Sideways to a given angle		
	Eyes	Exaggerated opening		Different eye movement at a given direction and pace of action
		Normal opening		
		Normal closing		
		Forceful closing		
		Closing-one-eye given on which side		
		Closing-two-eyes		
		Eye-blink given on left/right side		
		Blink		
	Eyebrows	Frown		Movement of eyebrows
		Raised		
	Mouth	Open		Opening and closing of mouth
		Closed		
	Lips	Corners up		Movement and direction of lips
		Corners down		
		Protruded		
		Retracted		
	Torso	Fixed		No movement of the spine and shoulder
		Left-rotation		Movement of the spine and shoulder to the left
		Right-rotation		Movement of the spine and shoulder to the right
		Forward-bend		Movement of the spine and shoulder forward and bending downwards
		Backward-bend		Movement of the spine and shoulder backwards and bending down
	Handedness	Hand	One-hand	The speaker uses one hand for gesturing
			Both hands	The speaker uses both hands for gesturing
			Bent	The speaker's hand is bent in elbow at a given angle and direction
			Straightened	The speaker's hand is straightened up at a given angle and direction
			Spiral movement	It is a complex movement of the hand spiral way at a given angle and direction
			Synchronized movement	The movement of both hands is synchronised at a given angle and direction
			Desynchronized movement	Each hand moves in a different angle and direction

			Complex movement	A very complex movement of one or both hands, needs a precise description
			Reaching at a given angle	One or both hand are involved in an act of reaching an object or person
		Palm	Up	Palm faces the sky
			Down	Palm faces the ground
			Neutral	Palm is in a neutral position beside the body
			Closed fist	Palm is closed, forming a fist
		Fingers	Spread fingers on a palm	All fingers on palm are spread apart
			Index-finger	The speaker uses only index finger for pointing, the rest of the fingers is formed in a closed palm
			Index-and-middle-finger	The speaker uses only index finger and middle finger usually for pointing, the rest of the fingers in formed in a closed palm
			Four-fingers	Only four fingers are visible, the thumb is pushed towards the inside of the palm and is not visible
			Thumb-up	Thumb-up gesture
		Accomplished hand gestures	Pointing	Speaker's hand is straightened up at a given direction and angle and uses given number of fingers (e.g. indexical-deictic gesture)
			Beat	Beat gesture of one or both hands when emphasizing something
			Iconic	Gestures that express some semantic features by similarity or homomorphism (e.g. length or weight of an object mentioned in the discourse)
			Touch motion	The speaker touchers a given object or a part of body (e.g. touching the Adam's apple)
	Body posture	Fixed		No movement detected in the body posture
		Various shifts		The body posture changes according to a given direction, precise description needed
	Locomotion	Walk		The speaker walks, precise description as to the direction and pace needed
		Run		The speaker runs, precise description of direction and pace is needed
	Hip	Fixed		No movement of hip detected

		Movement left/right		Movement of the hip toward a given direction
	Legs	Fixed		Neutral position, no particular movement
		Moving		Movement sideways/backwards/forwards
	Knee	Fixed		Neutral position, no particular movement
		Moving		Movement of the knee towards a given direction and at a given angle
	Ankle	Fixed		Neutral position, no particular movement
		Moving		Moving the ankle at a given angle
	Foot	Right/Left	Step-sideways	Movement at a given number, length and pace of steps
			Step-backwards	Movement at a given number, length and pace of steps
			Step-forwards	Movement at a given number, length and pace of steps
	Toes	Fixed		Neutral position, no particular movement
		Standing-on-tip-of-the-toes		Movement of a foot when the speaker stands on the tip of the toes
FML	Initiation	React, recognize, salut-distant, salut-close, initiate, EAP (Explicit Announcement of Presence)		Different categories that mark a start of a conversation or a turn
	Closing	Break-away, farewell		Different categories that mark closing of a conversation or a turn
	Turn-taking	Take-turn, give-turn, keep-turn, accept-turn, assign-turn, grab-turn, release-turn, request-turn, yeild-turn		Turn-management describing different actions
	Speech act	Implore, order, suggest, propose, warn, approve, appraise, recognize, disagree, agree, criticize, contradict, accept, advise, confirm, incite, refuse, question, ask, inform, request, announce, beg, greet, reject, evaluate, promise, elaborate, summarize, clarify, q&a, convince, find-plan		Different categories describing the purpose of the speaker's speech act
	Grounding	Initiate, continue, cancel, ack, request-ack, repair, clarification request (req-repair/ReqRepair), react, recognize, approach-react, giving-directions		Different categories describing the achievement of mutual understanding between speakers
	Discourse-structure	Topic, segment		Marks different segments or discourse topics

	Cognitive-process	Remember, infer, decide, idle	Speaker's observable cognitive process
	Meta-cognitive	Thinking, remembering, planning, listening	Speaker's observable meta-cognitives
	Emotion	Anger, disgust, embarrassment, fear, happiness, sadness, surprise, shame	Speaker's emotional statuses or reactions
	Time management	Stalling, pausing	Used to mark time changes in the speaker's action
	Emphasis	Speech-emphasis	When the speaker gives emphasis on speech
		Gesture-emphasis	When the speaker gives emphasis on gestures
Context	Participants	Identifier, name, gender, age, nationality, native-speaker, human-role/agent-role appearance, voice	Information about the speaker being annotated
		Felt, faked, leaked; confusion, shyness, insecurity, anger, despair, doubt, disgust, exaltation, fear, irritation, tense, stressed, joy, pain, sadness, happiness, serenity, surprise, worry, uninterested, disappointed, satisfied, neutral	Observable information about the speaker's mental/emotional state
		Interpersonal framing (e.g. empathy), relational stance (e.g. warmth)	Observable social aspect of the speaker
		Speaker-focus, object-focus	What the speaker is looking at (which people or objects)
		Where in the room: placement-mid/aside/front/back/etc., seating-posture, standing-posture, other-posture	Describe speaker's location
	Environment	Time, setting, topic, culture, language, history of interaction, weather-conditions, geographical-condition, geological-condition, political-condition	Define speaker's current environment on the basis of observable data

Appendix C – Questionnaire for the Pilot User Study

(Tick off □ the appropriate options or write an answer).

Participant:

1. Gender: □Male □Female □Other
2. Age: ______
3. Nationality: ______________________________
4. Level of Icelandic: □Beginner

   □Intermediate
   □Advanced
   □Proficient

5. Are you a permanent or temporary resident in Iceland?

   □Permanent □Temporary

6. What is the main reason you are learning Icelandic?

____

Game:

7. How did you find playing the game? (Choose one per line)

I found it…	□Enjoyable	□Neither	□Frustrating
I found it…	□Boring	□Neither	□Exciting
I found it…	□Difficult	□Neither	□Easy
I found it…	□Educational	□Neither	□Pointless

8. Other terms that describe your experience (max 3 terms):

   ____

9. Did you encounter any difficulties?

   □No
   □Yes. What was difficult?

   ____

10. How easy/difficult was it for you **to understand the agents** in the game when he/she spoke Icelandic to you? (Choose one)
□Very easy □Easy □Neither □Difficult]Very difficult

11. Did playing the game help you learn anything new about Icelandic language?

□No, it didn´t.
□Yes, I learned
____

12. Did playing the game help you learn anything new about Icelandic culture?

□No, it didn´t.
□Yes, I learned____

13. Any comments or suggestions?

____

Virtual Agent:

14. How did you perceive the agent´s behaviour regarding:

Spoken language	□Natural	□Neither	□Robotic
Facial expressions	□Natural	□Neither	□Robotic
Hand gestures	□Natural	□Neither	□Robotic
Body movement	□Natural	□Neither	□Robotic

15. Did you find the agents were… (Choose any that apply):

□Natural
□Disturbing
□Appropriate
□Robotic
□Friendly
□Inappropriate
Other terms (max 3 terms)____

16. Did you notice any particular expressions in the agent´s behaviour, e.g. particular facial expressions, hand gestures, body posture, etc.)?

□No
□Yes, it was______________________________________

17. True or false?

I think the agent did not understand what I said.	□True □False
The interaction with the agent felt natural.	□True □False
I think the agent really listened to me.	□True □False
I think that the agent understood what I said.	□True □False
The interaction with the agent(s) felt satisfying.	□True □False

18. The agent´s overall behaviour felt: (Choose any that apply)

□Natural
□Warm
□Fake
□Positive
□Disagreeable
□Spontaneous
□Negative
□Sincere
□Disinterested
□Agreeable
□Cold
□Interested