Augmenting Situated Spoken Language Interaction with Listener Gaze

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
an den Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von
Nikolina Mitev
aus Sofia

Saarbrücken, den 09. Januar 2019

Dekan:	Prof. Dr.Heinrich Schlange-Schöningen
Berichterstatter/innen:	Dr. Maria Staudte
	Prof. Dr. Alexander Koller
Tag der letzten Prüfungsleistung:	06. Juni 2019

Abstract

Collaborative task solving in a shared environment requires referential success. Human speakers follow the listener's behavior in order to monitor language comprehension (Clark, 1996). Furthermore, a natural language generation (NLG) system can exploit listener gaze to realize an effective interaction strategy by responding to it with verbal feedback in virtual environments (Garoufi, Staudte, Koller, & Crocker, 2016). We augment situated spoken language interaction with listener gaze and investigate its role in human-human and human-machine interactions. Firstly, we evaluate its impact on prediction of reference resolution using a mulitimodal corpus collection from virtual environments. Secondly, we explore if and how a human speaker uses listener gaze in an indoor guidance task, while spontaneously referring to real-world objects in a real environment. Thirdly, we consider an object identification task for assembly under system instruction. We developed a multimodal interactive system and two NLG systems that integrate listener gaze in the generation mechanisms. The NLG system *"Feedback"* reacts to gaze with verbal feedback, either underspecified or contrastive. The NLG system *"Installments"* uses gaze to incrementally refer to an object in the form of installments. Our results showed that gaze features improved the accuracy of automatic prediction of reference resolution. Further, we found that human speakers are very good at producing referring expressions, and showing listener gaze did not improve performance, but elicited more negative feedback. In contrast, we showed that an NLG system that exploits listener gaze benefits the listener's understanding. Specifically, combining a short, ambiguous instruction with contrastive feedback resulted in faster interactions compared to underspecified feedback, and even outperformed following long, unambiguous instructions. Moreover, alternating the underspecified and contrastive responses in an interleaved manner led to better engagement with the system and an effcient information uptake, and resulted in equally good performance. Somewhat surprisingly, when gaze was incorporated more indirectly in the generation procedure and used to trigger installments, the non-interactive approach that outputs an instruction all at once was more effective. However, if the spatial expression was mentioned first, referring in gaze-driven installments was as efficient as following an exhaustive instruction. In sum, we provide a proof of concept that listener gaze can effectively be used in situated human-machine interaction. An assistance system using gaze cues is more attentive and adapts to listener behavior to ensure communicative success.

Zusammenfassung

Natürliche Sprache ist unsere wichtigste Kommunikationsmethode und doch häufig vage und schwer zu deuten. Um festzustellen, ob Zuhörer ihre Äußerungen gehört und verstanden haben, beobachten Sprecher daher deren nonverbalen Verhalten (Clark, 1996). Daran kann erkannt werden, ob weitere Erklärungen nötig sind. Dieses Phänomen zeigt sich insbesondere für gesprochene zielorientierte Interaktionen, die in einem situativen Kontext eingebettet sind und in denen Effizienz wichtig ist. Dabei spielt der Blick des Zuhörers eine bedeutende Rolle, weil die Blicke ein Anzeichen des Sprachverstehens sind (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Referenzierende Ausrücke beziehen sich auf ko-präsente Objekte und beschreiben ihre Merkmale, damit der Zuhörer sie erkennen kann. Menschen können solche referenzierende Ausrücke ohne großen Aufwand planen und spontan artikulieren, weil sie den Sprachproduktionsprozess gut beherrschen. Dabei tendieren sie häufig dazu eine nicht minimale Beschreibung zu äußern, die mehr Objekteigenschaften als nötig beinhaltet (Pechmann, 1989). Besonderes häufig wird die Farbe eines Gegenstands erwähnt, obwohl sie manchmal redundant ist und dadurch eine überspezifizierte Beschreibung entsteht. Die Farbe ist eine absolute Charakteristik und wird gerne benutzt, weil man sie auf den ersten Blick wahrnehmen und daher schneller verarbeiten kann. Die Objektgröße dagegen ist eine relative Charakteristik, die von dem jeweiligen Kontext bestimmt wird. Die Blickrichtung des Zuhörers deutet darauf hin, wie ein referenzierender Ausdruck aufgelöst wird. In anderen Worten, schauen die Zuhörer auf Objekte, die sie in Betracht ziehen und als nächstes bearbeiten oder auswählen werden. Menschliche Kommunikation funktioniert meist dann gut, wenn verschiedene Informationskanäle gut synchronisiert sind. Für Maschinen hingegen stellt dies eine Herausforderung dar.

Mit der Generierung natürlicher Sprache (natural language generation NLG) befasst sich ein ganzes interdisziplinäres Forschungsgebiet. Denn diese ist für eine Reihe von Anwendungen wichtig: anfangend vom Wetterbericht bis zu intelligenten interaktiven Systemen. Aufgrund dieser verschiedenen Anwendungsbereiche haben NLG-Systeme verschiedene Ein- und Ausgabeformate und unterscheiden sich in ihrem Interaktivitätsgrad. Allerdings müssen alle NLG-Systeme folgende Teilaufgaben bewältigen: 1) Inhaltsauswahl (was gesagt werden soll), 2) Realisierung (wie es gesagt werden soll) und 3) Präsentation (ob die Ausgabe als Text oder Audio erfolgt). Bei der Konzeption eines NLG-Systems kommt es oft auf ein Kompromiss zwischen der Komplexität der Generierungsmethode und der

Laufzeit, die für die Berechnung benötigt wird, an. Je komplexer der Ansatz der Sprachgenerierung ist, desto höher ist der Rechenaufwand. Daher eignen sich für Echtzeit-Interaktionen eher weniger anspruchsvolle Techniken, während sich anspruchsvolle Techniken für Offline-Anwendungen anbieten. Interaktive Systeme, die natürliche Sprache benutzen, können multimodal gestaltet werden. So zeigten Garoufi et al. (2016) in der GIVE Umgebung, dass ein NLG-System effektiv den Zuhörerblick benutzen kann. Hier hatte das NLG-System den Zuhörer durch ein virtuelles Labyrinth geführt, indem es navigierende Anweisungen generierte. An jeder Wand in dem Labyrinth befindeten sich verschiedene Knöpfe. Das NLG-System generierte eindeutige Objektbeschreibungen, damit der Zuhörer bestimmte Knöpfe identifizieren und betätigen konnte, um mit der Aufgabe weiterzukommen. Sobald der Zuhörer ein Objekt betrachtete, wertete das System mit Hilfe von Eyetracking Technologie aus, ob das Zielobjekt im Fokus ist. In diesem Fall wurde eine Bestätigung (*"Ja, genau das!"*) geliefert, ansonsten eine Warnung (*"Nein, nicht das!"*). Dieses System wurde mit einem Basissystem verglichen und die Benutzung der Augenbewegungen hat eine signifikante Verbesserung der Erfolgsrate erzielt. Während dies zeigt, dass der Zuhörerblick nützlich sein kann, ist allerdings noch unklar, ob und wie menschliche Sprecher ihre verbalen Anweisungen in der "realen Welt" anpassen, wenn ihnen der Zuhörerblick zur Verfügung steht. Weiterhin ist unklar, ob NLG-Systeme in realen Setups den Zuhörerblick nutzen. Das Ziel dieser Dissertation ist es, diese Forschungsfragen zu untersuchen. Dafür wurden das Zusammenspiel von Sprache und Augenbewegungen in verschiedenen Mensch-Mensch und Mensch-Maschine Setups betrachtet. Diese Arbeit addressiert drei Szenarien, in denen die Rolle des Zuhörerblicks untersucht wurde.

Zuerst betrachten wir ein virtuelles Scenario und die Aufgabe, automatisch die Auflösung eines referenzierenden Ausducks vorherzusagen. Engonopoulos, Villalba, Titov, and Koller (2013) entwickelten dafür zwei probabilistische Modelle, machten sich Maschinelles Lernen zunutze und evaluierten diese in der GIVE Umgebung. Das erste Model wertet den linguistischen Kontext und das zweite Model den visuellen Kontext aus. Da beide Modelle komplämentäre Informationen verarbeiteten, war die beste Akkuratheit mit der Kombination von beiden erzielt. Wir erweiterten das zweite probabilistische Model, so dass die Blickbewegungen des Zuhörers berücksichtigt werden, z.B. wie oft ein Objekt angeschaut wurde. Dafür wurden Eyetracking Features entwickelt. Außerdem wurde die Trainigs- und Testmethode angepasst, um sequentielle Daten mit der `10-cross-fold-validation` Methode zu testen. Es zeigte sich, dass das Blickverhalten an sich nicht ausreicht, um eine sehr gute Genauigkeit zu erreichen. Die Kombination aus Blickbewegungen und den Features des Basismodells, die Salienz und Distanz zum Zielobjekt berechnen, führt jedoch zu

einer Verbesserung der Vorhersagegenauigkeit. Dies gilt insbesondere für unübersichtliche visuelle Kontexte mit mehreren Objekten, welche die gleichen Eigenschaften aufweisen und daher schlecht differenziert werden können. Diese zusätzliche Information ist zum früheren Zeitpunkt in der Interaktion aufschussreich, was wichtig für Interaktionen in Echtzeit ist, damit die Sprachausgabe rechtzeitig (z.B. bevor nach einem Objekt gegriffen wird) angepasst werden kann.

Das zweite Szenario ist "Schatzsuche" – Navigation in Gebäuden: Ein Feldexperiment mit zwei Teilnehmern, das sich mit der Frage befasst, ob die Verfügbarkeit des Zuhörerblicks zu einer besseren Mensch-Mensch Interaktion beiträgt. In diser explorativen Studie wurde das Zusammenspiel von Augenbewegungen und Sprache untersucht und dabei die Sichtbarkeit des Zuhörerblicks für den Sprecher manipuliert. Jedem Teilnehmer wurde eine Rolle zugeteilt, entweder als Sprecher, der spontan Richtungen angibt und identifizeirende Anweisungen produziert oder als Zuhörer, der die Anweisungen folgt, herumläuft und bestimmte Objekte nötig für neun unterschiedliche Alltagsszenarien, wie beispelsweise Brief schreiben, identifiziert. Es wurde erwartet, dass die Verfügbarkeit des Zuhörerblicks zu kürzeren Interaktionszeiten und mehr deiktischen Ausdrücken führt, weil Sprecher sehen, worauf gerade die Zuhörer schauen. Die Ergebnisse zeigen jedoch, dass der Zuhörerblick keinen Einfluss auf die Performanz hatte aber die Sprecher tendierten dazu, mehr negatives Feedback zu äußern, wenn sie ihn gesehen hatten, sie versuchten also, Missverständnisse zu verhindern.

Unser drittes Szenario is Modellbau unter Systemanweisungen. Der Kernaspekt dieser Arbeit ist die Entwicklung eines mulimodalen Assistenzsystems mit den zwei darin eingebetteten NLG-Systemen, "Feedback" und "Installments". Beide Systeme generieren automatisch identifizierende Anweisungen. Drei empirischen Studien wurden durchgeführt, die weitere wissenschaftliche Evidenz für die Nützlichkeit des Zuhörerblicks und seine Integrität in Interaktiven NLG Systemen liefern. Das Assistenzsystem berücksichtigt die Augenbewegungen des Zuhörers mithilfe eines mobilen Blickbewegungsmessers (Eye-Tracker) in einer realen Umgebung, um die Sprachausgabe in Echtzeit anzupassen. Um eine reale Szene in ein virtuelles Modell umzuwandeln und ermitteln zu können, wohin ein Zuhörer schaut, wurde die Technik der erweiterten Realität (Augmented Reality) verwendet. Durch diese Realisierung ist das Assistenzsystem aufmerksam und adaptiv hinsichtlich des Nutzerverhaltens. Da der Zuhörerblick ein zuverlässiger Hinweis auf Sprachverstehen ist, kann er dazu beitragen Mensch-Maschinen Interaktionen effektiver und angenehmer zu gestalten. Gleichzeitig ist das Augenbewegungssignal kontinuierlich,

sehr rapide, dynamisch und individuell. Aus diesen Gründen ist es nicht trivial die Augenbewegungen im Sprachkontext zu interpretieren. Um ein Blickbewegungssignal richtig zu deuten, braucht man das Wissen über die zugrundeliegenden Verarbeitungsprozesse. Zunächst werden sogenannte Eye-Tracking-Events extrahiert: Fixationen weisen auf das Betrachten eines Objekts hin; Sakkaden sind dagegen schnelle Bewegungen beider Augen, die einen neuen Fixationspunk erfassen. Diese Arbeit bezieht sich auf Inspektionen von Objekten, also längere Fixationen, deren Schwellenwert je nach Setup angepasst werden kann. Anhand dieser Inspektionen von Objekten kann das entwickelte System feststellen, ob eine Anweisung richtig verstanden wurde. Das erste NLG-System "Feedback" generiert entweder kurze, mehrdeutige oder lange, ausführliche Anweisungen und reagiert auf Objektinspektionen mit verbalem Feedback. Das Feedback hat unterschiedliche Spezifität, nämlich warnend und unspezifisch (z.B. "Nein, nicht das!") oder informativ und kontrastiv, indem die Position des Zielobjekts relativ zu dem jetzigen Fixationspunkt berechnet und als Richtungsanweisung ausgegeben wird (z.B. *"Weiter links!"*). Diese weiterführende Information soll die Suche eingrenzen und durch eine resultierende verkürzte Interaktionszeit eine effizientere Interaktion realisieren. Das zweite NLG System "Installments" implementiert die inkrementelle Generierung von identifizierenden Anweisungen und gibt eine Objektbeschreibung in aufeinanderfolgenden Phrasen aus (so genannte Installments). Jede Phrase liefert eine Teilbeschreibung des Zielobjektes, wobei es in Abhängigkeit des Blickverhaltens zu einer unterschiedlichen Anzahl der Installments in der Systemausgabe kommt. Des Weiteren kann das System eine lange, vollständige Objektbeschreibung generieren, indem es alle Phrasen aneinander zusammenfügt. Jedes System bietet zwei unterschiedliche Interaktionsstile an: passiv/nicht interaktiv gegenüber interaktiv. In der ersten empirischen NLG-Studie wurden Versuchspersonen eingeladen, mit dem NLG-System "Feedback" zu interagieren. Es wurde zum einen untersucht: ob eine mehrdeutige Anweisung kombiniert mit blickgesteuertem Feedback in realer Umgebung effizienter als eine ausführliche Anweisung sein kann. Witerhin wurde geprüft, wie sich die Spezifizität des Feedbacks auf die Performanz auswirkt. Dafür wurden zwei Gruppen getestet bei denen der Interaktionsstil als Within-Subject-Faktor und die Feedbackspezifizität als Between-Subject-Factor manipuliert wurden. Die Versuchspersonen wurden instruiert die Systemanweisungen zu befolgen und möglichst präzise bestimmte LEGO Duploteile zu identifizieren und mit denen ein kreatives Model zusammenzubauen. Die Ergebnisse zeigen, dass die Versuchspersonen in der ersten Gruppe signifikant schneller waren, wenn sie eine ausführliche Objektbeschreibung gehört haben als wenn sie eine

mehrdeutige Objektbeschreibung mit unspezifischem Feedback erhalten haben. Die Richtung dieses Haupteffekts wändete sich in der zweiten Gruppe: Eine unspezifische Anweisung mit informativem Feedback konnte eine ausführliche Anweisung übertreffen. In der zweiten empirischen NLG-Studie wurde die Rolle der Feedbackspezifizität näher betrachtet und diese als Within-Subject-Faktor abwechselnd für jede Anweisung manipuliert. Interessanterweise war die Kombination aus einer mehrdeutigen Beschreibung mit unspezifischem Feedback nun nicht mehr benachteiligt. Die Erwartungshaltung hinsichtlich der Systemfähigkeiten scheint für die Performanz ausschlaggebend zu sein. Wenn die Benutzer eine informative weiterführende Information erwarten, können sie eine nicht sonderlich informative Information besser verarbeiten und ebenso schnelle Interaktionszeiten erreichen. In der dritten empirischen NLG-Studie wurden die Versuchspersonen dazu eingeladen, mit dem NLG-System "Installments" zu interagieren. Es wurde untersucht, welcher Ansatz der Informationslieferung (Installments vs. NoInstallments) effizienter ist. Um eine inkrementelle Objektbeschreibung zu generieren, reagiert das NLG-System "Installments" auf Augenbewegungen eher indirekt (also nicht relativ zu dem Fixationspunkt), um die nächste Phrase auszulösen. Darüber hinaus variert die Informationsanordnung über die Position des Zielobjekts und über die Objekteigenschaften. Beide experimentellen Faktoren wurden als Within-Subject-Faktor manipuliert. Die Datenauswertung zeigt, dass hier die lange Variante schneller ans Ziel geführt hat als die schrittweise ausgegebene Beschreibung. Allerdings war die inkrementelle Variante genauso effizient wie die lange, ausführliche Variante, wenn die Objektposition am Anfang der Anweisung spezifiziert war, weil auf diese Weise der Suchraum von vornherein eingeschränkt war. Interessanterweise generierte das NLG-System in disem Fall mehr Installments, d.h. es liefert mehr Teilobjektbeschreibungen, als wenn die Position an der zweiten Stelle erschien. Das lässt sich durch die Tatsache erklären, dass die Zuhörer konkurrierende Objekte betrachteten, um beispielsweise die Größe eines Objekts in Relation zueinander zu setzen.

Zusammenfassend hat diese Arbeit die Rolle des Zuhörerblicks aus verschiedenen Blickwinkeln betrachtet und diese Modalität in Mensch-Mensch und Mensch-Maschine Interaktion integriert. Diese Integration war für menschliche Sprecher während der Sprachproduktion schwer zu interpretieren, weshalb sie sich nicht auf die sprachlichen Ausdrucken auswirkte. Die Information über das Blickverhalten des Zuhörers hat jedoch die automatische Vorhersage der Referenzauflösung verbessert. Des Weiteren liefert die vorliegende Arbeit den Wirksamkeitsnachweis, dass ein Assistenzsystem, welches identifizierende Anweisungen in natürlicher Sprache automatisch generiert, dem Zuhörerblick effektiv nutzen kann. Dies minimiert die Fehlerrate beim Greifen von Objekten und kann zu schnelleren

Interaktionszeiten führen. Der Zuhörerblick erweist sich auch hier als ein verlässliches Anzeichen des Sprachverstehens, welches eine positive Auswirkung auf Mensch-Maschine Interaktion hat.

Dedicated to my family.

“Когато сутрин се събудя и видя вашите лица ме обгръща най-могъщата сила на света и това е любовта!
Вие ме дарявате с енергия за стремеж и вдъхновение за духовен разтеж.”

Acknowledgements

First, I wood like to express my deepest gratitude to my supervisor Dr. Maria Staudte for the unique opportunity to pursue a PhD in the independent research group "Embodied Spoken Interaction" at Saarland University. Thank you for the excellent and tireless supervision, your patience, constructive critique, actively co-writing, letting me work on a topic I was interested in and encouraging me even when things did not go as expected.

Further, my special thanks goes to Prof. Dr. Alexander Koller and Martín Villalba. It has been a pleasure to collaborate with you! I am grateful for the productive discussions regarding interesting research directions and for providing me with hands on experience in the GIVE environment and corpus collections. For the latter thanks to Nikos Engonopoulos, as well.

I appreciate a very nice and productive collaboration with Dr. Thies Pfeiffer and Patrick Renner at Bielefeld University. Thank you for providing the EyeSee3D software and technical support with the mobile eye-tracking glasses, which enabled building an NLG system for real environment. Additionally, I would like to thank Dr. Sébastian Le Maguer for the support with connecting the Mary TTS system to enable auditive output.

I would like to extend my thanks to my colleagues Christine Ankener, Torsten Kai Jachmann and Mirjana Sekicki. I was pleased to work and brainstorm together with you. Danke für die morgentliche Unterhaltung und hilfreiche Ablenkung, wenn ich unter Stress war.

Moreover, I owe many thanks to Margaret De Lap for proofreading this thesis. Many thanks go to the MMCI Cluster office for the brilliant administrative management.

I am particularly grateful to my relatives and friends! Благодаря ви за хубавите задружни моменти и за това, че ме приемате такава каквато съм. Danke für die schönen Momente und aufmunternde Gespräche.

My sincerest and warmest thanks goes to my partner in life Georgi. Без твоята подкрепа нямаше да успея да се справя, благодаря ти за безусловната подръжка и за това, че винаги вярваш в мен!

Contents

Abstract iii

Zusammenfassung v

Acknowledgements xiii

Contents xv

List of Figures xix

List of Tables xxiii

1 Introduction 1
- 1.1 Motivation and Context . . . 1
- 1.2 Research Aims and Contributions . . . 5

2 Background 9
- 2.1 Natural Language Generation (NLG) . . . 9
- 2.2 Listener Gaze in Task-oriented Interaction . . . 14

3 Automated Prediction of Reference Resolution 17
- 3.1 Problem Definition . . . 18
- 3.2 Episodes and Feature Functions . . . 19
- 3.3 Prediction Models . . . 21
- 3.4 Dataset . . . 22
- 3.5 Evaluation and Results . . . 23
- 3.6 Discussion . . . 25

4 Listener Gaze in Human-Human Interaction 29
- 4.1 Method . . . 31

4.2 Results . . . 37
4.2.1 Performance . . . 37
4.2.2 Linguistic Analysis . . . 37
4.2.3 Visual Behavior Analysis . . . 43
4.3 Discussion . . . 45

5 GazInG: Gaze-driven Instruction Generation 49
5.1 Use Case and Task . . . 51
5.2 Gaze-sensitive Instruction Generation in the Real World . . . 52
5.2.1 GazInG: Multimodal Interactive System . . . 52
5.2.2 NLG System *"Feedback"*: Instructions Combined with Gaze-driven Verbal Feedback . . . 54
5.2.3 NLG System *"Installments"*: Gaze-driven Incremental Instruction Generation . . . 57
5.3 Summary . . . 60

6 Human-Machine Interaction: Effects of Gaze-driven Feedback 63
6.1 Experimental Method . . . 64
6.1.1 Setup and Apparatus . . . 64
6.1.2 Measures and Analysis . . . 65
6.2 Experiment 1: Interaction with the NLG System *"Feedback"* . . . 67
6.2.1 Participants . . . 68
6.2.2 Procedure . . . 68
6.2.3 Results . . . 71
6.2.4 Discussion . . . 79
6.3 Experiment 2: Interaction with the NLG System *"Feedback"* . . . 80
6.3.1 Participants . . . 80
6.3.2 Procedure . . . 81
6.3.3 Results . . . 81
6.3.4 Discussion . . . 85

7 Human-Machine Interaction: Listener Gaze for Incremental NLG 87
7.1 Experiment 3: Interaction with the NLG System *"Installments"* . . . 88
7.1.1 Method . . . 89
7.1.2 Results . . . 91
7.2 Discussion . . . 96

8 Conclusion 99
8.1 Summary . . . 99
8.2 Discussion . . . 102

A Micro Tasks used in the indoor guidance study **109**

B GazInG: Preliminary Studies **111**
B.1 **Object density** 111
B.2 **Timing and Usefulness of Gaze-based Feedback** 113

C Scene Layouts for Assembly **117**
C.1 Scene Layout 1 and Scene Layout 2 used in Experiment 1, 2 and 3 117
C.2 Scene Layout 3 and Scene Layout 4 used in Experiment 2 and 3 118

D Questionnaires für Experiment 1 **121**
D.1 Assessment of the Interaction 121
D.2 Comparison of the Two Interaction Strategies 121

E Questionnaire für Experiment 3 **123**
E.1 Assessment of the Interaction 123
E.2 Comparison of the Two Interaction Strategies 123

F Human-machine interaction Model Selection Results **125**

Bibliography **129**

List of Figures

1.1 Example visual contexts, where a deictic expression like "this" is resolved either to "a shelf panel" or most probably to "spinach". 2

3.1 The processing pipeline for automatic prediction of reference resolution. . . 18
3.2 The structure of the interactions. 20
3.3 The GIVE corpus: Example visual context of an easy (left picture) and a hard (right picture) referential scene in the virtual environment. 24
3.4 Accuracy as a function of training and testing time. 25

4.1 The experimental setup: the walker wearing a head-mounted eye tracker, following instructions and selecting objects (left picture) and the remote instructor giving directions, describing the next target and monitoring the walker's behavior (right picture) 31
4.2 The task completion time (log transformed with 95% CI error bars) for the individual trials in the micro task 38
4.3 The proportion of positive and negative feedback instances in the different conditions. The model fitted to that data is the following `feedbackType` $\sim$ `GazeAvailability + (1|Pair)` 39
4.4 The annotation scheme for categorizing referring expressions. 40
4.5 The average number of *sub-specific* referring expressions per trial in the micro task (95% CI error bars) 42

5.1 The workspace comprises 20 composed objects spread on a table (left picture); and a close-up view of a composed target object (right picture). . . . 51
5.2 This diagram depicts the modular software architecture of the GazInG system. 52
5.3 The generation mechnism implemented in NLG system *"Feedback"*. 55
5.4 The generation mechanism implemented in NLG system *"Installments"*. . . 59

6.1 Setup: Listener in front of a workspace before any objects are collected. The target is circled in green and competitors in red. The listener inspects the competitor object to the left as highlighted in the virtual 3D model. EyeSee3D is used to reconstruct the gaze ray in 3D (yellow). The target domain is modeled as a 3D situation model with boxes as proxies for the assembled structures (turquoise). 65

6.2 This diagram illustrates the interaction phases for both strategies: The spoken **instruction**, followed by **identification**, i.e. time to first target inspection, and the **grasp** of the object after a verbal confirmation is given. Visual search starts either during or after an instruction and can be interleaved with feedback, depending on the condition. 66
6.3 An example trial: System instructs the user by saying "Pick the big red building block." The listener identifies and grasps it. After that it is assembled to the other LEGO blocks (right picture). The circle represents the gaze cursor. 69
6.4 This plot depicts the task completion time (log transformed) from the instruction onset until the target is grasped in Experiment 1. 71
6.5 This plot depicts the time span from the instruction offset to the first target inspection in Experiment 1. 74
6.6 This plot depicts the number of negative feedback occurrences in Experiment 1. 75
6.7 This figure illustrates how a typical trial looks and the differentiation of *FeedbackSpecificity*. The red arrows indicate the time intervals analyzed for the sequential feedback analysis. 76
6.8 This plot depicts the time interval from the instruction offset to the onsets of the first negative (triggered by a competitor inspection) and first positive (triggered by a target inspection) feedback instances for the AMBIGUOUS *InteractionStrategy* in Experiment 1. 77
6.9 This plot depicts participants' perception and judgement of the interaction flow measured on a Likert scale for Experiment 1. 78
6.10 The task completion time measured in interactions obtained in Experiment 1 (left plot) and in Experiment 2 (right plot). 81
6.11 This plot depicts the time span from the instruction offset to the first target inspection in Experiment 1 (left plot) and in Experiment 2 (right plot). . . 83
6.12 This plot depicts the number of negative feedback occurrences in Experiment 2. 84
6.13 This plot depicts the time interval from the instruction offset to the onsets of the first negative and first positive feedback instances in Experiment 2. . 85

7.1 This diagram illustrates the interaction phases for both information delivery approaches NOINSTALLMENTS and INSTALLMENTS 90
7.2 This plot depicts the task completion time in Experiment 3. 91
7.3 This plot depicts the time interval from instruction onset to first target inspection in Experiment 3. 93
7.4 This plot depicts the time interval from instruction onset to first target inspection in Experiment 3. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 94
7.5 This plot depicts participants' perception and judgment of the interaction flow measured on a Likert scale for Experiment 3. 95

B.1 The GazInG setup: full (left picture) vs. reduced (right picture) referential scene. . . . 112
B.2 How natural did you find the spoken system instructions? . . . 113
B.3 How precise did you find the spoken system instructions? . . . 114
B.4 How adequate was the system's feedback? . . . 114
B.5 I think that the timing of the system's feedback was appropriate. . . . 115
B.6 Without the system's feedback I would not be able to find the right building blocks. . . . 115
B.7 Because the system reacted to my eye movements, it was easier for me to find the building blocks. . . . 115
B.8 The instructions did not contain enough information such that the system's feedback was crucial. . . . 116

C.1 First scene layout . . . 117
C.2 Second scene layout . . . 118
C.3 Third scene layout . . . 118
C.4 Fourth scene layout . . . 119

List of Tables

4.1 Features extracted from human visual behavior. 36

6.1 Interaction strategies (blocked) for each group in Experiment 1. 68
6.2 This table summarizes the number of trials remaining after outlier removal and in how many of them no wrong objects were grasped (presented in brackets). 71
6.3 This table summarizes the models fitted to the performance data and the model comparison results for Experiment 1. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 73
6.4 The mean durations in seconds of the interaction phases in Experiment 1 (see Figure 6.2). 73
6.5 This table summarizes the models fitted to the performance data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 82
6.6 Mean durations in seconds of the three interaction phases and the total time for Experiment 2 as depicted in Fig. 6.2. 83
6.7 This table summarizes the models fitted to the listener gaze data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 84

7.1 The design of Experiment 3. 90
7.2 This table summarizes the models fitted to the performance data and the model comparison results for Experiment 3. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 92

F.1 This table summarizes the models fitted to the time for identification data and the model comparison results of listener gaze behavior for Experiment 1. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 125
F.2 This table summarizes the model fitted to the feedback data and inferential statistics for Experiment 1.Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 126
F.3 This table summarizes the models fitted to the listener gaze data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 126

F.4 This table summarizes the models fitted to the performance data and the model comparison results. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 126
F.5 This table summarizes analysis and the model fitted to the speech data. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. 127

Chapter 1

Introduction

1.1 Motivation and Context

Natural language is our usual mean of communication. However, sometimes it can be difficult to interpret language without further cues, specifically in spoken interaction when referring to co-present objects. That is, the same linguistic expression can be resolved to different entities depending on the current situation and visual context. Consider, for instance, the utterance *"The next thing you need is this!"*: it contains a deictic expression *"this"*, which is unspecific and can be resolved to different entities depending on the visual context. In Figure 1.1 two sample scenarios are presented. In the situation depicted in the left picture the expression will be resolved to *"a shelf panel"* as opposed to the right, where it will be resolved most probably to *"spinach"*. Additionally, in the left picture there are multiple shelf panels and so other non-verbal cues like gestures and gaze play a very important role: They can be used to disambiguate an expression and thereby facilitate referential success. Specifically, the eye movements of the human interlocutors indicate their intentions. That is, human speakers look at co-present objects they are about to mention (Griffin & Bock, 2000) and listeners' eye movements mirror language comprehension, meaning that listeners inspect relevant objects matching a description (Tanenhaus et al., 1995). It often happens that such an ambiguous referring expression can be misunderstood. Depending on the context and task under consideration, a misunderstanding can have different consequences. In the cooking scenario, putting the ingredients in a


FIGURE 1.1: Example visual contexts, where a deictic expression like "this" is resolved either to "a shelf panel" or most probably to "spinach".

different order might affect the taste of the meal. On the other hand, in the assembly scenario, if the listener misunderstands a referring expression and grasps an incorrect object, which cannot be assembled into the available construction, then she should put it back and search further for the suitable one. This leads to longer interaction time because the speaker has to clarify, for example, by giving a more specific description. Thus, avoiding misunderstandings saves time and results in more efficient interaction. For this reason speakers monitor visual behavior and adapt their utterances to it, for example by providing verbal feedback such as *"No, I don't mean that!"* In other words, situated interaction involves various modalities to achieve communicative success because it takes place in a shared (physical) environment, which is particularly important for goal-oriented scenarios such as collaborative assembly, where a mutual goal has to be achieved. Although human interlocutors can align and interpret different modalities intuitively, this poses a challenge for assistance systems. Such systems aim to support a user in collaboratively solving a task and need to process multimodal cues automatically in order to be attentive to changes in the environment and adaptive to the user's behavior. But interpreting the numerious cues is not trivial and sometimes impossible even for humans.

In the following we identify three different scenarios touching upon three adjacent research areas before we formulate the research questions that we have tackeled in this thesis (in Section 1.2).

Human-human interaction In collaborative task solving, *grounding* is crucial to achieve communicative success. Grounding is the process of interlocutors' validation of each other's mental models. In other words, speakers observe listeners to detect if their communication message was received and understood (Clark, 1996). Specifically, they

monitor listener's understanding and the mapping of a meaning to the world by considering listener gaze (Clark & Krych, 2004; Brown-Schmidt, 2012). Gaze has been shown to be a reliable indicator of reference resolution (Cooper, 1974) because listeners look at objects they believe are being referred to by the speaker (Tanenhaus et al., 1995; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). Importantly, such gaze cues are closely time-locked to a referring expression (Allopenna, Magnuson, & Tanenhaus, 1998). Most of this evidence is based on very controlled laboratory settings using predefined utterances and simple visual scenes. In more dynamic setting that involve two interlocutors, the role of listener gaze is typically studied in face-to-face interactions and not interpreted from the egocentric perspective of the listener. A considerable exception is the work by Brown-Schmidt and Tanenhaus (2008), who present two experiments, in which they monitored gaze and speech, while pairs of naïve interlocutors engaged in a referential communication task. Their results demonstrate that gaze can be used to examine real-time processing during free interactive conversation. In another explorative study, Brennan, Schuhmann, and Batres (2013) investigated communication in the wild for outdoor navigation. They examined referring expressions and lexical choice during remote pedestrian guidance with human interlocutors, and reported that there is a strong degree of lexical entrainment and that the efficiency is affected by the direction giver's spatial ability. However, they did not take listeners' eye movements into account. Inspired by their setting, we identify our first scenario and design an experiment to investigate the interplay of spontaneous spoken instructions and listener gaze in an indoor guidance task.

Human-machine interaction An artificial speaker, also known as natural language generation (NLG) system, is capable of automatically planning and creating sentences, instructions or discourse from a machine representation. An NLG system can assist a user to solve a task collaboratively, as was proposed in the GIVE challenge (Koller, Striegnitz, Byron, et al., 2010). The effective use of listener gaze as an index of understanding has been shown to improve human-machine interaction in virtual environments (Koller, Staudte, Garoufi, & Crocker, 2012; Staudte, Koller, Garoufi, & Crocker, 2012; Garoufi et al., 2016). There, an interactive NLG system guided a human listener through a virtual maze, referred to specific buttons to be pressed and provided gaze-based feedback on button inspections. Importantly, interacting with the gaze-sensitive system resulted in better performance (lower error rate) than interacting with a baseline system that did not consider listener's eye movements and did not give feedback. An offline study

by Engonopoulos et al. (2013) on a corpus collection from the GIVE challenge investigated the problem of automatically predicting how a referring expression (RE) will be resolved. They achieved accurate prediction of reference resolution by combining two probabilistic log–linear models: a semantic model, evaluating the semantics of a given instruction, and an observational model, evaluating listeners' behavior. Notably, the best accuracy of the observational model was measured in a relatively late stage of the interaction. Similar observations are reported by Kennington and Schlangen (2014), who compared listener gaze and an incremental update model as predictors for the reference resolution. However, Engonopoulos et al. (2013) did not consider listener gaze. Thus we take this setup to be our second scenario for investigating the usefulness of listener gaze and augment the observational model with eye-tracking features to capture listener's attention.

Collaborative Assembly In contrast to human-machine interaction, where the pleasantness of the interaction and the politeness of a system are important aspects, the area of collaborative assembly focuses mostly on efficiency and does not necessarily use natural language to communicate which object is needed. There is a large body of work on assembly tasks in virtual and real setups, but less has been done to investigate the role of listener gaze in such scenarios. For example, Kopp, Jung, Leßmann, and Wachsmuth (2003) examined interactive assembly using a virtual agent. The agent is capable of instructing a human listener on how to build a pre-defined model. If it recognizes a failure, then the agent informs the listener about it and the virtual agent undoes the wrong step. Handling such errors takes additional effort and time. This required multimodal reference resolution in a dynamic virtual environment (Pfeiffer & Latoschik, 2004). Another study by Kirk, Rodden, and Fraser (2007) considered the role of remote gestures in human-human assembly tasks and showed that gestures offer positive benefits for collaborative performance. Neither study looked at the role of listener gaze, as they focused on other modalities. In contrast, Sakita, Ogawara, Murakami, Kawamura, and Ikeuchi (2004) considered human-robot collaboration using non-verbal cues and proposed a more flexible task management strategy for a LEGO assembly task by allowing a free choice the next assembly step. Tracking the assembler gaze to predict the next action allowed simultaneous assembly, which led to an efficiency gain. Further, Fischer et al. (2015) investigated social gaze in human-robot interaction for assembly, and demonstrated its importance for quicker engagement with the robot and feeling more responsible for the task performance. However, the exact temporal alignment of the user's object-directed

gaze with spoken instructions has not been assessed. This could be beneficial because gaze is an early indicator of the listener's intentions (Altmann & Kamide, 1999). An exception is the work by Fang, Doering, and Chai (2015), who proposed a collaborative referring expression generation algorithm for situated human-robot interaction. They focused more on embodiment and the robot's gestures, but also incorporated listener gaze to refer to objects incrementally in installments. Their results showed a performance drop when using listener gaze, which may be explained by the choice of the method they used to interpret the gaze signal.

Our third scenario is automatic, interactive instruction-giving for collaborative assembly in a real environment. Specifically, we developed a multimodal instruction-giving system that generates referring expressions to identify objects and interprets the listener's eye movements to adapt its verbal output. The listener grasps the objects and assembles them to an individual model.

1.2 Research Aims and Contributions

The main objective of this thesis is to investigate the utility of object-directed listener gaze for efficient communication.

We address this topic in different settings. Efficient communication is particularly important for goal-oriented scenarios, where misunderstandings could lead to mistakes that require correction, leading to longer interaction times. Thus we investigate if a speaker (human or artificial) who gives instructions to a human listener can effectively use listener gaze to better refer to co-present objects and reason about the listener's intentions. Human speakers would be then more rational by adapting to the listener's focus of attention and by using the gaze indicator to cooperate more effectively. An assistance system that tracks and integrates listener gaze into the automatic generation of identifying instructions offers an attentive and interactive behavior, which could lead to more efficient communication.

The first research question we pose is whether listener gaze can improve automatic prediction of reference resolution. We report on the extension of a probabilistic observational model to also consider a listener's gaze behavior. More precisely, we describe how we

implemented features that encode listener's eye movement patterns and in this way we take the listener's perspective into account. Then we evaluate their performance on a multimodal data collection for interactions in virtual environments. We show that such a prediction model, which is aware of the listener's gaze position, is more accurate especially when the referential scene is complex with many competitors available next to the target. The results from this study were published in the proceedings of the Association for Computational Linguistics ACL 2015 (Koleva, Villalba, Staudte, & Koller, 2015).

Our second aim was to assess if and how a human speaker uses listener gaze in a real world task because listeners gaze reliably indicates language understanding. For this, we designed an exploratory study that involves spontaneous spoken instructions in a real environment while we manipulated the listener's gaze availability to the speaker (in the form of a cursor). The speaker remotely guided a naïve listener though a hall to find the next table and collect specific objects associated with everyday tasks. Gaze availability had no effect on the performance, but human speakers were already very good at this task. However, gaze behavior differed before and after, but not while an instruction was being spoken, suggesting that it was used more deliberately. Further, we observed that speakers produced more negative feedback when they could see the gaze cursor. We observed that the manipulation of availability of listener gaze position to the speaker had a main effect on listener gaze before and after an utterance, but not while an instruction was being spoken. Gaze availabilty further affected the type and amount of feedback given by speakers. Our findings have been published in the proceedings of the Annual Meeting of the Cognitive Science Society 2015 (Koleva, Hoppe, Moniri, Staudte, & Bulling, 2015).

Our third aim, and potentially the largest contribution of this thesis, is to examine if an NLG system that uses listener gaze can lead to more efficient interactions. We designed, implemented and tested two interactive NLG systems that use augmented reality technologies (Pfeiffer, 2012; Pfeiffer & Renner, 2014) to monitor listener gaze in real environment. The scenario we consider is collaborative assembly. We created complex referential scenes such that the generation of a uniquely identifying description was challenging. We used a toy scenario, where participants had to identify specific building blocks for constructing a LEGO model. We provide as a proof of concept an assistance system that can use listener gaze in real setups to facilitate collaboration and improve performance. As has been shown in virtual environments, using listener gaze can minimize error rate. Our work

extends previous findings by splitting the information into more chunks rather than generating a one-shot reference. We hypothesize that the incremental approach would lead to quicker (more efficient) task solving as it monitors listener gaze behavior and adapts the verbal output. A cooking task or building a LEGO model are often considered for evaluating assistance systems (cf. Section 2.2). Beyond these toy scenarios there are a number of other applications where a small improvement can have a large impact. For example, in the manufacturing industry, a wrong step in the production on an assembly line could be propagated and damage the end product. An assistance system that detects such mistakes in advance and gives a warning before they happen can save resources (time and money).

Technical Contributions Our NLG system *"Feedback"* uses listener gaze to provide feedback proactively to the user. We replicate findings from virtual environments that gaze-based feedback is beneficial. We further provide evidence that splitting the information into an ambiguous instruction and more informative feedback, which can also be thought of as referring in interactive installments, improves task performance and even outperforms following an unambiguous, exhaustive instruction. These results have been published in the proceedings of the Annual Meeting of the Cognitive Science Society 2018 (Mitev, Renner, Pfeiffer, & Staudte, 2018). Further, we found that the informativity of the gaze-driven feedback determines the engagement of the listener with the system. These results have been published in "Attention in Natural and Mediated Realities", a special issue of the journal "Cognitive Research: Principles and Implications" of the Psychonomic Society. Our NLG system *"Installments"* uses listener gaze more indirectly and incorporates this non-verbal cue into the generation algorithm. It implements two information delivery approaches to refer to co-present objects, either in gaze-driven installments or by providing a full description at once. Our results showed that following a full instruction (all installments concatenated) was faster than gaze-driven installments triggered by object inspections. However, mentioning the position of the searched-for object first made installments equally efficient as acting out an exhaustive instruction, suggesting that it is an effective information delivery approach for collaborative task solving.

Outline of the thesis:

Chapter 2 gives background information and discusses related work for both topics, natural language generation and listener gaze in situated interactions. In Chapter 3 we present the extension of a probabilistic observational model targeted to automatically predict reference resolution. We propose eye-tracking features that capture listeners' attention and evaluate the performance of the model. Further, in Chapter 4, we present an exploratory study that aims at assessing if and how a human speaker uses listener gaze when spontaneously referring to co-present objects. The main focus of this thesis is how an artificial speaker can use listener gaze to tune natural language generation. In Chapter 5 we present a multimodal interactive system (`GazInG`) that monitors listener gaze, and two NLG systems, *"Feedback"* and *"Installments"* to generate identifying spoken instructions on the fly in real environments. In order to assess the usefulness of listener gaze in this scenario, we designed and conducted three experiments. In Chapter 6 we present the first two experiments investigating the interaction with the NLG system *"Feedback"* and report on the effects of gaze-driven verbal feedback on performance. The third experiment is described in Chapter 7; it assessed the interaction with the NLG system *"Installments"* and whether an incremental instruction generation benefits listener understanding. Finally, we discuss limitations and address directions for future research in Chapter 8 before making our final conclusions.

Chapter 2

Background

In this chapter we present the scientific background relevant to natural language generation and the role of gaze in interaction specifically for assistance systems.

2.1 Natural Language Generation (NLG)

Natural language generation (NLG) is a sub-field of computational linguistics and focuses on computational systems that automatically produce natural language texts and speech. NLG systems are important for various applications ranging from weather reports to intelligent interactive systems. Thus NLG systems have different input and output formats and support different degrees of interactivity. However, all types of NLG systems face the problems of *content selection* (what to say), *surface realization* (how to say it) and *presentation* of the generated material, i.e. output text, speech with suitable intonation or even non-verbal cues. Making decisions for all steps represents a challenge, as there is no single correct solution but rather multiple options, and deciding on a specific one could be influenced by subjective preferences and creativity, which complicates the evaluation of such a system (Stent & Bangalore, 2014). There is always a trade-off when designing and implementing an NLG system concerning the complexity of the generation algorithm and its run time, i.e. depending on the end application, one could optimize for speed using a shallow generation approach, or for the advancement of the generation technique using a deep generation approach. Hybrid approaches combine both in order to benefit from their advantages (e.g. Klarner & Ludwig, 2004).

There are numerous approaches proposed for natural language generation. Depending on the goal of the application one or the other could be more or less suitable due to what constraints are involved. For example, an NLG system that is targeted to be used in real-time interactions requires a fast generation method in order to compute the output on the fly during an interaction. In contrast, if the goal of the NLG system is to offer a more sophisticated generation method that, for instance, simulates processes of human language production (and the system's output is optimized for quality), then this could be computationally expensive and would need to be done offline. Such an approach is suitable for applications that output natural language but do not involve active interaction, e.g. automatic text summarization. Ideally an NLG system overcomes the disadvantages of both approaches; thus, hybrid generation approaches are becoming more popular.

Our work focuses on the generation of identifying instructions that contain referring expressions. Referring expressions are verbal descriptions of an entity that allow a comprehender to identify it. They are commonly used and are relevant for any type of interaction but are particularly important for situated communication. Human speakers are very good at producing a description of a target object such that it is distinguished from other co-present competitor objects. A referring expression has to be informative enough to enable unique identification of a target object. The semantic content of a referring expression is usually chosen to contrast the target object from competitors. There is evidence that human speakers tend to mention redundant attributes and produce so-called overspecified referring expressions; that is, they mention more attributes of the target object than are needed to uniquely describe it (Engelhardt, Bailey, & Ferreira, 2006; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Some object attributes are preferred over others. Specifically, speakers often mention absolute attributes like color even if it is redundant (Pechmann, 1989). Further, Belke and Meyer (2002) found out that color is more frequently used in overspecified descriptions than a relative attribute like size. On the other hand, automatic generation of referring expressions faces the problem of attribute selection, i.e. how to decide which attributes to mention such that the comprehender is able to identify a target. The discriminatory power of an attribute plays an important role, that is, how many objects would be excluded by mentioning a specific object attribute. Further evidence suggests that, in highly interactive settings, referring in installments is a common phenomenon. That is, speakers provide the information incrementally by presenting it not all at once, but in subsequent chunks, to the listener (Striegnitz, Buschmeier, & Kopp, 2012). An example from their study and data collection in the GIVE challenge is S: "the

blue button" ... L: [moves and then hesitates] ... S: "the one you see on your right" ... L: [starts moving again] ... S:"press that one". Indeed, speakers often start speaking before they have planned the entire utterance, especially if they are under time pressure. They are thereby able to adapt to changes in the surroundings and the listeners' signals. However, whether an interactive system can also successfully adapt to the listener's behavior remains unclear. In the following, we review various approaches proposed to solve the problem of automatically generating referring expressions in natural language.

Already two decades ago, Dale and Reiter (1995) proposed an incremental algorithm for generating simple referring expressions similar to those produced by human speakers in accordance with the Gricean maxims (Grice, 1975). Their algorithm does not encode the ranking of attributes, but models human preferences based on empirical evidence. However, depending on the task and domain, the preferences could be different. The problem of automatically generating referring expressions is usually divided into three steps: 1) selection of the expression type, 2) selection of pre- and post-modifiers specifying object attributes like color, size etc., and 3) their realization in the form of linguistic expressions (Reiter & Dale, 2000). Krahmer, van Erk, and Verleg (2003) proposed a graph-based approach and framed the problem as finding the sub-graph that minimizes cost. Their approach also assumes that some attributes are preferred over others and are thus associated with a lower cost.

Simple approaches provide computationally efficient generation that is suitable for real time applications. In order to realize an adaptive behavior, the system has to accommodate its verbal output to the user's behavior and changes in the environment. Thus the generation and output of natural language expressions should happen on the fly and cannot be done in advance. Another important aspect for designing an interactive system is domain independence, i.e. switching to another domain does not require re-implementation of the generation algorithm, and existing modules are portable to other applications.

A very simple approach is to use canned text that is defined prior to runtime and is presented whenever triggered without any adaptation because the linguistic output is static. For this, template-based realization is used; that is, the templates are pre-defined and during runtime slot filling is applied (e.g. Channarukul, 1999). Another method is to use a rule-based approach for generation by defining a grammar that encodes the syntactic structure of the utterances an NLG system can generate (e.g. DeVault, Traum, & Artstein, 2008). There are some systems that use a hybrid approach by combining template and

rule-based generation for spoken dialogue applications (e.g. Stent, 2001; Galley, Fosler-Lussier, & Potamianos, 2001). In the existing literature the opposition of template vs. real generation has been discussed (Reiter, 1995; Van Deemter, Krahmer, & Theune, 2005). Template-based systems are not as easy to extend and maintain as linguistically-based systems, where building additional functionality does not require major changes such as rewriting templates.

Later on, the GIVE challenge addressed more advanced NLG and tested different methods for instruction generation in situated communication (Koller, Striegnitz, Byron, et al., 2010). There, an NLG system guided a human listener through a virtual maze, referred to specific buttons to be pressed and thus proposed objective evaluation metrics for an NLG system. Different approaches for generating referring expressions have been developed and tested in the GIVE framework, which offers more objective evaluation of NLG systems in virtual environments. For such a dynamic task, the adaptation of the instruction generation to the constantly changing visual context is necessary. Stoia, Shockley, Byron, and Fosler-Lussier (2006) presented a machine learning approach that interleaved navigational and discrimination information to better control the situated context. Further, Garoufi and Koller (2010) presented a natural language generation method that made use of AI planning techniques. They exploited non-verbal context in situated interactions and guided the listener to a location, which is convenient for the generation of simple referring expressions with context-dependent adjectives. Both approaches plan and output a reference as a single noun phrase, the so-called "one-shot" reference. However, splitting a referring expression into shorter information chunks can be beneficial. For instance, Mitchell, van Deemter, and Reiter (2013) proposed a method for generating expressions to refer to co-present objects, and they separated absolute form relative properties, which often resulted in overspecified expressions. Their algorithm was evaluated in two domains and was shown to outperform previously proposed algorithms by Dale and Reiter (1995), Krahmer et al. (2003) and Viethen, Dale, Krahmer, Theune, and Touset (2008). Another study by Kelleher and Kruijff (2006) focused on generating spatial expressions incrementally. Their work focuses on expressions that describe the spatial relation of a target object to a reference object, also known as a landmark. Object descriptions that specify the position of the target object relative to a landmark could be computationally expensive due to the high number of combinations. They address this issue and exemplify their approach on a static scene with the long-term goal to apply such an algorithm in

situated dialog for the development of conversational robots. The production of viewer-centered expressions (such as "on the left") involves perspective-taking, which abstracts from spatial relations. Depending on the setting, such expressions might be preferred, as they might be more appropriate. For example in the GIVE-2 challenge, there were more viewer-centered spatial expressions than expressions containing a spatial relation (Koller, Striegnitz, Gargett, et al., 2010). Our approach is to use the listener's gaze position and to specify the relative position of the target object; in this manner, we avoid the issue of searching for a landmark. More recently, the generation of installments, that is, referring expressions delivered piece-wise instead of being output all at once, was also shown by Zarrieß and Schlangen (2016) to improve performance on object identification in real-world pictures. They first output an easy expression; if it is not understood, then try to combine it with another one or paraphrase it. Their findings suggest that such a generation approach enhances identification of real objects depicted in static images and has a stable success rate over time. Further, Villalba, Teichmann, and Koller (2017) proposed the generation of contrastive referring expressions. They presented a static scene to the user, asking them to select an object that matches a written description. Their system detects misunderstandings of a referring expression whenever the wrong object was selected. Then it generates contrastive referring expressions that emphasize other object attributes in order to achieve communicative success. This strategy was shown to be effective and also preferred by the users'.

In this thesis, we investigate NLG for situated spoken interaction inspired by the GIVE challenge but switching to a real environment. Further, we examine whether interpreting listener gaze could be utilized for incremental generation of identifying instructions in dynamic setups, as opposed to Zarrieß and Schlangen (2016), who consider static scenes. The task we used is collaborative assembly in a real environment. We focus on interpreting listener gaze behavior with respect to referring expression resolution. For that, we built a multimodal interactive system and developed two NLG systems embedded in it that use gaze cues to detect misunderstandings early on and either proactively generate verbal feedback (see Section 5.2.2, NLG system *"Feedback"*) or trigger the next installment (see Section 5.2.3, NLG system *"Installments"*).

2.2 Listener Gaze in Task-oriented Interaction

Previous research has shown that listeners follow speakers' verbal references (as well as their gaze in face-to-face situations) to rapidly identify a referent (Eberhard et al., 1995; Hanna & Tanenhaus, 2004). Listener gaze reveals a lot about how the listener processes a given word, namely quickly, incrementally and in a way that is tightly linked to the visual context. Keysar, Barr, Balin, and Brauner (2000) looked at mental processes that underlie perspective taking in comprehension. They observed that although listeners know that some objects are not visible to the speaker, they do not restrict the visual search, but also consider non-visible objects when trying to establish a reference. As soon as the listener becomes aware of an error, she uses common ground to correct it, i.e. information about the speaker's perspective is used while interpreting an utterance. In contrast to these findings, Barr (2008) showed that the listener uses common ground solely before receiving the message and not during its interpretation. Further, Brown-Schmidt (2009) demonstrated that other factors influence the listener's initial interpretation, as well. Specifically, it is sensitive to the partner and depends on the identity of the speaker and the experience of interacting with them. The reaction of the speaker to referential eye movements, however, was considered in only a few studies. Clark and Krych (2004), for instance, aimed to grasp this *reciprocal* nature of an interaction in a study using a collaborative block building task and manipulating whether participants could see each other or each other's workspaces. Their results suggested that the joint workspace was more important than seeing each other's faces. Using the GIVE setup, Staudte et al. (2012) conducted a study in which users were guided by an NLG system through a virtual world to find a trophy. The system either gave feedback on the users' eye movements, or not. This controlled setting allowed the observation of dynamic and interactive (gaze) behavior while maintaining control over one interlocutor (the NLG system). The results of this study suggest that it can be beneficial for task performance when listener gaze is exploited by the speaker to give feedback. It remains unclear, however, whether (human) speakers indeed provide such feedback and how the availability of listener gaze *recursively* affects the spoken instructions and, possibly, the gaze behavior itself.

Assistance Systems Gaze-based assistive technologies have a long tradition in command-like desktop interfaces for the physically challenged, but with advances in mobile eye tracking technologies, they have moved into less controlled environments in the last

decade (Pfeiffer, 2013). Our work is related to work in attentive assistance systems (Maglio, Matlock, Campbell, Zhai, & Smith, 2000) or human-robot/human-agent interaction, where gaze is also relevant for the social aspects of interaction (Sidner, Kidd, Lee, & Lesh, 2004; Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005) as well as for grounding verbal utterances using mechanisms of joint attention (Imai, Ono, & Ishiguro, 2003). The focus of our work is more on assistance systems and user gaze behavior for understanding collaborative comprehension processes. Gaze has already been used in previous work on assistance systems to tune verbal or visual feedback. For example, a prototype of an attentive mobile eye tracking system has been presented, which monitored eye movements in real time and provided feedback to guide the user back to a given track on a map (Eaddy, Blasko, Babcock, & Feiner, 2004). There is, however, no report on a systematic evaluation of the system and it is not stated to what extent the natural language feedback was generated automatically. This kind of interaction is typical for perceptual user interfaces. For example, Turk and Robertson (2000) consider gaze-assisted interaction and the quote *"No, not that one!"* they suggested in their article is actually realized by our working system. Smart Eyewear has been identified as a key technology for assistance systems (Pfeiffer, Feiner, & Mayol-Cuevas, 2016) and recently has been combined with a real-time analysis of eye tracking to support assembly tasks (Renner & Pfeiffer, 2017; Blattgerste, Strenge, Renner, Pfeiffer, & Essig, 2017).

Gaze has been used for HCI, but in rather non-verbal interactions. It is shown to be a faster indicator in the context of object selection task than a hand movement (Kosunen et al., 2013). Carter, Newn, Velloso, and Vetere (2015) built a gaze and gesture system to investigate the role of showing the gaze cursor (referred to as feedback) to the user during an object selection task when playing a game. In a user study, they found that people dislike the version with a visualized gaze cursor. On the other hand, (Garkavijs, Okamoto, Ishikawa, Toshima, & Kando, 2014) showed that gaze feedback improves satisfaction in an exploratory image search.

Further, Torrey, Fussell, and Kiesler (2013) investigated the usefulness of adaptive robot behavior. The robot instructed experts and novices on which cooking tools to select next. The robot responded to users' typed input. They observed a benefit of the adaptive behavior for the users, especially when they were under time pressure. Later on, Andrist, Gleicher, and Mutlu (2017) considered bidirectional gaze in face-to-face communication

and mechanisms to coordinate gaze cues of a virtual character with a human user identifying ingredients for making a sandwich. They showed that the virtual character can produce quick and effective non-verbal references by responding to users' gaze. Their interactive system is based on interactions obtained from a human-human study. The virtual agent initially provides a verbal reference to identify a target, but it is not automatically generated. An error in such toy scenarios might not be fatal, but for more serious applications, minimizing the risk can have a larger impact. For example, Reynal, Colineaux, Vernay, and Dehais (2016) present a study involving pilots in the cockpit, aiming to assess how the crew supervises the flight deck. They found that both pilots (flying and monitoring) looked more at the primary than at the secondary flight parameters; also, a similar visual behavior of both pilots was observed. Their findings suggest that the visual behavior of the pilot monitoring attention could be suboptimal. Another study by Campana et al. (2001) extended a dialogue system integrated in a simulated version of the Personal Satellite Assistant to also monitor user gaze. If an underspecified command is given by the user, the system asks for clarification before any action is performed. However, such clarification may appear unnatural if the user looks at the intended target. They expected to see a reduction in task completion times and turns taken during the interaction, but no systematic evaluation of their approach was reported.

If the user is performing actions in such a critical use case, where it is crucial not to make mistakes, an assistance system could prevent mistakes from happening by exploiting listener gaze to detect misunderstandings.

Chapter 3

Listener Gaze for Automated Prediction of Reference Resolution

Interactive systems that generate natural language to collaborate on a task with a human listener aim at effective and efficient interaction. Ideally they should model the grounding process, that is, monitor listener behavior and respond to it. If the listener intends to perform an incorrect action, it would be useful if the system could detect a misunderstanding and react with a warning in order to prevent a wrong step that would have to be undone. In this manner, an interactive system would be more attentive and importantly would ensure more efficient interaction. The first step towards realizing such a clever mechanism is to address the problem of automatic prediction of reference resolution. That is, we aim to automatically predict how the listener has resolved a referring expression by evaluating her visual behavior. Engonopoulos et al. (2013) proposed two statistical models to solve a grounding problem, i.e. to predict (mis-)understandings of a referent described by an automatically generated object description: a semantic model P_{sem} computing predictions based on the linguistic content, and an observation model P_{obs} computing predictions based on listener behavior features.

In this chapter, we present joint work with Alexander Koller and Martín Villalba and report on the extension of the observational model P_{obs} introduced by Engonopoulos et al. (2013). We address the research question of how to automatically predict a referring expression (RE) resolution, i.e., answering the question of which entity in a virtual environment has been understood by the listener after receiving an instruction. While

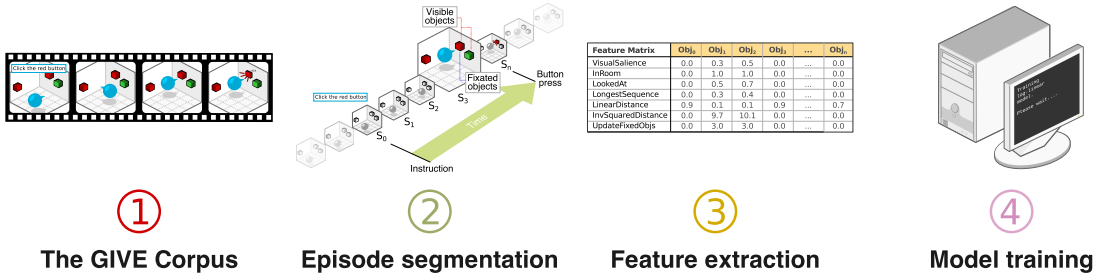

FIGURE 3.1: The processing pipeline for automatic prediction of reference resolution.

the linguistic material in instructions carries a lot of information, even completely unambiguous descriptions may be misunderstood. A robust NLG system should be capable of detecting misunderstandings and preventing its users from making mistakes. Language comprehension is mirrored by interlocutors' non-verbal behavior, and this can help when decoding the listener's interpretation. Precise automatic estimates may be crucial when developing a real–time NLG system, as such a mechanism would reliably predict the next action to be taken by an instruction follower. In the case of detecting a misunderstanding, the system can plan and output a corrective response aiming at more effective interaction. Specifically, we implement features that encode the listener's eye movement patterns and extend the P_{obs} model to evaluate their performance on a multimodal data collection (the GIVE Corpus). We show that the extended observational model, as it takes an additional communication channel into account, provides more accurate predictions, especially when dealing with complex, more cluttered scenes where more competitors next to the target object are available. These results have been published in the proceedings of the Association for Computational Linguistics ACL 2015 (Koleva, Villalba, et al., 2015).

3.1 Problem Definition

Figure 3.1 illustrates our processing pipeline for the automatic prediction of reference resolution. We segment the collected interactions into episodes consisting of the beginning of an instruction (speech onset) until the target button is pressed (action). The next step is to extract the observational and eye tracking features. After that the prediction models are trained to correctly predict how the reference is resolved, i.e. which button will be pressed.

More formally, let's assume a system generates an expression r that aims to identify a target object o_t among a set O of possible objects, i.e. those available in the scene view. Given the state of the world s at time point t, and the observed listener's behavior $\sigma(t)$ of the user at time $t \geq t_b$ (where t_b denotes the end of an interaction), we estimated the conditional probability $p(o_p|r, s, \sigma(t))$ that indicates how probable it is that the listener resolved r to o_p.

This probability can be also expressed as follows:

$$P(o_p|r, s, \sigma(t)) \propto \frac{P_{sem}(o_p|r, s)P_{obs}(o_p|\sigma(t))}{P(o_p)}$$

Following Engonopoulos et al. (2013) we make the simplifying assumption that the distribution of the probability among the possible targets is uniform and obtain:

$$P(o_p|r, s, \sigma(t)) \propto P_{sem}(o_p|r, s)P_{obs}(o_p|\sigma(t))$$

We expect an NLG system to compute and output an expression that maximizes the probability of o_p. Due to the dynamic nature of our scenarios, we also require the probability value to be updated at certain time intervals throughout an interaction. Tracking the probability changes over time, an NLG system could proactively react to changes in its environment. Henderson and Smith (2007) show that accounting for both fixation location and duration are key to identify a player's focus of attention.

3.2 Episodes and Feature Functions

The data for our experiment was obtained from the GIVE Challenge (Koller, Striegnitz, Gargett, et al., 2010), an interactive task in a 3D virtual environment in which a human player (instruction follower) is guided through a maze, locating and pressing buttons in a predefined order, aiming to unlock a safe. While pressing the wrong button in the sequences doesn't always have negative effects, it can also lead to restarting or losing the game. The instruction follower receives instructions from either another player or an automated system (instruction giver). The instruction follower's behavior was recorded every 200ms, along with the instruction giver's instructions and the state of the virtual world. The result is an interaction corpus comprising over 2500 games and spanning over

340 hours of interactions. These interactions were mainly collected during the GIVE-2 and the GIVE-2.5 challenges. A laboratory study conducted by Staudte et al. (2012) comprises a data collection that contains eye-tracking records for the instruction follower. Although the corpus contains both successful and unsuccessful games, we have decided to consider only the successful ones.

We define an *episode* in this corpus as a typically short sequence of recorded behavior states, beginning with a manipulation instruction generated by the instruction giver and ending with a button press by the instruction follower (at time point t_b). In order to make sure that the recorded button press is a direct response to the instruction giver's instruction, an episode is defined such that it contains no further utterances after the first one. Both the target intended by the instruction giver (o_t) and the one selected by the instruction follower (o_p) were recorded.

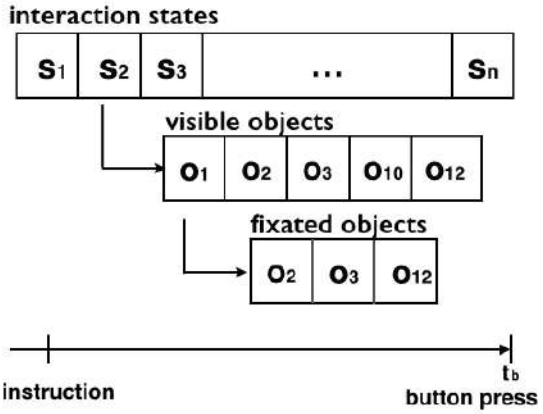

FIGURE 3.2: The structure of the interactions.

Figure 3.2 depicts the structure of an episode when eye-tracking data is available. Each episode can be seen as a sequence of interaction states $(s_1, \ldots, s_n)$, and each state has a set of visible objects $(\{o_1, o_2, o_3, o_{10}, o_{12}\})$. We then compute the subset of fixated objects $(\{o_2, o_3, o_{12}\})$. We update both sets of visible and fixated objects dynamically in each interaction state with respect to the change in visual scene and the corresponding record of the listener's eye movements.

We developed feature functions over these episodes. Along with the episode's data, each function takes two parameters: an object o_p for which the function is evaluated, and a parameter d seconds that defines how much of the episode's data the feature is allowed to analyze. Each feature looks only at the behavior that happens in the time interval $-d$

to 0. Henceforth we refer to the value of a feature function over this interval as its value at time $-d$. The value of a feature function evaluated on episodes with length less than d seconds is undefined.

3.3 Prediction Models

Given a referring expression uttered by an instruction giver, the *semantic* model P_{sem} estimates the probability for each possible object in the environment to have been understood as the referent, ranks all candidates and selects the most probable one in a current scene. This probability represents the semantics of the utterance, and is evaluated at a single time point immediately after the instruction (e.g. "press the blue button") has been uttered. The model takes into account features that encode the presence or absence of adjectives carrying information about the spatial or color properties (like the adjective "blue"), along with landmarks appearing as post-modifiers of the target noun.

In contrast to the semantic model, the *observational* model P_{obs} evaluates the changes in the visual context and the player's behavior after an instruction has been received. The estimated probability is updated constantly before an action, as the listener in our task-oriented interactions is constantly in motion, altering the visual context. The model evaluates the distance of the listener position to a potential target, whether it is visible or not, and also how salient an object is in that particular time window.

Interlocutors constantly interact with their surroundings and point to specific entities with gestures and, importantly, with their eyes. Gaze behavior is also driven by the current state of the interaction. As we have seen above, eye movements provide useful information indicating language comprehension. That is, they are tightly aligned with the linguistic input and give some insights about listener intentions. In particular, for goal-oriented interactions they can reveal what the listener is about to do next. Thus, we extend the basic set of P_{obs} features and implement eye tracking features that capture gaze information. We call this the *extended observational* model P_{Eobs} and consider the following additional features:

1. *Looked at:* this feature counts the number of interaction states in which an object has been fixated at least once during the current episode.

2. *Longest Sequence:* detects the longest continuous sequence of interaction states in which a particular object has been fixated.

3. *Linear Distance:* returns the Euclidean distance *dist* on screen between the gaze cursor and the center of an object.

4. *Inv-Squared Distance:* returns $\frac{1}{1+dist^2}$.

5. *Update Fixated Objects:* expands the list of fixated objects in order to consider the instruction follower's focus of attention. It successively searches in 10-pixel steps and stops as soon as an object is found (the threshold is 100 pixels). This feature evaluates to 1 if the list of fixated objects has been expanded and 0 otherwise.

When training our model at time $-d_{train}$, we generate a feature matrix. Given a training episode, each possible (located in the same room) object o_p is added as a new row, where each column contains the value of a different feature function for o_p over this episode at time $-d_{train}$. Finally, the row based on the target selected by the instruction follower is marked as a positive example. We then train a log-linear model, where the weights assigned to each feature function are learned via optimization with the L-BFGS algorithm. By training our model to correctly predict a target button based only on data observed up until $-d_{train}$ seconds before the actual action t_b, we expect our model to reliably predict which button the user will select. Analogously, we define accuracy at testing time $-d_{test}$ as the percentage of correctly predicted target objects when predicting over episodes at time $-d_{test}$. This pair of training and test parameters is denoted as the tuple (d_{train}, d_{test}).

3.4 Dataset

We evaluated the performance of our improved model on data collected by Staudte et al. (2012) using the GIVE Challenge platform. Both training and testing were respectively performed on a subset of the data obtained during a collection task involving worlds created by Gargett, Garoufi, Koller, and Striegnitz (2010), designed to provide the task with varying levels of difficulty. This corpus provides recorded eye-tracking data, collected with a remote faceLAB system. In contrast, the evaluation presented by Engonopoulos et al. (2013) uses only games collected for the GIVE 2 and GIVE 2.5 challenges, for which

no eye-tracking data is available. Here, we do not investigate the performance of P_{sem}, but concentrate on the direct comparison between P_{obs} and P_{Eobs} in order to find out if and when eye tracking can improve the prediction of an RE resolution.

We further filtered our corpus in order to remove noisy games following Koller et al. (2012), considering only interactions for which the eye-tracker calibration detected inspection of either the target or another button object in at least 75% of all referential scenes in an interaction. The resulting corpus comprises 75 games, for a combined length of 8 hours. We extracted 761 episodes from this corpus, amounting to 47m 58s of recorded interactions, with an average length per episode of 3.78 seconds ($\sigma = 3.03 sec.$). There are 261 episodes shorter than 2 sec., 207 in the 2-4 sec. range, 139 in the 4-6 sec. range, and 154 episodes longer than 6 sec.

3.5 Evaluation and Results

The accuracy of our probabilistic models depends on the parameters (d_{train}, d_{test}). At different stages of an interaction the difficulty of predicting an intended target varies as the visual context, and in particular the number of visible objects, changes. As the weights of the features are optimized at time $-d_{train}$, it would be expected that testing also at time $-d_{test} = -d_{train}$ yields the highest accuracy. However, the difficulty of making a prediction decreases as $t_b - d_{test}$ approaches t_b, i.e. as the player moves towards the intended target. We expect that testing at $-d_{train}$ works best, but we need to be able to update continuously. Thus we also evaluate at other timepoints and test several combinations of the (d_{train}, d_{test}) parameters.

Given the limited amount of eye-tracking data available in our corpus, we replaced the cross-corpora-challenge test setting from the original P_{obs} study with a ten-fold cross-validation setup. As training and testing were performed over instances of a certain minimum length according to (d_{train}, d_{test}), we first removed all instances with length less than $max(d_{train}, d_{test})$, and then performed the cross-validation split. In this way we ensured that the number of instances in the folds were not unbalanced. Moreover, each instance was classified as *easy* or *hard* depending on the number of visible objects at time t_b. An instance was considered *easy* if no more than three objects were visible at that point, or *hard* otherwise (see Figure 3.3 for examples). For $-d_{test} = 0$, 59.5% of

all instances are considered *hard*, but this proportion decreases as $-d_{test}$ increases. At $-d_{test} = -6$, the number of hard instances amounts to 72.7%.

FIGURE 3.3: The GIVE corpus: Example visual context of an easy (left picture) and a hard (right picture) referential scene in the virtual environment.

We evaluated both the original P_{obs} model and the P_{Eobs} model on the same dataset. We also calculated accuracy values for each feature function, in order to test whether a single function could outperform P_{obs}. We included as baselines two versions of P_{obs} using only the features *InRoom* and *Visual Salience* proposed by Engonopoulos et al. (2013).

The accuracy results in Figure 3.4 show our observations for $-6 \leq -d_{train} \leq -2$ and $-d_{train} \leq -d_{test} \leq 0$. The graph shows that P_{Eobs} performs similarly as P_{obs} on the *easy* instances, i.e. the eye-tracking features are not contributing in those scenarios. However, P_{Eobs} shows a consistent improvement on the *hard* instances over P_{obs}.

For each permutation of the training and testing parameters (d_{train}, d_{test}), we obtain a set of episodes that fulfill the length criteria for the given parameters. We apply P_{obs} and P_{Eobs} on the obtained set of instances and measure two corresponding accuracy values. We compared the accuracy values of P_{obs} and P_{Eobs} over all 25 different (d_{train}, d_{test}) pairs, using a paired samples t-test. The test indicated that the P_{Eobs} performance ($M = 83.72$, $SD = 3.56$) is significantly better than the P_{obs} performance ($M = 79.33$, $SD = 3.89$), $(t(24) = 9.51, p < .001, Cohen's\ d = 1.17)$. Thus, eye-tracking features seem to be particularly helpful for predicting to which entity an RE is resolved in hard scenes.

The results also show a peak in accuracy near the -3 seconds mark. We computed a 2x2 contingency table that contrasts correct and incorrect predictions for P_{obs} and P_{Eobs}, i.e. whether o_i was classified as the target object or not. Data for this table was collected from all episode judgements for models trained at times in the $[-6\,sec., -3\,sec.]$ range and tested at -3 seconds. McNemar's test showed that the marginal row and column

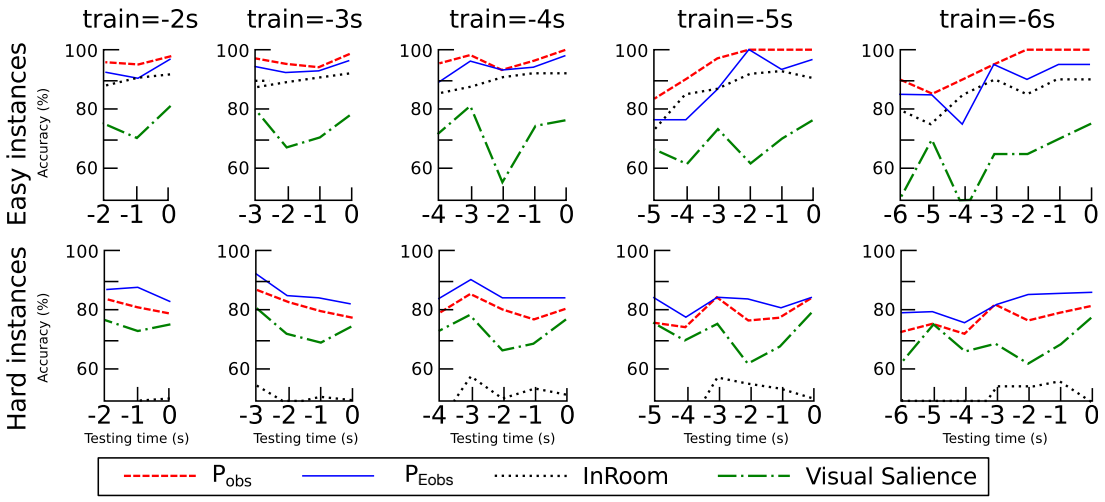

FIGURE 3.4: Accuracy as a function of training and testing time.

frequencies are significantly different ($p < 0.05$). This peak is related to the average required time between an utterance and the resulting target manipulation. This result indicates that our model is more accurate precisely at points in time when we expect fixations to a target object.

3.6 Discussion

In this chapter, we have demonstrated that accuracy increases when considering eye tracking features in the context of predicting the resolution of an RE. Eye movements are a good indicator of language comprehension because they are tightly connected to the visual scene and are driven by the semantics of a given word as well as the goal to identify and press the referenced object (see Chapter 2). In addition, we observed that our extended observational model P_{Eobs} proves to be more robust than the basic observational model P_{obs} when the time interval between the prediction ($t_b - d_{test}$) and the button press (t_b) gets larger, i.e. gaze is especially beneficial in an early stage of an interaction. This approach shows a significant accuracy improvement on hard referential scenes where more objects are visible and thus each can be seen as a potential target.

We have also established that gaze is particularly useful when combined with some other simple features, as the features that capture the listener's visual behavior are not powerful

enough to outperform even the simplest baseline. Gaze only benefits the model when it is added on top of features that capture the visual context, i.e. the current scene.

In other words, gaze alone is not sufficient to accurately predict how a reference will be resolved. However, it provides precise information that is temporally aligned with an utterance as listeners look at what they hear (Tanenhaus et al., 1995). Specifically in task-oriented scenarios that involve reference resolution, this signal encodes listeners' intentions (Staudte et al., 2012) and thus is a reliable information source.

A future line of research is the combination of our P_{Eobs} model with the semantic model P_{sem}, in order to test the impact of the extended features in a combined model, which is out of the scope of this thesis. If successful, such a model could provide reliable predictions for a significant amount of time before an action takes place. This is of particular importance when it comes to designing a system that automatically generates and outputs feedback online to confirm correct and reject incorrect listener intentions.

Testing with users in real time is also an area for future research. An implementation of the P_{obs} model is currently in the testing phase, and an extension for the P_{Eobs} model would be the immediate next step. The model could be embedded in an NLG system to improve the automatic language generation in such scenarios.

As corpora collections containing eye-tracking data are sparse, here it remains open whether this effect applies only to the considered domain, or would be evident in other interactive scenarios as well as in a real environment. Indeed, it would be interesting to ask if a human instruction giver could benefit from the predictions of P_{Eobs}. We could study whether predictions based on the gaze (mis-)match between both interlocutors are more effective than simply presenting the instruction follower's gaze to the instruction giver and trusting the instruction giver to correctly interpret this continuous signal. If such a system proved to be effective, it could point out misunderstandings to the instruction giver before either of the participants becomes aware of them.

Our study builds on previous work from virtual environments and considers an automated speaker, an NLG system. Garoufi et al. (2016) showed that an NLG system can exploit this non-verbal cue and react to it with verbal feedback, which led to more effective interaction. However, whether human speakers actively use eye movements and react to them with feedback, and if their feedback would be even more informative, or simply different, is still unclear.

Thus we address this research question in the next chapter, where we switch to a real environment. Specifically, we investigate whether and how a remote human speaker uses the additional information of where the listener is currently looking while spontaneously referring to real-world objects the listener has to select, and if so, how this influences performance.

Chapter 4

Human-Human Interaction: Availability of Listener Gaze in an Indoor Guidance Task

We constantly direct our gaze to different parts of the visual scene to be able to perceive objects of interest with high acuity. These eye movements can be driven internally, i.e. by some self-initiated goal or intent, or externally, by something that attracts our visual attention (Yantis & Jonides, 1990). External factors that drive a listerner's attention can be the saliency of visible objects or another person's utterances that direct our eyes to a co-present object or an event. The latter has been exploited in many psycholinguistic studies in order to study language comprehension processes (e.g. Cooper, 1974; Tanenhaus et al., 1995). Conversely, a listener's gaze may also signal (mis-)understanding back to the speaker. Taking the listener's behavior into account when planning and making utterances is an important aspect of collaborative, goal-oriented interaction. In this sense, listener's eye movements can be both a result of a comprehension process, i.e. a "symptom", and/or a "signal" and feedback channel to the speaker, who can then react to it by adapting their next utterance.

This chapter describes an explorative real-world study on indoor guidance. The study investigated whether and how showing the listener gaze to a human speaker influences interlocutors' behavior during task-oriented interactions. Specifically, a remote speaker gave instructions to a naïve listener to localize, identify and collect specific objects while

being eye-tracked. The speaker was asked to verbally guide the listener and together they solved nine tasks. Firstly, the experiment was aimed at revealing whether the availability of listener gaze position to the speaker would affect the production of verbal feedback. Secondly, if gaze was used as a *signal*, which listeners control and use deliberately, then the option to do so (and thereby evoke speaker reactions) would ubiquitously change listener gaze. If gaze was more generally a *symptom* of other processes and deliberate control was (too) difficult, listener gaze would change according to tasks or events rather than based on *GazeAvailability*. Finally, if gaze was used as a *signal*, variations of listener gaze behavior should mainly occur prior to an utterance. If gaze was a reaction to changes in the utterances (i.e. a *symptom*), gaze behavior should instead change after an utterance.

We obtained a multimodal data collection consisting of the videos from the listeners' perspective, their gaze data, and instructors' utterances. We analyze the changes in instructions and listener gaze with respect to *GazeAvailability* when the speaker can see 1) only the video (NoGaze), 2) the video and the gaze cursor (Gaze), or 3) the video and a manipulated gaze cursor, i.e. one not displayed on the exact gaze position but randomly shifted with an offset of $\pm 0,2$ (ManGaze). Our results show that listener visual behavior mainly depends on utterance presence but also varies significantly before and after instructions. Additionally, we observed that more negative feedback occurred in condition 2). While piloting a new experimental setup, our results provide an indication for gaze reflecting both a symptom of language comprehension, and a signal that listeners employ when it appears useful, and which therefore adapts to our manipulation. Our findings have been published in the proceedings of the Annual Meeting of the Cognitive Science Society 2015 (Koleva, Hoppe, et al., 2015).

We expected to encounter different types of instructions changing with the availability of listener gaze to the speaker. As a consequence, listeners may even consciously use their gaze, similar to a pointing gesture, for instance in order to point to an object when the hands are full. Further, we assumed that showing listener gaze would lead to more efficient interactions. For this reason, we tested if the speaker could use the information about the current focus of the listener's attention and so utter more precise instructions. Additionally, we investigated whether a speaker needs the exact gaze position (Gaze) or could also make use of the general area where the listener looks (ManGaze), or if this would be confusing for the speaker.

4.1 Method

We designed a task that combines a dynamic, interactive setting with the possibility to conduct exact and detailed analyses, in particular on eye movement behavior, in order to assess the mutual influence of listener gaze and speech in human-human interaction. Naïve participants either became an instructor (speaker) or a walker (listener). The speaker instructed the listener to perform a series of tasks. These tasks consisted of a navigational part, i.e. finding the next out of nine tables in a hall, which we call the *macro* task, and the identification of some objects at each table, referred to as the *micro* task. Each pair of participants experienced all three *GazeAvailability* conditions in a different order according to a Latin square.

FIGURE 4.1: The experimental setup: the walker wearing a head-mounted eye tracker, following instructions and selecting objects (left picture) and the remote instructor giving directions, describing the next target and monitoring the walker's behavior (right picture)

Figure 4.1 depicts our setup with the two roles of the interlocutors: the remote instructor received a plan of the route needed to solve the macro tasks and a static picture of the tabletop where the next target object for the *micro* task was highlighted. The walker listened to the instructions via headset and wore a head-mounted eye tracker through which the speaker could see the scene from the listener's perspective without, with or with a shifted gaze cursor. The purpose of manipulating *GazeAvailability* was to reveal whether the availability of listener gaze to the speaker affected a) the produced utterances and b) the listener's gaze behavior. We included ManGaze in order to investigate whether slightly perturbed gaze would be considered either uninformative (more like NoGaze) or

even distracting, or whether the speaker would be robust towards slight imprecisions of the gaze cursor and treat it more like the GAZE condition.

Here is an example trial starting with a *macro* task consisting of navigational instructions:

(1) "So, und jetzt musst du nochmal laufen, und zwar wieder zurück an diesen vorherigen Tisch ... nochmal zurück durch die, durch die Tür, in diesen Konferenzraum. Un vor dem Tisch standest du grad eben schon ... das ist der, wo so Scheren und Post-its drauf liegen. Nein, das iss er nicht, der andere, auf der andern Seite von, also quasi gegenüber ... Genau, genau."
(So and now you should walk again, namely back again to the previous table ... again through the door into the conference room. And you were just now standing in front of this table ... this is it, where the scissors and the post-its are placed. No, it is not that one ... the other one on the other side, opposite. ... Right, exactly!)

As soon as the walker reached the right location, the instructor continued with the *micro* tasks and started describing the first target object that was supposed to be collected from the current setup.

(2) "Also, da sollen wir ne Schere suchen, und zwar ist das, die zweite schwarze Schere von oben auf der rechten Seite... die liegt neben einem grünen Stift... Genau die... ok. Danke."
(Alright, here we should search for a pair of scissors, namely this is the second black pair of scissors from above on the right side... it is located next to a green pen... Exactly that one!... OK. Thanks!)

Materials

The nine micro tasks were associated with daily duties such as doing office work or cooking. Office scenarios included writing a letter using envelopes, pens, paper and glue; kitchen scenarios, for example, making a cake using milk, sprinkles, mixing spoons and a carton of eggs. To make the task sufficiently complex, i.e. to have hard referential scenes, and elicit the production of detailed referring expressions, which uniquely identify the target

object, at least two competitor objects for each target were also included in each setup. In total, 234 everyday objects were used; 36 of them were target objects (see Appendix C).

Participants

Twelve pairs of participants (16 females) took part in this study. The average age was 26.6 and all but one were in the age range 18–40. All participants were German native speakers and received a payment of €10. They reported normal or corrected-to-normal vision. A session lasted between 30 and 45 minutes.

Procedure

Participants were first asked about their preference for role assignment and assigned to the walker/instructor role accordingly. Two experimenters instructed both participants separately from each other. The participants received a brief description (one page) of their role. Specifically, the instructor was shown the route and tables but was not told how to refer to the target objects in order to avoid priming. Then, the instructor was led to a remote room from which she guided the walker. During the experiment the instructor saw a picture of the current target object, a map of the hall, and the scene view of the walker (see right picture in Figure 4.1). Neither walker nor instructor were informed about our manipulation.

Apparatus

We used a Pupil Pro monocular head-mounted eye tracker for gaze data collection (Kassner, Patera, & Bulling, 2014). The tracker is equipped with a high-resolution scene camera (1280 x 720 pixels) and eye camera (640 x 360 pixels). We extended the Pupil software with additional functionality needed for our study, namely to hide and display a manipulated gaze cursor to the instructor.

Two notebooks were used: one for the walker and one for the instructor. The instructor notebook was connected to two displays, one for the instructor and one for the experimenter. The experimenter sitting next to the instructor used a control panel to send commands to the eye-tracking software to switch between conditions. The mobile eye tracker was connected to the walker notebook, which was a `MacBook Pro Mid 2013` on which `Ubuntu 14.04` was running. The eye tracker was connected to the walker notebook on which we recorded the incoming sound, i.e. the instructions the listener heard. At the same time, the walker's speech was muted such that we ensured only non-verbal responses from the walker. For this, the sound was redirected (the output sound was assigned to the input channel in order to record the incoming instructions). A command line audio recorder, `SoX`, was launched in parallel for each new recording.
Both audio and video signals were streamed using Skype. In addition, the walker was equipped with a presenter to signal success (finding a target object) by pressing a green button or confusion (when something was unclear) by pressing a red button. When the green button was pressed, the picture of the next target object was updated and a new recording was started. If the red button was pressed, then a picture depicting confusion was shown on the instructor's screen. The communication of the different software components was implemented using custom client-server software, but all recordings were carried out on the walker machine.

Measures and Analysis

We collected a multimodal corpus of interactions and derived various measures to analyze interlocutors' behavior. We evaluated the linguistic material produced by the instructors and the eye movements of the listener and assessed the influence of *GazeAvailability* on those two modalities.

Linguistic Data

To prepare the recorded data for further processing, we applied a standard linguistic preprocessing pipeline. We first transcribed the audio signal, which was a manual step as the discourse collected in our study was very specific and also contained ungrammatical utterances and disfluencies. We then aligned the text to the audio signal by applying the

forced alignment technique (Kisler, Schiel, & Sloetjes, 2012). We performed lemmatization and part-of-speech (POS) tagging followed by linguistic annotation using shallow syntactic analysis. These annotations were automatically carried out using the TreeTagger (Schmid, 1995). Further, two types of feedback instances, positive and negative, were automatically recognized by searching for words that express feedback (simple string matching), e.g. "Ja, genau!" (*Yes, exactly!*) is a positive instance and "Nein, falsch!" (*No, that is wrong!*) is a negative one. However, in some cases those words did not express feedback, but had a different grammatical function and meaning, e.g. "ne" was used to express a negation "Nein!"(*No!*) but also as an abbreviated feminine indefinite article in German "eine" (*a*). Therefore, a manual post-correction was carried out to filter out incorrectly detected instances, and also to add a few other words that are not typical for feedback but had this function in a particular context, or different spellings more usual for spoken language, e.g. "jep". Then we were able to assess the proportion of positive and negative feedback instances per condition. Lastly, we evaluated what kind of referring expressions were produced by the speakers and whether our manipulation had an influence on that. We came up with an annotation scheme suitable for our setting and labeled the descriptions present in the corpus. The span of each micro-scale task was also manually annotated, i.e. from when the walker had reached a target location (table) until all target objects were found and selected, which essentially distinguished two activities: walking around and the stationary search for target objects.

Statistical analyses were conducted in the R statistical programming environment (R Core Team, 2014). We assessed statistical significance utilizing linear mixed-effects models using the lme4 package in R and model selection in order to determine the influence of *GazeAvailability*. As proposed by Bates, Kliegl, Vasishth, and Baayen (2015), we started out with the maximal model fitting our assumptions with respect to the random effects structure.

Eye Movement Data

We first detected fixations using a standard dispersion-based fixation detection algorithm as in Salvucci and Goldberg (2000) that declares a sequence of gaze points to be a fixation if the maximum distance from their joint center is less than 5% of the scene camera width and the sequence has a minimum duration of 66 msec. Eye movements between two

Fixation	rate, mean, max, variance of durations; mean, variance of variance within one fix.
Saccades	rate, ratio of (small/large/right/left) sacc.; mean, max, variance of amplitudes
Combined	ratio of saccades/fixations
Wordbooks	number of non-zero entries; maximum and minimum entries as well as their difference for n-grams with $n \leq 4$
Ratios	all fixation, saccade and combined features in ratio to the value over the whole trial for a particular pair and condition.

TABLE 4.1: Features extracted from human visual behavior.

fixations were considered saccades without further processing. Blinks were not included as video-based eye trackers, such as Pupil, do not record them by default. We then used a sliding window approach with a window size of 500 msec and step size of 250 msec to extract eye movement features, resulting in a dataset consisting of 18841 time windows.

For each window, we extracted a subset of 45 features of those previously proposed for eye-based recognition of visual memory recall processes (Bulling & Roggen, 2011) and cognitive load (Tessendorf et al., 2011). We added 21 additional features relating current gaze behavior to the overall gaze behavior of the current person in the current experiment, e.g. the ratio of the small saccade rate in the whole experiment to the small saccade rate in this time window. Inspired by Bulling, Ward, Gellersen, and Tröster (2011) we extracted the set of features shown in Table 4.1. For feature selection we used the minimal-redundancy-maximal-relevance criterion (mRMR) which aims to maximize the feature's relevance in terms of mutual information between target variable and features while discarding redundant features (Peng, 2007). For our analyses we relied on data driven method and used the consistently top-ranked features for target variables such as *GazeAvailability*, *Pair* or *FeedbackPresence* and fitted linear mixed-effects models to the top-ranked feature according to mRMR (saccade rate). Similar results can also be achieved based on further top-ranked features such as the ratio of small to large saccades (where a saccade is considered small if its amplitude is less than twice the maximum radius of a fixation).

4.2 Results

The results presented in this section are based on the measures collected during *micro* tasks. The reason for analyzing only this data is because the eye tracker was calibrated for the micro setting, as calibration for the macro task would require adjustment of the scene camera during the experiment.

4.2.1 Performance

Our overall performance measure is the task completion time measured from the beginning of an instruction until the success button was pressed by the walker to mark the time point of target identification. There were only a few confusion button presses, and all tasks were correctly solved. We analyzed the task completion time in each condition to reveal whether listener gaze was used to complete a task more efficiently. There were no significant differences obtained for the performance measure ($\chi^2(2) = 1.722, p = 0.423$). Figure 4.2 illustrates that participants were comparably fast in all three conditions, NoGaze ($M = 15.030\,sec, SD = 6.750\,sec$), Gaze ($M = 15.300\,sec, SD = 7.105\,sec$) and ManGaze ($M = 14.609\,sec, SD = 6.352\,sec$). Since the average interaction time was generally very low, a floor effect may have prevented a distinction of the conditions.

4.2.2 Linguistic Analysis

Length of Utterances Next we examined the intuition that the utterances can differ in length; that is, we expected shorter instructions in the Gaze condition compared to the other conditions due to possible usage of deixis, given that the current focus of attention was provided. So we investigated the number of words needed to describe a target object involved in a *micro* task. There were no significant differences with respect to the different levels of *GazeAvailability* Gaze vs. NoGaze ($\beta = -0.052,\ SE = 0.057,\ z = -0.90,\ p = 0.367$) and Gaze vs. ManGaze ($\beta = -0.023,\ SE = 0.052,\ z = -0.43,\ p = 0.665$), respectively. Specifically, the amount of words uttered by the speaker including disfluencies and feedback was similar in the NoGaze ($M = 21.76\ words, SD = 10.79\ words$), Gaze ($M = 23.12\ words, SD = 11.89\ words$) and ManGaze ($M = 22.36\ words, SD = 9.53\ words$) conditions.

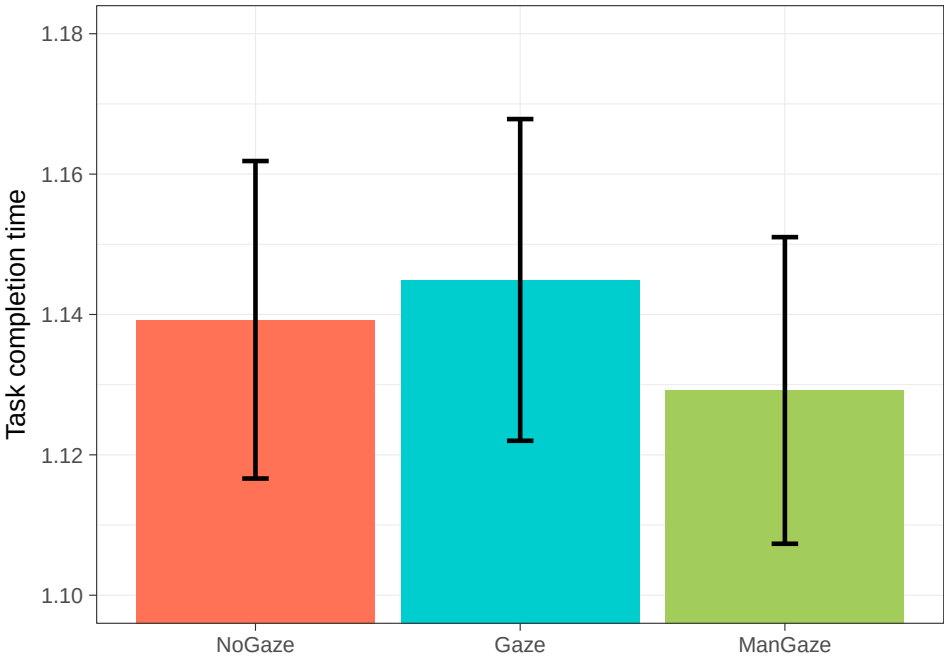

FIGURE 4.2: The task completion time (log transformed with 95% CI error bars) for the individual trials in the micro task

Verbal Feedback We then investigated the proportion of the detected feedback instances. As already mentioned, there are two types of feedback, positive (**pos**) and negative (**neg**). Positive feedback confirms correct understanding of an instruction and occurred more frequently in this setting. Negative feedback aims at signaling misunderstandings and introduces repairs in the linguistic content of a previous utterance. To test if the difference in the proportions was significant, we constructed a generalized linear mixed-effects model (with a logit link function) fitted to *FeedbackType* with *GazeAvailability* as a fixed effect.

Figure 4.3 depicts a graph that shows the proportion of feedback in the different gaze conditions and gives the model specification. The amount of data points (feedback instances per pair) does not license the inclusion of a random slope in the model, so we include only the random intercept for *Pair*.

This model shows a difference between the GAZE and NOGAZE condition that approaches significance ($\beta = 0.574,\ SE = 0.314,\ z = 1.829,\ p = 0.067$). This marginally significant

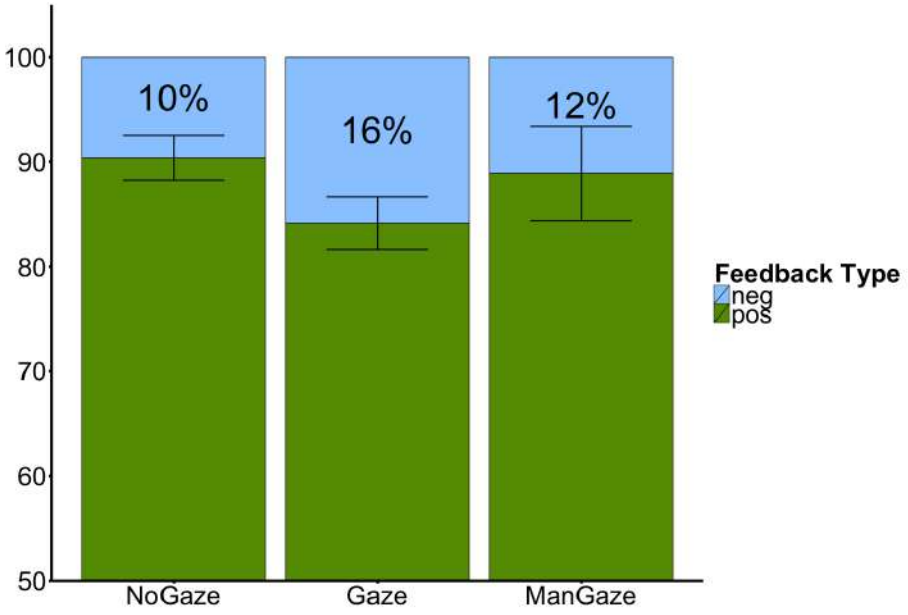

FIGURE 4.3: The proportion of positive and negative feedback instances in the different conditions. The model fitted to that data is the following `feedbackType ~ GazeAvailability + (1|Pair)`

difference indicates that speakers make use of the exact gaze positions of the listeners and that they utter more negative feedback to signal misunderstandings. MANGAZE (12%) falls somewhat inbetween GAZE (16%) and NOGAZE (10%).

Moreover, a negative feedback instance is usually followed firstly by a repair, i.e. an additional description that either provides complementary information that was not mentioned in the instruction before, or an alternative description that describes a distractor which is usually underspecified. Secondly, a positive feedback instance often follows to confirm the successful resolution of the repair. Example (3) illustrates that repeated pattern.

(3) "ne das andere ... Genau" (*no the other one ... exactly*)

We further explored if these repairs differed with the availability of gaze: We measured the length (in words) of the repairs and compared them across all conditions. For this measure there were also no significant differences found.

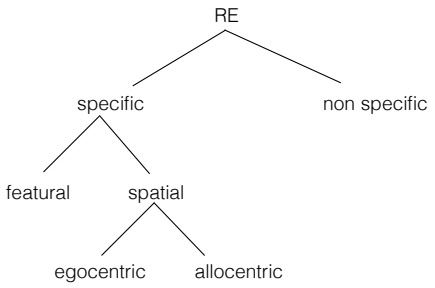

FIGURE 4.4: The annotation scheme for categorizing referring expressions.

The number of words is a rather coarse measure that investigates the quantity of the recorded speech but cannot capture the semantic content of the utterances. Thus we further investigated the types of referring expressions that were produced by the instructors during the interactions.

Referring Expressions Speakers used referring expressions to describe the target objects. The type of the produced referring expressions provides some insights about the kind of information speakers chose to describe an object. In order to evaluate whether there was a systematicity of mentioning specific object properties, like color or position on the table, and relate it to our manipulation, we developed an annotation scheme to categorize a referring expression.

In Figure 4.4 the annotation scheme is depicted. There is a general differentiation between *specific* vs. *non-specific* references and we expected to observe more *specific* expressions given the task under consideration. Further, a *specific* expression can be sub-categorized into *featural* or *spatial*. That is, we investigated whether speakers tended to use object properties or its spatial location to describe it for the listener who had to identify and collect it. For this distinction, we expected to see more *featural* descriptions because human speakers tend to select absolute object attributes (Belke & Meyer, 2002), which would be perceived more easily and then if needed, include a spatial description for clarification. Lastly, we split the *spatial* category into *spatial-egocentric* or *spatial-allocentric* in order to examine if the speaker takes the perspective of the listener (egocentric) or instead considers the visual scene and relates closer objects to the target position.

The possible categories for a referring expression were defined as follows:

- **non-specific**: general, non-exhaustive expressions, e.g. *the pen*
- **specific**: explicit or definite expressions
  - **featural**: not comparable, uses the properties of the object to describe it, e.g. *the blue pen*
  - **spatial-egocentric**: linked to the walker's current location, e.g. *the pen that is furthest away from you*
  - **spatial-allocentric**: linked to a reference frame based on the visible scene and independent of the walker's current location in it, e.g. *the pen that is next to the notepad*

We sampled a small random subset of our corpus (11 descriptions) and asked two annotators to assign a referring expression occurring in the description to one of the categories. They had a very high agreement; all but one of the expressions were assigned to the same category. We then split the corpus in two parts and each annotator labeled one half.

We compared the mean occurrences of *specific* vs. *non-specific* referring expressions per trial. For the statistical analysis, we constructed a generalized mixed-effects model (with Poisson distribution) fitted to *REsOccurrences* with *GazeAvailability* and *Category* as fixed effects.

There were no significant differences with respect to *GazeAvailability*: Gaze vs. NoGaze ($\beta = -0.021,\ SE = 0.082\ z = -0.250,\ p = 0.803$) and Gaze vs. ManGaze ($\beta = 0.015,\ SE = 0.082,\ z = 0.183,\ p = 0.855$). This suggests that speakers did not incorporate the listener gaze position while planning their utterances. However, in agreement with our expectations, the majority of the produced referring expressions were *specific*: NoGaze ($M = 2.48\ inst, SD = 0.95\ inst$), Gaze ($M = 2.54\ inst, SD = 0.86\ inst$) and ManGaze ($M = 2.57\ inst, SD = 0.81\ inst$) and fewer *non-specific* ones were identified: NoGaze ($M = 1.52\ inst, SD = 0.65\ inst$), Gaze ($M = 1.46\ inst, SD = 0.68\ inst$) and ManGaze ($M = 1.38\ inst, SD = 0.57\ inst$). Specifically, there was a main effect of *Category* ($\beta = -0.553,\ SE = 0.097,\ z = -5.709,\ p < 0.001$). This result is not surprising and can be explained by the nature of the task under consideration. In order to enable the

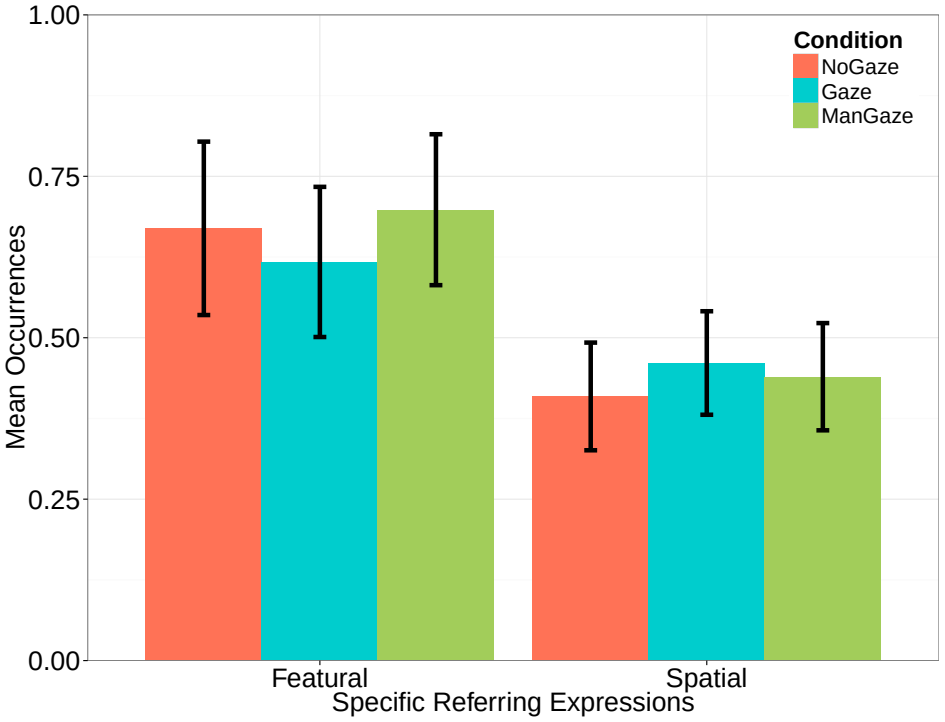

FIGURE 4.5: The average number of *sub-specific* referring expressions per trial in the micro task (95% CI error bars)

listener to precisely identify a target object, *specific* expressions were predominated. The presence of *non-specific* expressions can be explained by the fact that speakers described the objects spontaneously and sometimes they were unsure about an object type, e.g. "das Blatt oder der Block" (*the sheet or the notepad*).

Figure 4.5 depicts the mean occurrences one level deeper in the annotation scheme, namely how many referring expressions were categorized as *featural* vs. *spatial*. The former specifies identifying object features such as color, size and type, whereas the latter specifies the location of the searched-for target object. Again there were no significant differences between GAZE and NOGAZE ($\beta = 0.081$, $SE = 0.163$, $z = 0.498$, $p = 0.619$) and GAZE and MANGAZE ($\beta = 0.123$, $SE = 0.163$, $z = 0.757$, $p = 0.449$). However, the model revealed a marginal effect of *Category* ($\beta = -0.292$, $SE = 0.153$, $z = -1.907$, $p = 0.056$). That is, speakers tended to produce more *featural* expressions (NOGAZE ($M = 0.67$ *inst*, $SD = 0.75$ *inst*), GAZE ($M = 0.62$ *inst*, $SD = 0.63$ *inst*) and MANGAZE ($M = 0.70$ *inst*, $SD = 0.63$ *inst*)) than *spatial* ones (NOGAZE ($M =$

0.40 $inst, SD = 0.66$ $inst$), GAZE ($M = 0.46$ $inst, SD = 0.61$ $inst$) and MANGAZE ($M = 0.43$ $inst, SD = 0.64$ $inst$)), which conforms to our assumptions. In other words, the speakers specified a property of the target object to contrast it from other co-present objects and shift the listener's attention to a relevant object. Then, they further described the location of the objects as complementary information to ensure finding the current target.

Further, the specific spatial expressions can be assigned to two sub-categories indicating the type of information they contain (*allocentic* vs. *egocentric*). For this subset, we obtained the same pattern as for the other categories, namely that there was no influence of *GazeAvailability* on what kind of referring expressions were uttered: GAZE vs. NOGAZE ($\beta = -0.416,\ SE = 0.493,\ z = -0.845,\ p = 0.398$) and GAZE vs. MANGAZE ($\beta = 2.262,\ SE = 0.332,\ z = -0.028,\ p = 0.978$). However, our analysis revealed a main effect of *Category* ($\beta = -0.553,\ SE = 0.097,\ z = 6.811,\ p < 0.001$). Specifically, we observed that speakers described the spatial location more often using *allocentric* expressions (NOGAZE ($M = 0.76$ $inst, SD = 0.75$ $inst$), GAZE ($M = 0.83$ $inst, SD = 0.63$ $inst$) and MANGAZE ($M = 0.79$ $inst, SD = 0.70$ $inst$), and only rarely using *egocentric* expressions (NOGAZE ($M = 0.06$ $inst, SD = 0.23$ $inst$), GAZE ($M = 0.09$ $inst, SD = 0.28$ $inst$) and MANGAZE ($M = 0.09$ $inst, SD = 0.28$ $inst$)). This could presumably be because taking the listener's perspective is more difficult than focusing on the visual scene when planning an utterance and might not be very efficient in such setups.

4.2.3 Visual Behavior Analysis

To assess the role of listener gaze in this scenario, we examined the interplay of utterances, listener gaze and the *GazeAvailability* manipulation.

First, we fitted a linear mixed-effects model with a random intercept and random slope for *pair* to the dataset consisting of all (sliding) time windows (18841 in total). We found a significant main effect of *UtterancePresence* through model selection ($\chi^2(1) = 9.54, p < 0.01$). *GazeAvailability*, in contrast, had no effect on model fit.

We then considered feedback expressions which are a specific form of utterance and commonly occur in situated and spoken interaction: Such phrases typically form a direct and closely time-locked response to changes in the situation or, more crucially, the listener's

behavior. Similarly to the analysis of utterances in general, we fitted a linear mixed-effects model, this time with *FeedbackPresence* as a factor. We observed a main effect ($\chi^2(1) = 80.63, p < 0.001$) and an interaction with *GazeAvailability* ($\chi^2(2) = 9.38,\ p < 0.01$). The interaction suggests that the manipulation of gaze availability has some effect on how listeners move their eyes during verbal feedback, compared to before or after it. This observation also seems to be in line with the results of the linguistic analysis according to which the proportion of positive and negative feedback instances vary in the different levels of *GazeAvailability* to the speaker.

Taken together, the results from gaze behavior in *UtterancePresence* and *FeedbackPresence* indicate that gaze patterns differ depending on whether speech is happening or not, i.e. when the listener is processing speech compared to when she is not currently listening to an utterance, and that this is relatively independent of *GazeAvailability*. In light of the symptom-signal distinction, this suggests that language comprehension processes drive the ocular system (*symptom*) but that deliberate control of gaze, e.g. using it as pointer in the GAZE but not the NOGAZE condition (*signal*), hardly affects overall gaze patterns.

Furthermore, we attempted to break up the reciprocal nature of the interaction between listener gaze and speech by considering the temporal order of gaze events and speech events. Examining how gaze affects utterances and then, in turn, how the utterances affect eye movements helps us to shed light onto the dual role of listener gaze: On the one hand, it can be seen as a sign that helps the walker to communicate with the instructor (as the instructor can observe the walker's behavior but cannot hear the walker). In this case, gaze patterns may differ between the GAZE and NOGAZE conditions *before* an utterance, since in the former condition gaze may be more frequently used as a *signal* to which the speaker reacts. On the other hand, gaze may be mostly a *symptom* that reflects language processing and which therefore may also reflect when the speaker adapts to seeing listener gaze (GAZE condition) and produces utterances accordingly. In that case, gaze patterns are likely to differ with *GazeAvailability* immediately *after* utterance offset.

Thus, analogously to the analyses above, we fitted linear mixed-effects models on a subset of the data, namely the time windows immediately *before* the onset and *after* the offset of an utterance. Both subsets consist of 954 instances and we found that the factor *GazeAvailability* significantly contributes to a better model fit, not only *before* an instruction ($\chi^2(2) = 9.77,\ p < 0.01$) but also *after* it ($\chi^2(2) = 10.89,\ p < 0.01$). The same analysis was carried out for the time windows *before* and *after* positive and, additionally,

before and *after* negative feedback occurrences. However, no effect of *GazeAvailability* was observed (which may also be due to the lower number of samples).

To conclude, we observed no significant difference in gaze behavior with the *GazeAvailability* manipulation, but gaze patterns were distinct from each other in the presence and absence of utterances in general and feedback in particular. The analyses taking temporal aspects of the gaze and speech events into consideration showed that listener gaze significantly differs *before* and *after* instructions. This evidence supports the view that listener gaze can not only be seen as a *symptom* of language comprehension but also a non-verbal *signal* to the speaker. The latter role is comparable to the role of verbal deictic expression like *"Do you mean that one there?"* which could have been used in a bidirectional verbal dialogue.

4.3 Discussion

In this exploratory study, we observed that the manipulation of the availability of listener gaze position to the speaker had a main effect on listener gaze *before* and *after* an utterance, but not while an instruction was spoken. *GazeAvailability* further affected the type and amount of feedback given by speakers. In particular, GAZE differed significantly from NOGAZE, with MANGAZE being inbetween those two conditions with respect to the amount of negative feedback uttered by the speaker. This suggests that manipulated gaze was used somewhat less than natural gaze, but was not ignored either. Surprisingly, *GazeAvailability* did not affect the type of referring expressions produced in the course of the interactions. This observation suggests that the speakers were not able to integrate that information while planning an object description. It was possibly too demanding for them to constantly follow and interpret the gaze cursor, which is continuous and could rapidly move, while they were concentrated on how to precisely describe a target object. However, we have seen that the majority of the produced expressions are *specific* in all conditions and that this is task dependent. More general, *non-specific* expressions cannot uniquely identify a target and are consequently not often used. Interestingly, speakers used more *featural* than *spatial* descriptions to specify a target, suggesting that mentioning the featural characteristics is essential for the object identification, whereas spatial

expressions complement them. Additionally, most of the spatial expressions were *allocentric* and only a few *egocentric* ones were produced. This indicates that speakers focused on the visual context, and encoded information about the location of the target relative to the location of the other objects, rather than relative to the listener's location.

Based on the combination of gaze effects *before* and *after* an utterance and the lack of such an effect on eye movements *during* an utterance, we further assume that listener gaze can be seen as both a *signal* from listeners for conveying some sort of information to the speaker and as a *symptom* that reflects the language comprehension processes. The tendency of speakers to produce more negative feedback with gaze availability also supports the role of listener gaze as a *signal* to which instructors actively react. These feedback instances have the potential to quickly eliminate wrong beliefs by the listener about intended referents. We did not find an improvement of performance in terms of time needed for task completion in the GAZE condition, but we believe that this could be due to a ceiling effect.

Similarly, we did not find a significant effect of *GazeAvailability* on other coarse-grained measures of the spoken material such as utterance length (in words). However, many words do not necessarily carry more information. Further, *GazeAvailability* did not influence the type of referring expressions, but we observed some task-specific patterns that allow the walker to precisely identify a target.

In sum, human instruction givers seem to be very efficient at producing referring expressions that uniquely identify a target object among many others in real, hard referential scenes. Unlike results in the joint attention literature investigating face-to-face social interactions, where gaze is a helpful information source and facilitates coordination of turn taking and reasoning about intentions of the conversational partner (e.g. Foulsham, Cheng, Tracy, Henrich, & Kingstone, 2010), here the availability of listener gaze that indicates the current focus of attention does not contribute to faster task solving. This could be due to the specific setting, with listener gaze projected as a cursor. Another explanation for why they could not constantly exploit the gaze information is possibly because speakers were concentrating on producing a unique description of a target in an overloaded scene. In contrast, an NLG system (as its output can be fully controlled) may take advantage of this additional information. Specifically, it can provide proactive feedback andthus optimize the interaction by achieving better performance in virtual environments as shown by Garoufi et al. (2016). However, the open question remains whether

these findings are evident in real environments that are noisier than virtual ones and do not abstract from individual differences, e.g. head and hand movements of each person are different. Furthermore, listener gaze can be integrated in an REG algorithm to incrementally deliver an object description to the user in subsequent chunks and thereby realize the notion of referring in installments. But another open question is whether this approach could lead to efficient human-machine collaboration. To our knowledge, only Fang et al. (2015) addressed this issue, but they focused mainly on gesture and embodiment in their work. Hence we present two NLG systems dedicated to answering the above-mentioned open research questions in the next chapter.

Chapter 5

Gaze-driven Interactive Instruction Generation in the Real World

Designing an interactive system that communicates with the user in natural language is a challenging task. What could be especially difficult are human-machine interactions that take place in a shared physical environment where many factors influence communicative success. Therefore a system that should take actions depending on the interaction state could benefit from exploiting non-verbal cues. Interactive systems can in this way become more attentive, efficient and friendly.

As mentioned in Section 1.1, gaze is an important indicator of language comprehension and can be used to predict 1) what the speaker is about to say next and 2) how the listener will resolve a reference. However, gaze is a very rapid, continuous signal and thus it is not quite straightforward to decide when such eye movements are informative and in particular how to react to them. Listener gaze can be also misleading because it is dynamic, fast and continuous information source. A listener looks not only at an intended and understood object but to other co-present objects that are similar to it and share some features like type, color or size. Thus this information can sometimes be as noisy and ambiguous as language. The key question is to identify informative eye movements that correspond to an utterance. This is challenging in dynamic situated communication, where many utterances occur and listeners perform visual search to identify and find specific objects. It is not clear how long a fixation has to last in order to be considered to reflect the intention of the listener to perform an action. However, human speakers can

often successfully interpret such signals immediately and adequately react to them. There is evidence that an NLG system can use gaze in virtual environments and, importantly, that this facilitates collaboration and improves performance (Garoufi et al., 2016; Koller et al., 2012; Staudte et al., 2012). However, whether this is similar in real environments, which are noisier and motion cannot be as controlled as in virtual environments remains unclear. Further, it is unexplored whether using listener gaze as a trigger for incremental generation of natural language instruction can lead to smooth and efficient interaction.

In this chapter we present a multimodal interactive system (`GazInG`) and two NLG systems embedded in it. The system assists and verbally guides a human listener during an assembly task in the real world. Both NLG systems use listener gaze aiming to improve referential success and to offer more interactive communication thereby encouraging the user to better engage with the system during collaborative task solving. The NLG system *"Feedback"* gives either long, UNAMBIGUOUS or short, AMBIGUOUS instructions and uses listener gaze to proactively generate verbal feedback on object inspections. This system uses listener gaze directly to provide either UNDERSPECIFIED ("No, not that one!") or CONTRASTIVE feedback ("Further left!"), i.e. it specifies the position of the target relative to the current gaze position of the listener. This system is presented in our conference paper in the proceedings of the Annual Meeting of the Cognitive Science Society 2018 (Mitev et al., 2018) and in a journal paper for the special issue "Attention in Natural and Mediated Realities" of the journal "Cognitive Research: Principles and Implications" of the Psychonomic Society. The NLG system "Installments", in contrast, integrates the listener gaze rather indirectly into the generation mechanism. Specifically, it provides an object description incrementally, in subsequent chunks, and specifies all features of the target object as well as its absolute (viewer-centered) position. In other words, this system refers to objects in gaze-driven INSTALLMENTS, but it can generate a long description containing all chunks and output them at once (NOINSTALLMENTS). Moreover, it presents the information in different order by either including the *SpatialDescriptor*, i.e. the location of the searched-for target in the workspace, on the FIRST or the SECOND position in the instruction.

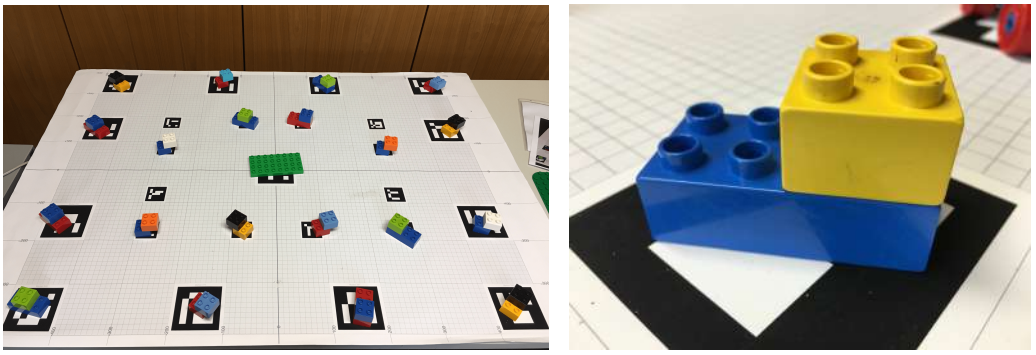

FIGURE 5.1: The workspace comprises 20 composed objects spread on a table (left picture); and a close-up view of a composed target object (right picture).

5.1 Use Case and Task

The use case we consider, for which an assistance system would be appropriate, is assembly. This scenario involves goal-oriented teamwork. In particular, a speaker gives instructions to a listener who performs actions. An important step before putting together any pieces to assemble a whole object is to identify the right missing element at any time in the assembly process. To avoid having to undo wrong steps, which negatively influences performance, it is important to select a specific object. This may not always be easy, especially if the workspace is overloaded and many similar objects are available, but tracking listener gaze can help. We designed a task that involves such an interaction in a dynamic setting and allows us to study the mutual influence of listener gaze and speech. The target domain of our scenario is LEGO DUPLO. This domain is suitable for our setup as the building blocks are of convenient size, while allowing a multitude of combinations and various ways of assembly. The workspace is overloaded such that it is challenging to automatically generate a unique identifying referring expression. Our findings presented in Chapter 3, namely that listener gaze is beneficial in hard referential scenes, further motivate the complexity of the visual context for this task.

Figure 5.1 depicts the workspace (left picture). A layout consists of 20 composed objects in total and eight targets to be collected. Each composed object comprises two simple building blocks (see Figure 5.1 (right picture) for a close-up view). For each target object there are at least two competitors available in the workspace.

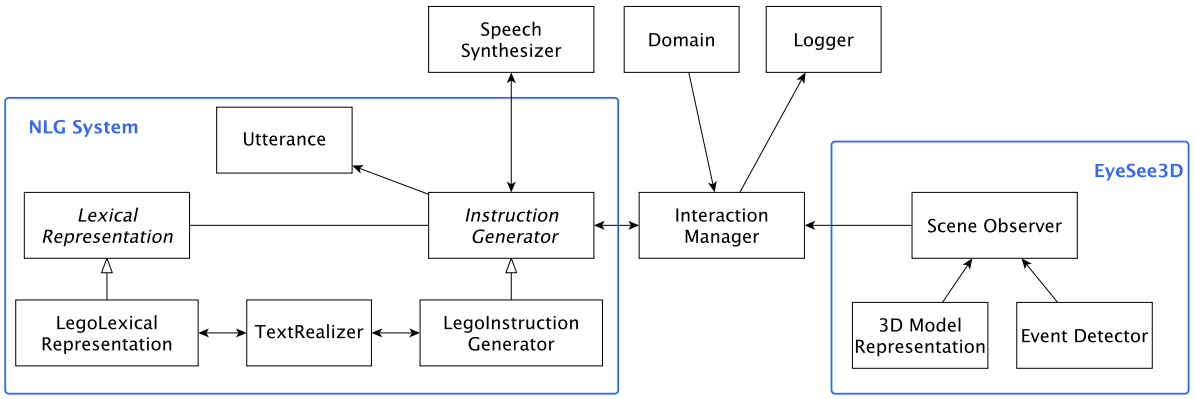

FIGURE 5.2: This diagram depicts the modular software architecture of the GazInG system.

5.2 Gaze-sensitive Instruction Generation in the Real World

In this section we introduce the assistance system, `GazInG` (Gaze-driven Instruction Generation), that supports a user during a real-world object identification task. We present two NLG systems that make use of listener gaze to augment their referring expression generation algorithms.

5.2.1 GazInG: Multimodal Interactive System

The multimodal interactive system `GazInG` monitors listener gaze and interprets object inspections and so attents to the listener. In other words, whenever the listener looks at a co-present object and considers to picking it, the system evaluates the gaze signal and responds to the listener respectively. The two different NLG systems embedded in it are targeted at generating instructions in natural language and contain referring expressions describing a specific target object. Both systems implement two different methods for generating an instruction: they can either provide a long, exhaustive description or split the description into different chunks and output them sequentially.

The modular design of the system makes it flexible, easily extendable and adaptable to other domains. Figure 5.2 depicts the system's architecture. The core software component is the `InteractionManager` which steers the interaction flow. It is coupled to the

EyeSee3D module that transfers the real scene into a 3D virtual model, which is necessary for the semantic mapping of object inspections. The `InteractionManager` has access to the `Domain` knowledge, where the characteristics of the real-world objects are stored. On the basis of the properties of a target object (color, size and position), the `InstructionGenerator` uses the `LexicalRepresentation` and generates an `Utterance`, i.e. instruction in natural language. The `InstructionGenerator` is connected to the `SpeechSynthesizer` (MaryTTS (Schröder, Charfuelan, Pammi, & Steiner, 2011)) in order to obtain an auditory version of the generated text and output spoken instructions on request by the `InteractionManager`. The target language we used is German and the `TextRealizer` we chose is SimpleNLG (Gatt & Reiter, 2009).

The programming language used for the implementation of the NLG systems is Java. The different modalities are synchronized and aligned by making use of thread programming and the interaction flow is realized with event-based programming.

Augmented Reality: EyeSee3D Module

EyeSee3D was developed to enable real-time analysis of mobile gaze-based experiments. The central idea is to model the environment as a 3D situation model in which the stimuli are represented (see turquoise arrow in Figure 6.1). The model can be created by scanning the environment (e.g. using a Microsoft Kinect) or, as done here, manually using abstract geometries like boxes.

In order to integrate the user's head position and orientation into the model, the environment is instrumented with low-cost printable fiducial markers (see cloth table in Figure 6.1). These are located in previously known positions relative to the stimuli. The scene camera of the mobile eye tracker is then used to detect and track the markers. If at least one marker is visible in the scene camera, the head position and orientation can be calculated.

Computed from the user's gaze direction, a 3D gaze ray can then be cast into the situation model (see yellow arrow in Figure 6.1). By testing intersections of the ray with the modeled stimuli, objects of interest being gazed at can be identified. In this way the semantic mapping of the listener's inspections is realized, i.e. onto which real-world object

inspections are detected. For more technical details of the approach see Pfeiffer and Renner (2014).

Modality Alignment and Synchronization

For such situated interactions, the temporal alignment of the different modalities is crucial and challenging. The synchronization can be an issue if it fails. The eye-tracking signal is continuous and noisy because eye movements are rapid, i.e. they can quickly jump from one object to another. Thus, it is important to decide on which fixations should be interpreted and used to trigger feedback, and when to output this feedback. A critical parameter is the inspection threshold, i.e. how long a fixation should last to be considered as an indicator of the listener's intention to pick up the gazed-at object. Inspired by Garoufi et al. (2016), who dealt with long distances between listener and target, i.e. targets which were not directly within the participant's reach, we set the inspection threshold initially to 300 ms. However, we adjusted the threshold to 200 ms on an empirical basis as we are dealing with short distances between user and targets, so that objects can be reached very quickly. System instructions were not self-interrupted and feedback occurred only after an instruction ended (offset). In advance, we ran two preliminary studies to find an appropriate object density and to determine whether the latency of the eye-tracking data streaming allows the generation of feedback on time. We used human-authored instructions in the form of canned text to refer to the different objects the listeners should identify. These were final strings that were presented to the user without changing them when the appropriate trigger (object inspection) was detected (see Appendix B).

5.2.2 NLG System *"Feedback"*: Instructions Combined with Gaze-driven Verbal Feedback

In this section, we present our first NLG system. We use a heuristic approach and implement two interaction strategies: generating short, AMBIGUOUS or long, UNAMBIGUOUS instructions. Furthermore, the system provides gaze-driven feedback triggered by inspections of competitors or the target. The feedback triggered by inspections of competitors can be of different specificity: either UNDERSPECIFIED or CONTRASTIVE.

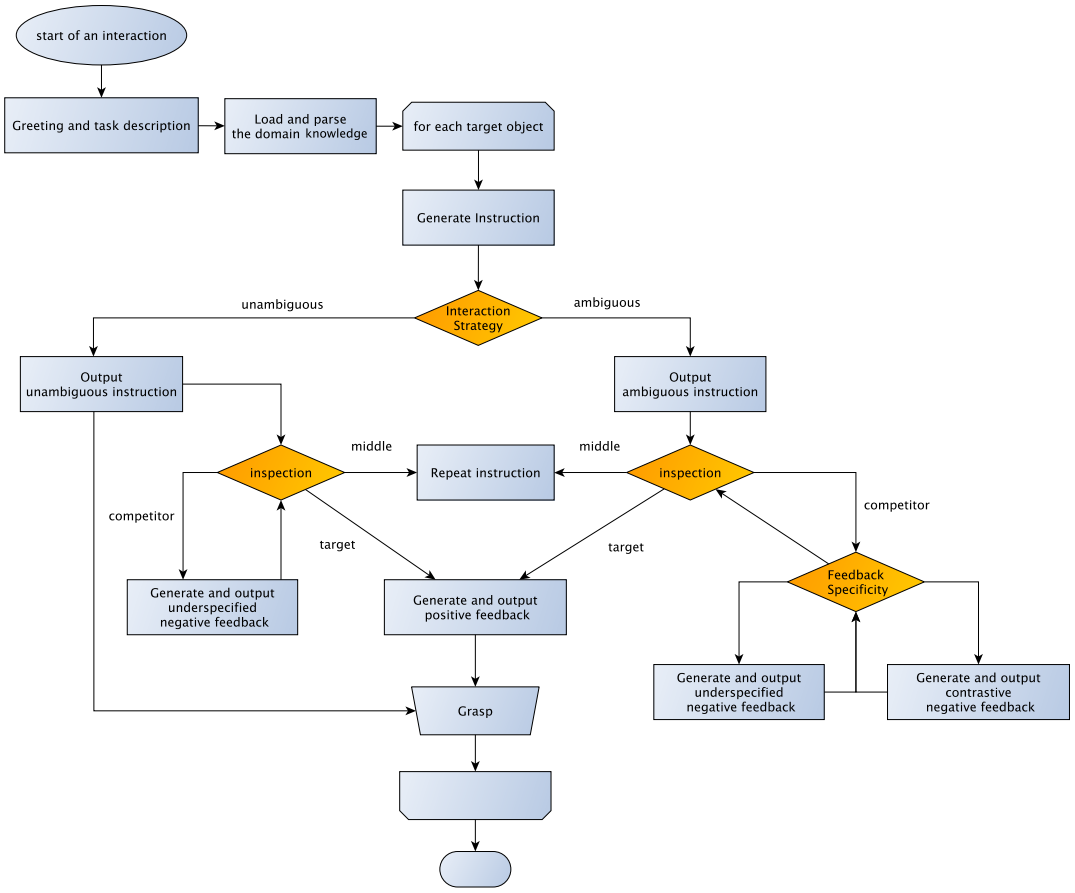

FIGURE 5.3: The generation mechnism implemented in NLG system *"Feedback"*.

On Figure 5.3 the workflow diagram of the generation mechanism is depicted. The system generates, depending on the interaction strategy, either an UNAMBIGUOUS or an AMBIGUOUS instruction. The system observes gaze behavior and interprets it. If the listener looks at a competitor object, the system generates a negative feedback instance. If the listener inspects the target object, the system generates a positive feedback instance to confirm the correct interpretation of an instruction. Typically after hearing a confirmation, listeners grasp the target and assemble it onto the other building blocks. Listeners had the option to look at the middle of the workspace (fixating on a small green LEGO plate) if they were confused in order to get further help from the system. In this case, the system repeated the initial instruction. An UNAMBIGUOUS instruction is followed by either no or UNDERSPECIFIED feedback, and after an AMBIGUOUS instruction the system provides the user with either UNDERSPECIFIED or CONTRASTIVE feedback.

NLG Heuristics Our system uses a heuristic approach to generate an instruction containing a referring expression (RE) that describes a composed object available in the workspace. The syntactic structure of the instructions is predefined. The system is able to generate AMBIGUOUS instructions consisting of a main clause that describes the bottom object:

(1) "Nimm den großen roten Baustein!" (*Pick the big red building block.*)

Size and color are used as pre-modifiers and the head noun is randomly chosen from a set of synonyms suitable for the type of object such that the instructions are not too monotonous. In order to output an UNAMBIGUOUS instruction the algorithm appends two post-modifiers additionally to describe the top object, 1) a prepositional phrase (PP):

(2) "Nimm den großen roten Baustein mit dem kleinen gelben Duploteil darauf!"
(*Pick the big red building block with the small yellow one on top.*)

or a relative clause (RelClause)

(3) "Nimm den großen roten Baustein, auf dem ein kleiner gelber Duploteil steckt"
(*Pick the big red building block, on which a small yellow one is placed.*)

and 2) an adverbial phrase containing absolute position information.

(4) "Nimm den großen roten Baustein, auf dem ein kleiner gelber Duploteil steckt, hinten links!"
(*Pick the big red building block, on which a small yellow one is placed, at the back toward the left.*)

The workspace is divided into four squares a) at the back toward the left or b) the right and c) in the front toward the left or d) the right. Providing the spatial expression disambiguates the instruction.

Verbal Feedback The system is capable of generating either UNDERSPECIFIED or CONTRASTIVE feedback. Inspections of target objects trigger positive feedback (e.g. *"Yes"*, *"Exactly"* etc.), and inspections of competitors trigger negative feedback signaling that the listener is considering the wrong object: UNDERSPECIFIED, such as *"No, not that one!"* or CONTRASTIVE, providing relative position information to compensate, such as *"Further left!"*. In the former case, the listener can exclude only the inspected competitor, which might be sufficient for simple scenes where fewer competitors are available in the visual context. In the latter case, the listener's attention is directed towards the target relative to the current gaze position. The system thereby avoids inspections of other competitors until the target is found. This makes such an interaction approach comparable to the notion of referring in installments as it splits the information into different chunks. However, in this case the second piece of information is instead in the form of feedback and is related to the current gaze position of the listener. In the next section we present our second NLG system that implements true installments, i.e. it splits a full instruction into three chunks and provides them subsequently depending on which objects the listener inspects.

5.2.3 NLG System *"Installments"*: Gaze-driven Incremental Instruction Generation

In this section, we present our second NLG system, *"Installments"*, that describes real-world objects needed by a listener for assembly in a real-time and environment. The system includes a mechanism to interpret listener gaze in the REG algorithm. We implemented two approaches for *InformationDelivery*, i.e. how to provide the listener with the required information to identify an object. Our system generates and outputs 1) the whole description at once (NOINSTALLMENTS) vs. 2) incrementally in subsequent chunks (INSTALLMENTS) triggered by object inspections. Furthermore, it varies the order of the information presented to the listener by providing a *SpatialDescriptor* as either the FIRST or the SECOND installment, i.e. the position specification of a target object is mentioned before or after a featural descriptor.

NLG Heuristics Analogously to NLG system *"Feedback"* we use a heuristic approach to generate instructions, which have a predefined syntactic structure. However, the order

of the information varies with respect to the appearance of the *SpatialDescriptor*. The system initially generates a polite request:

(5) "Nimm bitte den folgenden Baustein!" (*Pick the following building block, please!*)

After that, the description of the composed target object is specified. In this setup both color and size of the respective target need to be specified in order to identify an object. To be as concise as possible, the object type is not mentioned, because it does not contribute to the unique reference (all objects are of the same type). Instead the head noun in a referring expression is a nominalized color adjective, and the adjective that specifies the size is used as a pre-modifier. Figure 5.4 depicts the second generation mechanism implemented in the GazInG system. For each target object, initially all installments are generated, and depending on the *InformationDelivery* approach, are presented in a different manner to the listener. Two *InformationDelivery* approaches are implemented: referring to objects in NoInstallments vs. Installments (triggered by the listener's object inspections). For the former, all installments are concatenated and output at once, while the latter gives the next installment as a response to a competitor inspection. A competitor object has the same characteristics as a target object, except for location or color, depending on when the *SpatialDescriptor* is mentioned. All other objects are considered as distractors and the system does not react to distractor inspections. This system generates verbal confirmation. That is, inspections of a target object trigger positive feedback like *"Yes"*, *"Exactly"* etc. to encourage the user to grasp it. Usually after hearing an exhaustive description (NoInstallments), listeners would not consider a competitor, but if they do, our system outputs negative feedback such as *"No, not that one!"* to prevent a wrong grasp.

Additionally, as the position of the target object disambiguates a referring expression, we flip the first two installments to investigate whether mentioning the position right at the beginning influences performance. That is, *SpatialDescriptor* is generated first (2) vs. second (3).

(6) "Hinten links" ... < *competitor inspection* > ... "den großen Blauen" ...
< *competitor inspection* > ... ["mit dem kleinen Gelben darauf"] ...
< *target inspection* > ... "Ja!"

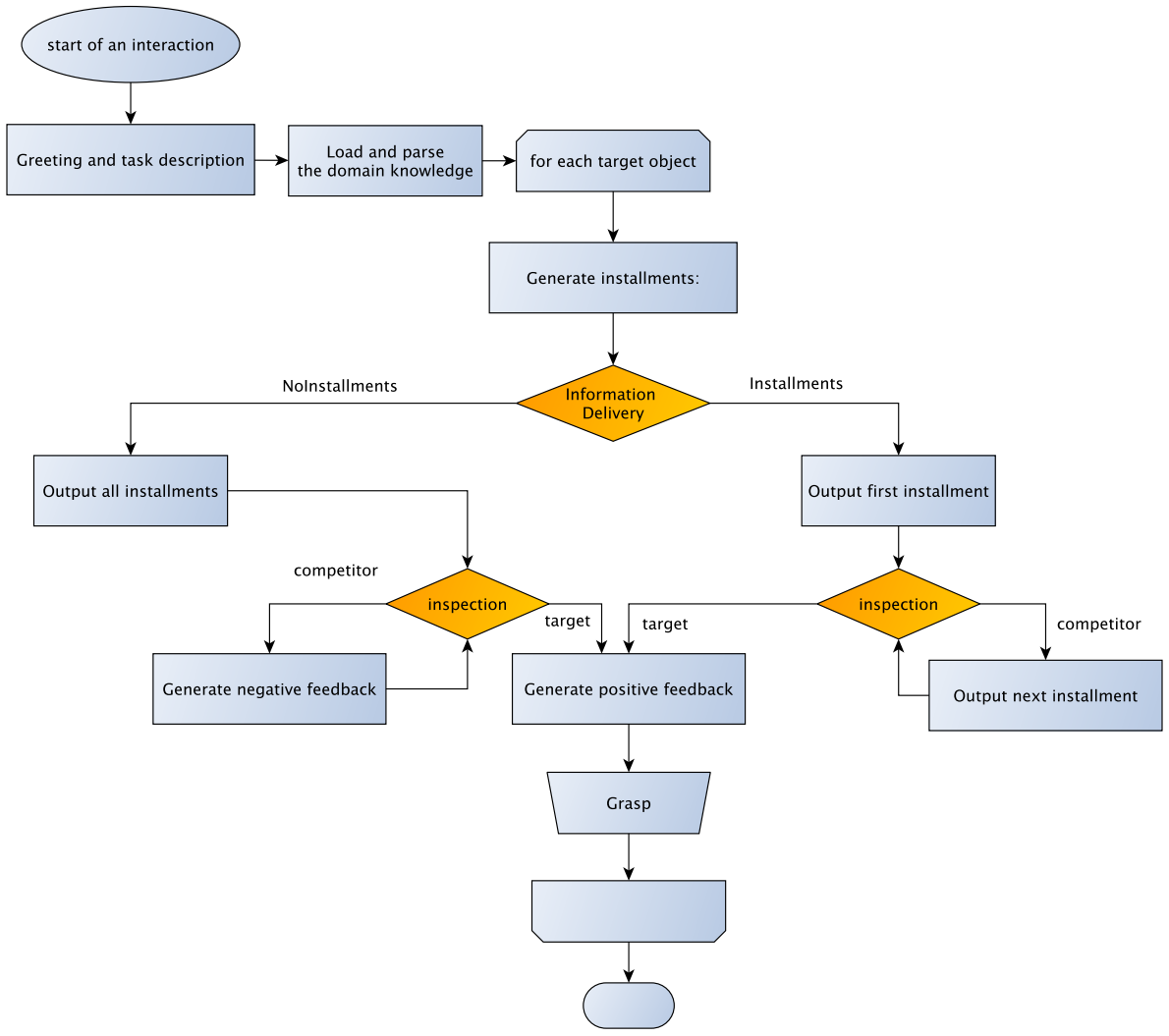

FIGURE 5.4: The generation mechanism implemented in NLG system *"Installments"*.

(At the back toward the left ...< *competitor inspection* > ... the big blue one ...< *competitor inspection* > ... [with the small yellow one on top of it] ... < *target inspection* > ... Yes!)

(7) "Den großen Blauen" ...< *competitor inspection* > ... "hinten links" ... < *competitor inspection* > ... ["mit dem kleinen Gelben darauf"] ... < *target inspection* > ... "Ja!"
(*The big blue one ...< competitor inspection > ... at the back toward the left ...* < *competitor inspection* > ... *[with the small yellow one on top of it]...* < *target inspection* > ... *Yes!*)

For the NoInstallments approach, there are two corresponding versions with respect to the order of information; see examples (4) and (5).

(8) "Hinten links, den großen Blauen mit dem kleinen Gelben darauf." ...
< *target inspection* > ... "Ja!"
(*At the back toward the left, the big blue one, with the small yellow one on top of it...*< *target inspection* > ... *Yes!*)

(9) "Den großen Blauen, hinten links, mit dem kleinen Gelben darauf." ...
< *target inspection* > ... "Ja!"
(*The big blue one, at the back toward the left, with the small yellow one on top of it...*< *target inspection* > ... *Yes!*)

5.3 Summary

In this chapter, we have introduced a multimodal instruction-giving system (`GazInG`) that is targeted to assist a user by giving verbal instructions. The use case we consider is collaborative assembly; that is, the system and the user team up to find and collect specific co-present objects needed for assembly. The system is in the role of a speaker and the user is the listener. Furthermore, our system tracks listener gaze and reacts to it. This enables consideration of attention shifts and adaptive behavior aiming at efficient communication. We developed two NLG systems to automatically generate identifying instructions. The NLG system *"Feedback"* reacts to listener's eye movements with verbal feedback, which can be either underspecified, i.e. just warning, for example, *"No, not that one!"*, or contrastive, i.e. providing additional information by specifying the spatial location, for example *"Further left!"*. In contrast, the NLG system *"Installments"* uses listener's eye movements to output an instruction incrementally. That is, it first outputs a phrase that gives partial information, and then, depending on where the listener looks, either further describes the intended target with another phrase (not in the form of feedback) or outputs a confirmation. Both systems also implement a non-interactive object description generation and so output an instruction that specifies at once all object attributes needed to identify a target, i.e. color, size and position. Our system provides a proof-of-concept that listener gaze can be used for adaptive NLG in real environments. However, it is not yet

clear if using a gaze-driven strategy for NLG is 1) beneficial for the user, i.e. it benefits the listener's understanding and leads to more efficient interactions and 2) if it is preferred by the user over a non-interactive strategy, i.e. if it feels better, namely more appropriate and natural. Importantly, `GazInG` can be used to investigate these research questions; we address them in the next two chapters. Specifically, we present two experiments investigating the interaction with the NLG system *"Feedback"* and how gaze-driven feedback affects performance and engagement in Chapter 6. Further, in another experiment presented in Chapter 7, we observed how listeners interacted with the NLG system *"Installments"* and examined if gaze-driven installments benefit understanding and if spatial information determines efficiency.

Chapter 6

Human-Machine Interaction: Effects of Listener Gaze on Performance and Engagement

Listener gaze is a reliable indicator of language understanding. We address the question of whether a listener gaze can successfully be used as a non-verbal feedback cue for adaptive instruction generation and integrate listener feedback into the interaction loop. There is some evidence from studies in virtual environments that feedback from the *artificial speaker* based on listener gaze can increase interaction efficiency (Koller et al., 2012; Staudte et al., 2012; Garoufi et al., 2016). However, there are two remaining questions that we address in the present chapter: (1) whether the successful use of *listener gaze* can be replicated in real environments, which are much more complex, noisy, and less controlled to handle technically, and (2) whether gaze-aware natural language generation can be used to generate adaptive *repairs* in the form of contrastive feedback, which further describes a target.

In this chapter we present two experiments that are designed to test the usefulness of listener gaze for adaptive feedback generation. Both experiments were designed to investigate the interaction with the first NLG System *"Feedback"*. Our results indicate that listener gaze can reliably be used to anticipate, in the real world as well, which object the listener is considering to grasp. In both experiments we obtain a very low error rate and so validate that gaze can be used to prevent wrong steps that would need to be undone.

Further, in Experiment 1 we show that CONTRASTIVE feedback improves performance and speeds up task solving. Specifically, when it is combined with an AMBIGUOUS instruction it outperforms acting out an UNAMBIGUOUS instruction, suggesting that distributing the information into subsequent chunks is beneficial. These results have been published partly published in the proceedings of the Annual Meeting of the Cognitive Science Society 2018 (Mitev et al., 2018). In Experiment 2, we found that the presence of CONTRASTIVE gaze-driven feedback influences the engagement in the interaction with the the instruction-giving system and listeners' information uptake. Both aspects influence how the listener can progress through the task. Thus, even in the more difficult condition when feedback was UNDERSPECIFIED, listeners were as fast as when they received CONTRASTIVE feedback. The findings from both experiments have been published in "Attention in Natural and Mediated Realities", a special issue of the journal "Cognitive Research: Principles and Implications" of the Psychonomic Society.

6.1 Experimental Method

In order to investigate how listener gaze can be used in a dynamic task-oriented interaction and in particular how listeners engage with an artificial speaker, we conducted two experiments. Both experiments were targeted to evaluate the interaction with the NLG system "Feedback" and to examine whether exploiting listeners' gaze to generate verbal feedback facilitates communication and contributes to efficiency. We tested two different *InteractionStrategies* (UNAMBIGUOUS vs. AMBIGOUS) and two levels of *FeedbackSpecificity* (UNDERSPECIFIED vs. CONTRASTIVE) and how these factors influence task performance and listeners' engagement with the instruction-giving system.

6.1.1 Setup and Apparatus

Figure 6.1 depicts our setup. Our system is designed to describe real-world objects to a naïve listener, who is asked to select these in real time. The system does not provide guidelines on how to put together the selected elements but leaves this to the listener's creativity.

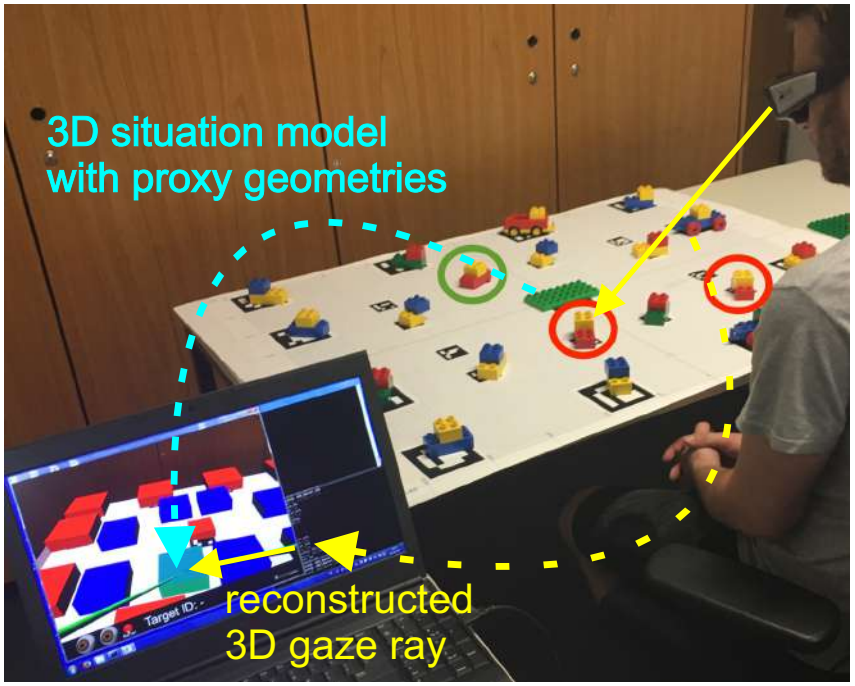

FIGURE 6.1: Setup: Listener in front of a workspace before any objects are collected. The target is circled in green and competitors in red. The listener inspects the competitor object to the left as highlighted in the virtual 3D model. EyeSee3D is used to reconstruct the gaze ray in 3D (yellow). The target domain is modeled as a 3D situation model with boxes as proxies for the assembled structures (turquoise).

We used an SMI Eye Tracking Glasses binocular head-mounted eye tracker for gaze data collection. The tracker is equipped with a high-resolution scene camera (1280 x 960) at 24Hz and two eye cameras recording at 30Hz. The eye tracker was connected to a notebook. The EyeSee3D augmented reality software (see Section 5.2.1) and the NLG system run on a Dell Precision M4800 15,6" WORKSTATION with processor I7 4900MQ at 2.8GHZ and with 16GB RAM. The speech synthesizer was located on a remote server. The communication was implemented using a client-server architecture.

6.1.2 Measures and Analysis

Both experiments included almost the same core set of objective and behavioral measures to assess the quality and effectiveness of the interaction. In Experiment 1 we assessed additionally subjective measures concerning listeners' perception of the interactions with the GazInG system.

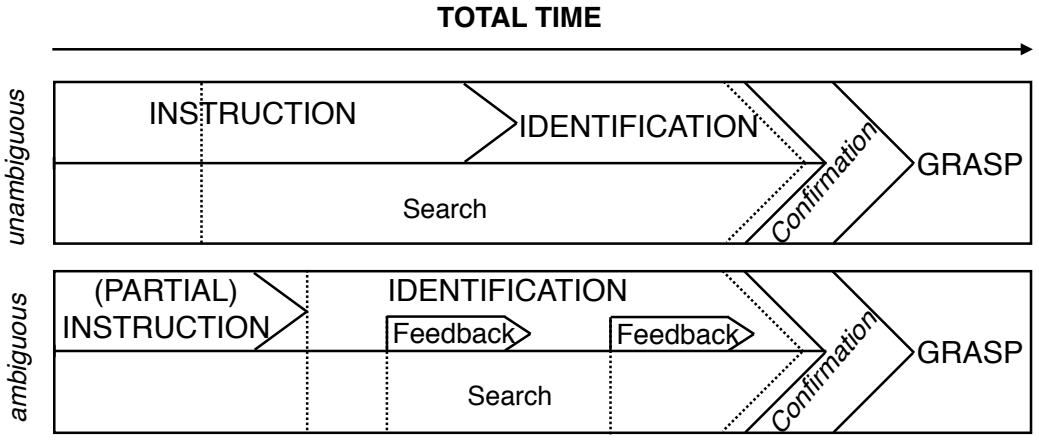

FIGURE 6.2: This diagram illustrates the interaction phases for both strategies: The spoken **instruction**, followed by **identification**, i.e. time to first target inspection, and the **grasp** of the object after a verbal confirmation is given. Visual search starts either during or after an instruction and can be interleaved with feedback, depending on the condition.

All measures were collected on a per item basis. Performance was measured by total time needed for task completion and success rate.

The total time was further divided into three phases, which differ depending on the *InteractionStrategy* (see Figure 6.2). The first phase is determined by the duration of the spoken **instruction**, from speech onset to speech offset. Secondly, we assessed the time for **identification**, i.e. the time needed from the offset of the instruction to make the first inspection to the target in Experiment 1 and 2. Finally, the time from the first target inspection until the **grasp** of the target determined the duration of the third phase. For the exhaustive, UNAMBIGUOUS instructions and NOINSTALLMENTS, the visual search starts while the system is still speaking out the instruction. This is not the case for the AMBIGUOUS interaction strategy.

Further, we derived various metrics from the eye-tracking data. For the object inspections, we examined the average number per trial, i.e. how often participants looked at the target or at one of the competitors. We obtained the duration of the inspections during and after an instruction (until finding the target) and also for the whole time span (task completion time). For speech, the only variable is feedback occurrences, as this modality

was controlled throughout the experiments. Besides the total count of feedback occurrences, we analyzed the time intervals from instruction offset to feedback onset of the first positive and also first negative feedback instance because they correspond to visual search behavior.

Statistical analyses were conducted in the R statistical programming environment (R Core Team, 2014). We assessed statistical significance using linear mixed-effects models using the `lme4` package in R and model comparison in order to determine the influence of *InteractionStrategy* and *FeedbackSpecificity*. As proposed by (Bates et al., 2015), we started out with the maximal model fitting our assumptions with respect to the random effects structure. When the models failed to converge, our approach for simplifying the random structure was to first remove the correlations between random slopes and intercepts, followed by the intercept terms, starting with the random effect for items (if present).

6.2 Experiment 1: Interaction with the NLG System *"Feedback"*

In this experiment, we manipulated the *InteractionStrategy*: UNAMBIGUOUS vs. AMBIGUOUS instructions within participants and the *FeedbackSpecificity*: UNDERSPECIFIED vs. CONTRASTIVE feedback between participants, i.e. group one was provided with no or unspecific feedback (e.g. *"No, not that one!"*) while group two received spatial information relative to the user's current fixation point (e.g. *"Further right!"*). We hypothesized that providing CONTRASTIVE feedback complementing an AMBIGUOUS instruction will shorten total interaction time compared to when feedback is UNDERSPECIFIED. The former guides listeners attention and narrows down the search space, while the latter only provides warnings to prevent wrong actions. Further, we expected to find that distributing the information in different chunks would be similarly effective as following long, UNAMBIGUOUS instructions.

	Interaction Strategy	
	Ambiguous	Unambiguous
Group 1	Underspecified Feedback	No Feedback
Group 2	Contrastive Feedback	Contrastive Feedback

Table 6.1: Interaction strategies (blocked) for each group in Experiment 1.

6.2.1 Participants

Forty-eight participants, mainly students enrolled at Saarland University, took part in the experiment. Twenty-four were assigned to group one (19 female) and the other twenty-four formed group two (18 female). The average age of the first group of participants was 25 years with a range of 19–35, and of the second, 24 years with a range of 20–31. All participants were German native speakers and reported normal or corrected-to-normal vision and no red-green color blindness. Their participation was compensated with €8 (first group) and €5 (second group) with the difference being due to the slightly shorter duration of the second group's experiment.

6.2.2 Procedure

Participants were seated in front of the workspace and asked to carefully listen to and follow the system's instructions. They were instructed to act as a team with the system and solve the task together as precisely as possible, i.e. to avoid taking the wrong building blocks. Then participants were equipped with a pair of eye-tracking glasses and a 3-point calibration procedure followed. Calibration was repeated between layouts and whenever needed. Before performing the actual task, a short practice session was completed: participants had to collect three targets among six objects in total in order to familiarize themselves with the task and the system's pace.

The experimental part consisted of two blocks, one for each interaction strategy (see Table 6.1). The order was balanced across participants. Each block consisted of one layout with a total of 20 composed objects and eight targets to be collected in each (see Appendix A). Our system did not give instructions on how to assemble the identified building blocks. However, participants were encouraged to put effort into building an

FIGURE 6.3: An example trial: System instructs the user by saying "Pick the big red building block." The listener identifies and grasps it. After that it is assembled to the other LEGO blocks (right picture). The circle represents the gaze cursor.

individual LEGO model, as an additional reward was given for the most creative one. An example trial is presented in Figure 6.3: The system gives the instruction "Pick the big red building block". The listener identifies the target (left picture). After receiving a confirmation based on looking at the target object ("Yes, that one!"), the listener takes it, hears "Well done!" and assembles it with the other blocks (right picture).

In the different experimental conditions the interaction is typically as follows:

Using the UNAMBIGUOUS *InteractionStrategy*, the system gives a long description. The listener identifies the target, and in the no-feedback condition people just grasp the uniquely described target, or get a confirmation. UNDERSPECIFIED feedback is given after an exhaustive RE, to encourage the listener to grasp the target.

(1) SYSTEM: Pick the big red building block with a small yellow piece on top of it at the back toward the left.
LISTENER: [*inspects the target*]
SYSTEM: [*no reaction or*] Yes, that one! (underspecified or contrastive)
LISTENER: [*grasps the target*]
SYSTEM: Well done!

In contrast, the AMBIGUOUS *InteractionStrategy* is more interactive and includes more turns. Initially, a partial description, which specifies the characteristics of the bottom object, is given. Then, if a competitor is inspected, the system warns the listener with UNDERSPECIFIED feedback that a wrong object is being considered, and finally the target is found and grasped.

(2) SYSTEM: Pick the big red building block.
LISTENER: [*inspects a competitor*]
SYSTEM: No, not that one! (underspecified)
LISTENER: [*inspects a competitor*]
SYSTEM: No, not that one! (underspecified)
LISTENER: [*inspects the target*]
SYSTEM: Yes, exactly!
LISTENER: [*grasps the target*]
SYSTEM: Well done!

Providing CONTRASTIVE feedback directs listeners' attention in the right direction and may require fewer turns.

(3) SYSTEM: Pick the big red building block.
LISTENER: [*inspects a competitor*]
SYSTEM: Further toward the left! (contrastive)
LISTENER: [*inspects the target*]
SYSTEM: Yes, that one!
LISTENER: [*grasps the target*]
SYSTEM: Well done!

After finishing one layout, each participant filled in a questionnaire assessing participants' perception and impressions of the interaction with the system. Finally, they answered questions about the comparison of both interaction strategies they experienced. The experiment lasted between 30 and 45 minutes.

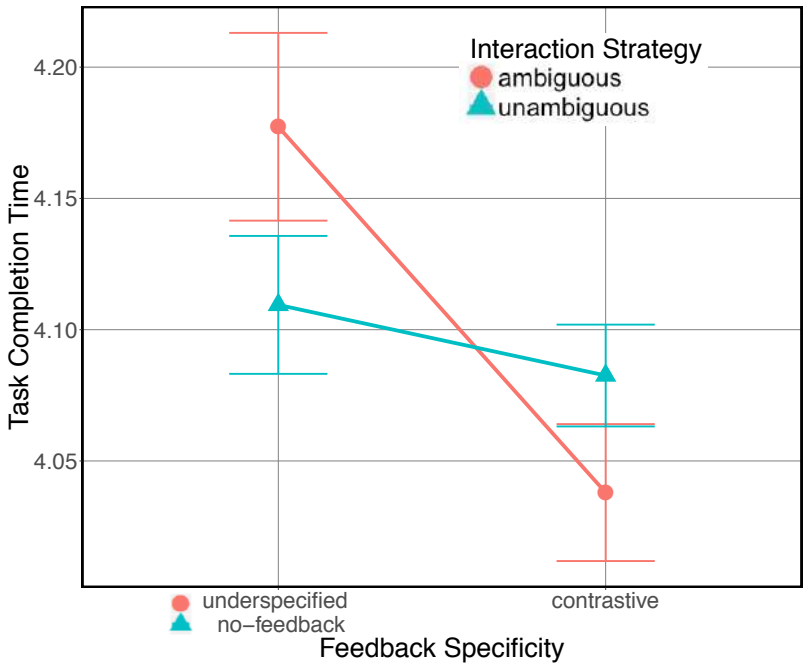

FIGURE 6.4: This plot depicts the task completion time (log transformed) from the instruction onset until the target is grasped in Experiment 1.

6.2.3 Results

The results reported in this section are based on 722 unique trials remaining after outlier removal (filtering out data points that were 2.5 standard deviations above or below the mean) from a total of 768; the outliers amounted to 6% of the data. Table 6.2 summarizes the number of trials for each condition and each group of participants. The number of correct trials indicates that it was unproblematic for the participants to identify a target in the UNAMBIGUOUS condition even if no feedback was provided by the system.

Performance The total time to solve each task, i.e. find and collect a building block, indicates efficiency of the communication with the system. All tasks were solved, and

Interaction Strategy	Group 1	Group 2
UNAMBIGUOUS	180 (166)	183 (176)
AMBIGUOUS	175 (151)	184 (166)

TABLE 6.2: This table summarizes the number of trials remaining after outlier removal and in how many of them no wrong objects were grasped (presented in brackets).

there were only a few wrong grasps (8.7%), and almost no need for repetition of an instruction, validating that both interaction strategies are effective. Figure 6.4 depicts the main findings: Participants were faster when they received CONTRASTIVE feedback after an UNAMBIGUOUS instruction as opposed to when no feedback was given after an UNAMBIGUOUS instruction (blue line). Additionally, we found that an AMBIGUOUS instruction in combination with CONTRASTIVE feedback was acted out faster compared to the combination with UNDERSPECIFIED feedback and, surprisingly, even outperforms the UNAMBIGUOUS interaction strategy (red line).

Specifically, the first group of participants was faster at solving the task listening to an UNAMBIGUOUS instruction ($M = 14.31\,sec, SD = 8.60\,sec$) than in the AMBIGUOUS condition with UNDERSPECIFIED feedback ($M = 17.56\,sec, SD = 10.44\,sec$). For the second group, the effect changed its direction: the AMBIGUOUS condition now led to shorter task completion times ($M = 11.96\,sec, SD = 5.61\,sec$) compared to the UNAMBIGUOUS one ($M = 12.75\,sec, SD = 4.75\,sec$).

More precisely, we constructed an individual model for each group with *InteractionStrategy* as a fixed effect and with random intercepts and slopes for subjects and items. Table 6.3 summarizes the inferential statistics. Both comparisons revealed main effects of *InteractionStrategy* for the first group exposed to underspecified feedback ($\chi^2(1) = 4.008, p < 0.05$) and for the second group exposed to contrastive feedback ($\chi^2(1) = 4.502, p < 0.05$).

For the AMBIGUOUS subset, we fitted a linear mixed-effects model with *FeedbackSpecificity* as fixed effect and included random intercepts and slopes for subjects and items. There was a main effect of *FeedbackSpecificity* on total time revealed by model comparison ($\chi^2(1) = 15.907, p < 0.001$), that is, CONTRASTIVE feedback improved task completion time over UNDERSPECIFIED feedback.

Listener gaze Next, we analyzed the identification time needed to find and inspect the intended target after instruction offset. Unsurprisingly, participants were quicker at identifying a target in the UNAMBIGUOUS instruction as it contains all object characteristics and also specifies its absolute position, so the search started while the instruction was being spoken (see Figure 6.5). Table 6.4 summarizes the mean reaction times per trial.

Analogously to the analysis of the total time, we fitted linear mixed-effects models for each dataset with the same random structure. Table F.1 summarizes the inferential statistics.

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	P($>\chi^2$)
model0	8	-116.36	-85.39	66.18	-132.36			
model1	9	-118.37	-83.52	68.19	-136.37	4.01	1	0.0453*

Group 1
Model 0: *totalTime ~ 1 + (InteractionStrategy | Subject) + (InteractionStrategy | Item)*
Model 1: *totalTime ~ InteractionStrategy + (InteractionStrategy | Subject) + (InteractionStrategy | Item)*

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	P($>\chi^2$)
model0	8	-305.54	-274.30	160.77	-321.54			
model1	9	-308.05	-272.90	163.02	-326.05	4.50	1	0.0338*

Group 2
Model 0: *totalTime ~ 1 + (InteractionStrategy | Subject) + (InteractionStrategy | Item)*
Model 1: *totalTime ~ InteractionStrategy + (InteractionStrategy | Subject) + (InteractionStrategy | Item)*

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
model0	8	-99.77	-68.70	57.88	-115.77			
model1	9	-113.67	-78.73	65.84	-131.67	15.91	1	<0.001***

Grpup 1 and Group 2 AMBIGUOUS condition
Model 0: *totalTime ~ 1 + (FeedbackSpecificity | Subject) + (FeedbackSpecificity | Item)*
Model 1: *totalTime ~ FeedbackSpecificity + (FeedbackSpecificity | Subject) + (FeedbackSpecificity | Item)*

TABLE 6.3: This table summarizes the models fitted to the performance data and the model comparison results for Experiment 1. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Model selection revealed main effects of *InteractionStrategy* for the first group ($\chi^2(1) = 60.257, p < 0.001$) and for the second group ($\chi^2(1) = 92.868, p < 0.001$). Additionally, a

		Instruction	**Identification Time**	**Grasp**	**Total Time**
first group	UNAMB.	7.21	2.17	4.93	14.31
	AMB.	2.81	7.22	7.52	17.56
second group	UNAMB.	7.23	1.27	4.25	12.75
	AMB.	2.80	4.21	4.94	11.96

TABLE 6.4: The mean durations in seconds of the interaction phases in Experiment 1 (see Figure 6.2).

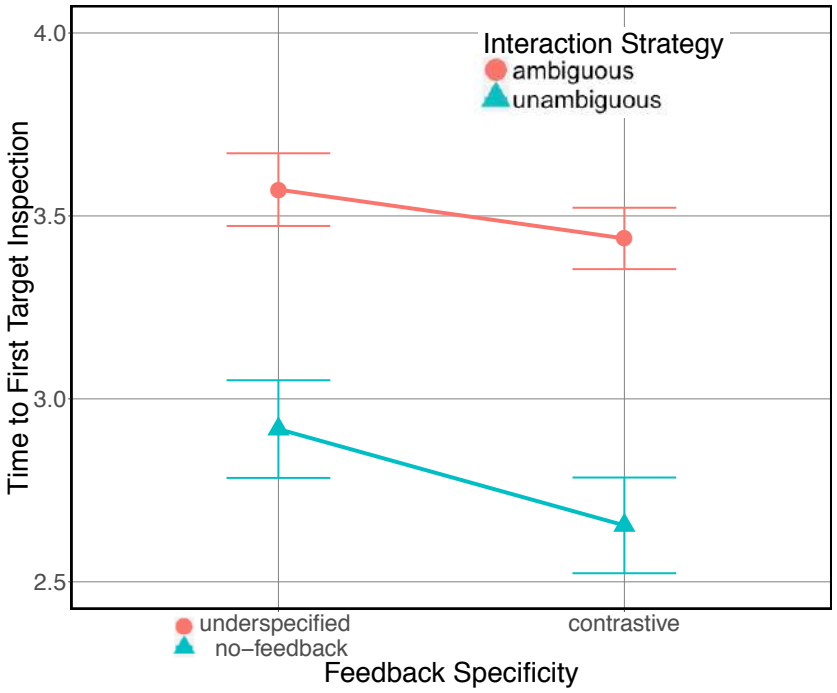

FIGURE 6.5: This plot depicts the time span from the instruction offset to the first target inspection in Experiment 1.

main effect of *FeedbackSpecificity* for the AMBIGUOUS condition ($\chi^2(1) = 4.172, p < 0.05$) was observed. In other words, listeners needed three times longer after hearing an AMBIGUOUS instruction ($M = 7.22\,sec, SD = 8.37\,sec$) to find the target object than after listening to an UNAMBIGUOUS one ($M = 2.17\,sec, SD = 5.12\,sec$). This time span was shortened dramatically when gaze-driven CONTRASTIVE feedback followed the instructions, though listeners still inspected the intended target sooner after the UNAMBIGUOUS interaction strategy ($M = 1.27\,sec, SD = 2.21\,sec$) than in the AMBIGUOUS case ($M = 4.21\,sec, SD = 3.80\,sec$).

Speech As we had full control of the speech modality, the only variation can be encountered in the feedback instances output by the system. For the UNAMBIGUOUS strategy, there is no comparison between groups because in the first part of the experiment, no feedback was given to the listener. We analyzed the number of negative feedback instances which occurred in the AMBIGUOUS condition across groups. To test if there was a significant difference, we constructed a generalized linear mixed-effects model (with a

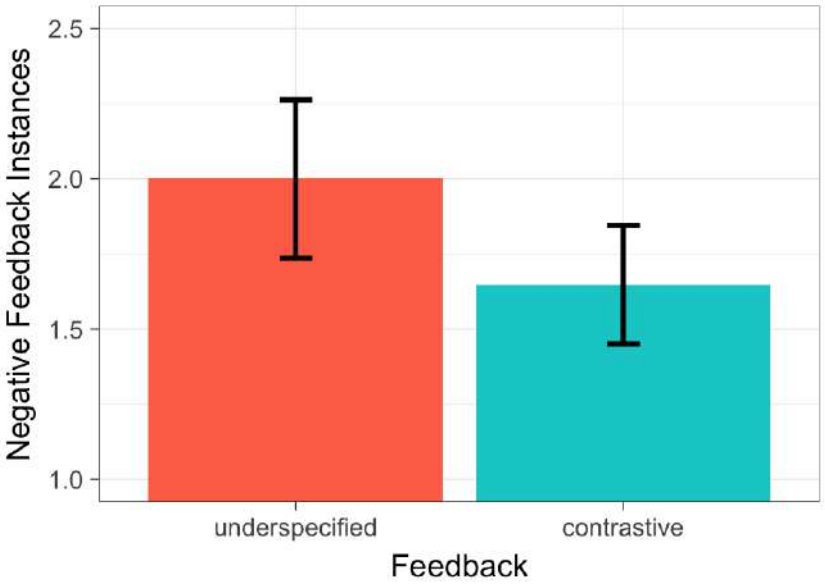

FIGURE 6.6: This plot depicts the number of negative feedback occurrences in Experiment 1.

logit link function) fitted to *FeedbackOccurrences* with *FeedbackSpecificity* as a fixed effect. Surprisingly, there was no significant difference with respect to our manipulation ($\beta = -0.038,\ SE = 0.086,\ z = -0.443,\ p = 0.658$). Overall there were more positive than negative instances in general ($\beta = -0.094,\ SE = 0.048,\ z = -1.948,\ p = 0.051$), which can be explained by the fact that whenever the listener is reaching for a target, she keeps looking at it and this triggers positive feedback affirming understanding. This pattern was also observed in our human-human interaction study (see Chapter 4), which suggests that feedback proportions are dependent on the setup (tabletop within reach). Surprisingly, there was no significant difference in the number of negative instances with respect to our manipulation ($\beta = -0.038,\ SE = 0.086,\ z = -0.443,\ p = 0.658$)

As the setting is very dynamic, the number of feedback occurrences might not be a good indicator of task performance and the involvement in the interaction. After carefully inspecting samples of the video material collected during the experiment, we observed that negative feedback instances can also occur after a confirmation of an object inspection (positive feedback instance) because listeners turned quickly to place the found building block on the LEGO model. Additionally, in such a setup an artifact is that no neutral

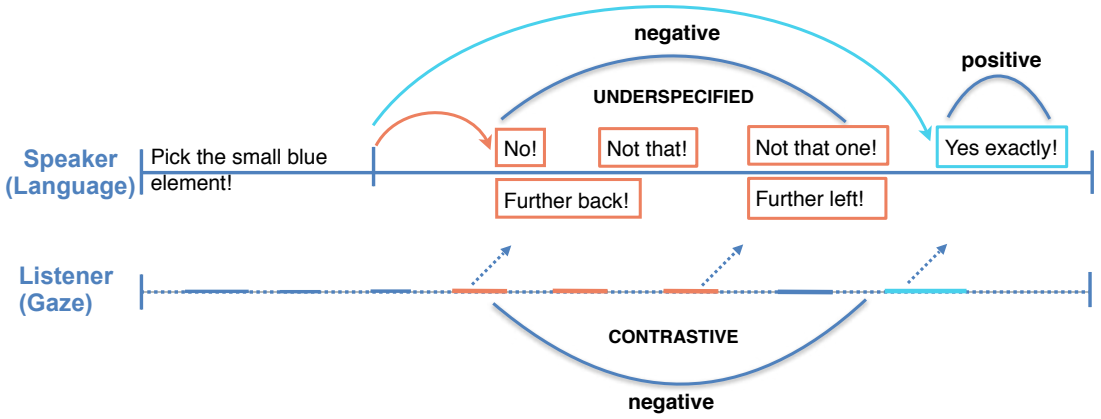

FIGURE 6.7: This figure illustrates how a typical trial looks and the differentiation of *FeedbackSpecificity*. The red arrows indicate the time intervals analyzed for the sequential feedback analysis.

fixation position exists, such as a fixation cross in the visual world paradigm. The projection of the gaze vectors hits an area of interest almost all of the time, and given the high density of similar objects, this can also trigger a reaction by the system. Additionally, there may be a larger participant variation in the eye movements which trigger the verbal feedback, and thus we analyzed the proportional feedback per trial, i.e. the number of negative feedback instances normalized by the total number of feedback instances that each participant triggered in each trial. There was no effect of *FeedbackSpecificity* ($\chi^2(1) = 1.179, p = 0.277$).

Further, we investigated the sequential order of feedback occurrences, i.e. how long after hearing an instruction listeners received the first negative and the first positive feedback instance, which are triggered by inspecting relevant objects (see red arrows in Figure 6.7). This mirrors visual search behavior during the task and also hints at how well participants engaged with the instruction-giving system.

Figure 6.8 depicts the mean time intervals from instruction offset to the onset of feedback instances for the AMBIGUOUS condition. By design of the interaction, positive feedback occurred later than negative feedback, which is reflected in a main effect of *FeedbackType* ($\chi^2(1) = 123.455, p < .001$). Importantly, the analysis showed that there is a main effect of *FeedbackSpecificity* ($\chi^2(1) = 18.416, p < 0.001$). As expected, the pattern observed in the listener gaze evaluation (time to first fixation) persists for the time to first positive feedback instance because this inspection triggers the first positive feedback

instance. In the UNDERSPECIFIED feedback condition, listeners induced later positive feedback ($M = 10.33sec, SD = 16.91sec$) than in the case of CONTRASTIVE feedback ($M = 5.43sec, SD = 5.97sec$). This demonstrates how more specific feedback narrowed down the search for the target object and shortened the time until finding it. Furthermore, the investigation of the first occurrence of a negative feedback instance revealed that listeners also inspected a competitor fitting the description faster in the CONTRASTIVE ($M = 1.97sec, SD = 2.68sec$) than the UNDERSPECIFIED ($M = 4.07sec, SD = 5.77sec$) condition. This suggests that listeners' expectation of an informative response elicits more deliberate and controlled use of gaze because the system constantly reacts with additional, useful information to their back channels.

Perception Participants answered 13 questions to judge each interaction strategy. 8 questions were using a five-point Likert scale (1 indicating a very good and 5 a poor score), e.g. "How good/precise did you find the spoken instructions?" or "How flexible did you find the interaction?". There were 5 yes/no questions like "Was the system's

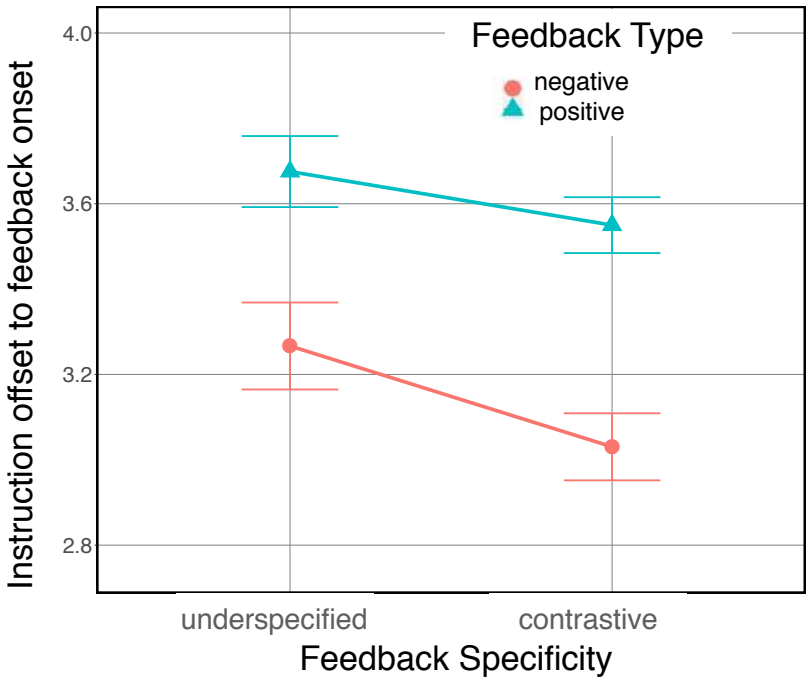

FIGURE 6.8: This plot depicts the time interval from the instruction offset to the onsets of the first negative (triggered by a competitor inspection) and first positive (triggered by a target inspection) feedback instances for the AMBIGUOUS *InteractionStrategy* in Experiment 1.

feedback confusing?" to assess if the interaction with the system felt natural or "Were the instructions exhaustive, i.e. you were able to identify a target upon hearing the instruction?" to check whether participants paid attention. In a final questionnaire they answered 5 yes/no questions to compare both interaction strategies and assess user preferences. Overall the interaction with the system was perceived as natural and gaze-driven feedback was rated as helpful and not confusing. In sum, we conclude that in terms of pace and flow, the interaction was well perceived. This can be interpreted as validation of our design and choice of parameters. In order to assess whether participants paid attention, they were asked if they noticed differences in the type of spoken instructions. In addition, we asked which one of the interaction strategies they preferred. Interestingly, there was a clear preference in both groups for listening and following an UNAMBIGUOUS instruction. All participants (100%) in the underspecified feedback group and most of the participants in the contrastive feedback group (87.5%) stated that they prefer UNAMBIGUOUS instructions and indicated them as more pleasant, although the contrastive feedback group was faster when experiencing the AMBIGUOUS strategy. However, a simple linear regression ran on the responses to "How good did you find the interaction flow?"

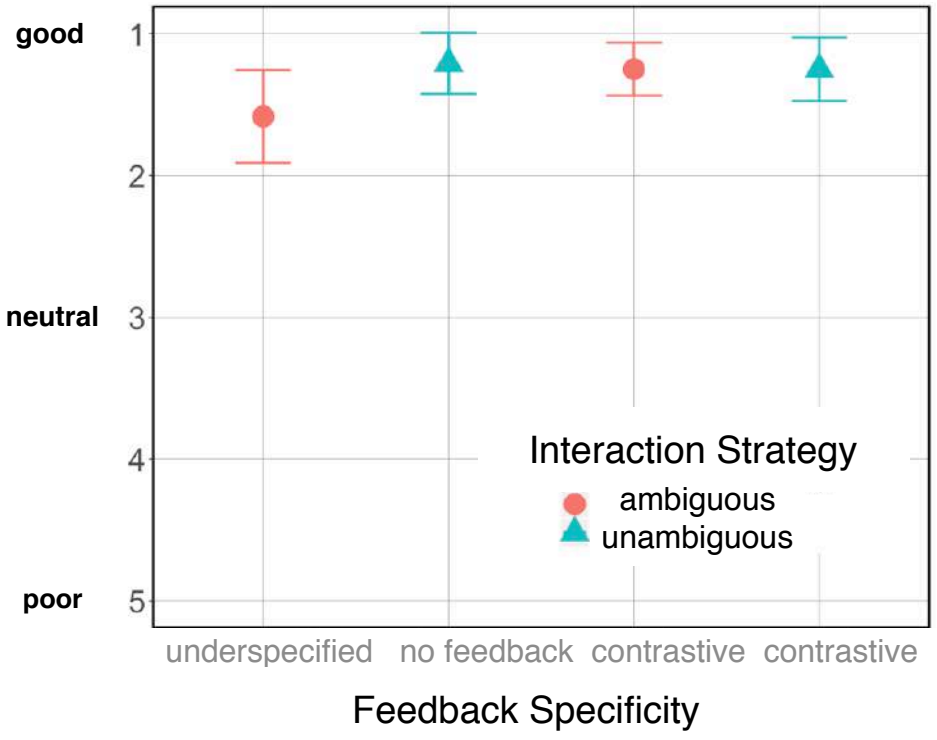

FIGURE 6.9: This plot depicts participants' perception and judgement of the interaction flow measured on a Likert scale for Experiment 1.

revealed main effects of *FeedbackSpecificity* ($\beta = -0.333,\ t(92) = -1.996,\ p < 0.05$) and *InteractionStrategy* ($\beta = -0.750,\ t(92) = -2.008,\ p < 0.05$) (see 6.9). That is, when CONTRASTIVE feedback followed an AMBIGUOUS instruction, it was judged to be better ($M = 1.25, SD = 0.44$) than when UNDERSPECIFIED feedback was provided ($M = 1.58, SD = 0.77$). The former assessment was similar to the perception of the UNAMBIGUOUS instructions by the contrastive ($M = 1.25, SD = 0.53$) and underspecified feedback group ($M = 1.20, SD = 0.51$). This demonstrates that the informativity of the verbal feedback improves the *InteractionStrategy* giving initially partial, AMBIGUOUS instructions and so listeners experience it as smoother.

6.2.4 Discussion

The analyses of the data collected during interactions with the GazInG system provide some evidence for successful use of listener gaze in a real-world task. An *InteractionStrategy* that refers to objects incrementally and reacts to listeners' gaze can be used to identify objects in the shared space. Moreover, performance results indicate that *FeedbackSpecificity* is essential for efficiency. The results reveal that CONTRASTIVE feedback benefits task performance because it not only warns the listener against grasping a wrong object, but also includes a relative direction in which to look for the target. In contrast, UNDERSPECIFIED feedback solely prevents the user from wrong grasps and does not facilitate search. Notably, the combination of AMBIGUOUS instructions with CONTRASTIVE feedback even outperformed following UNAMBIGUOUS instructions, which contain all characteristics including the position of the target object.

Interestingly, there was a mismatch in the perception and performance measures with respect to the UNAMBIGUOUS and AMBIGUOUS interaction strategies. Apparently, listeners felt more confident in their own performance when following UNAMBIGUOUS instructions. One explanation for this perception might be that the UNAMBIGUOUS strategy allowed participants to be more passive during the interactions. After an AMBIGUOUS instruction, in contrast, they had to actively engage with the system in order to make progress in the task. The former is obviously considered as more convenient despite being apparently less efficient compared to the more interactive strategy with specific responses.

In sum, listener gaze is important and can be used to split information by proving first a partial description and then supplementary, more informative feedback as a reaction to

object inspections. Task performance depends on the informativity of system responses, i.e. CONTRASTIVE feedback leads to shorter task completion time.

6.3 Experiment 2: Interaction with the NLG System *"Feedback"*

This experiment was intended to further examine the impact of *FeedbackSpecificity* on task performance, still giving AMBIGUOUS instructions but with a different distribution than in the previous experiment. Here, *FeedbackSpecificity* was manipulated within participants and UNDERSPECIFIED vs. CONTRASTIVE feedback occurred in an interleaved and randomized order on an item-by-item basis. Thus, participants did not know in advance which type of feedback they might receive and a strategic adaptation to the specific system behavior was impossible. This aimed at assessing whether participants benefited from the CONTRASTIVE feedback in the first experiment, because more information was indeed conveyed in the form of CONTRASTIVE feedback, so that this system is *inherently* more efficient—or whether participants more generally adapted to the system, e.g., by increasing their attentiveness or willingness to collaborate and thus to really take up and process the provided information efficiently. If the former hypothesis holds, then performance in the CONTRASTIVE feedback condition would remain high (and higher than with UNDERSPECIFIED feedback), even when it occurred in an interleaved manner. If the latter hypothesis is true, we would expect to see either low performance in both conditions (when engagement decreases altogether) or high performance in both conditions (when engagement is high and leads to more efficient information uptake).

6.3.1 Participants

Twenty-four German native speakers participated in the experiment. The average age of participants was 24 years with a range of 18–32. They reported normal or corrected-to-normal vision and no red-green color blindness, and were compensated with €7.

6.3.2 Procedure

The task was the same as in Experiment 1 and the procedure was almost identical. This time, the experimental part consisted of four blocks and so two more layouts were designed (see Appendix B). In contrast to the procedure in Experiment 1, there was no questionnaire, but after finishing all four blocks participants answered two questions: whether they noticed any differences and if they had a particular strategy for inspecting objects. The experiment lasted around 40 minutes.

6.3.3 Results

Performance Figure 6.10 depicts the time needed to finish the task given AMBIGUOUS instructions for both experiments. In contrast to Experiment 1 (left plot) there is no significant difference in performance observed in Experiment 2 (right plot). When participants received UNDERSPECIFIED feedback, task completion time was slightly longer ($M = 12.63\,sec, SD = 6.83\,sec$) than following CONTRASTIVE feedback ($M = 12.33\,sec, SD = 6.52\,sec$). We fitted a model with *FeedbackSpecificity* as a fixed effect and with random intercepts and slopes for subjects and items, but there was no significant effect of *FeedbackSpecificity* ($\chi^2(1) = 0.666, p = 0.414$). There were twice as many targets in Experiment 2 as in Experiment 1 and thus we split the interactions in two parts (first vs. second

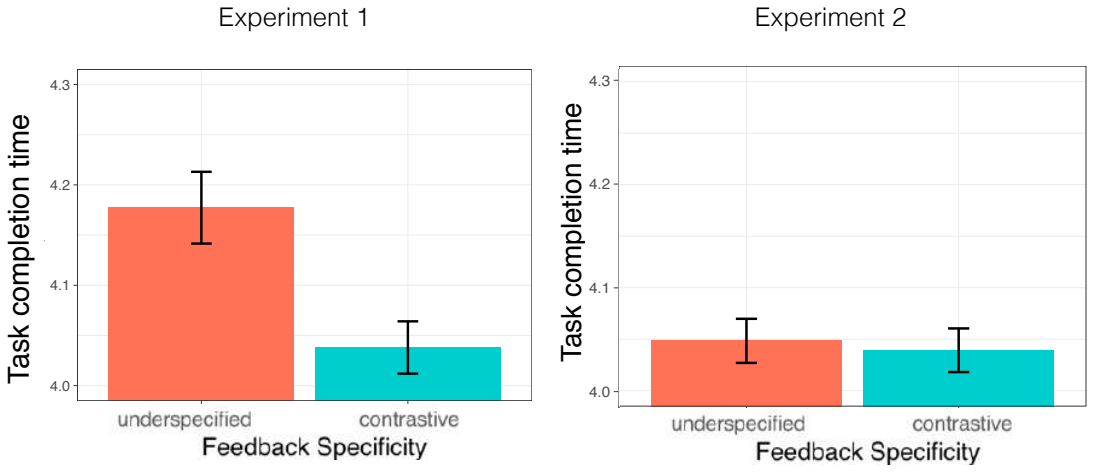

FIGURE 6.10: The task completion time measured in interactions obtained in Experiment 1 (left plot) and in Experiment 2 (right plot).

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
Model 0	12	-346.84	-291.92	185.42	-370.84			
Model 1	13	-345.51	-286.01	185.75	-371.51	0.67	1	0.4144
Model 2	14	-348.87	-284.79	188.43	-376.87	5.36	1	0.0206*
Model 3	15	-348.37	-279.72	189.18	-378.37	1.50	1	0.2200

Model 0: *totalTime* $\sim$ *1 + (FeedbackSpecificity * Half | Subject)*
Model 1: *totalTime* $\sim$ *FeedbackSpecificity + (FeedbackSpecificity * Half | Subject)*
Model 2: *totalTime* $\sim$ *FeedbackSpecificity + Half + (FeedbackSpecificity * Half | Subject)*
Model 3: *totalTime* $\sim$ *FeedbackSpecificity * Half + (FeedbackSpecificity * Half | Subject)*

TABLE 6.5: This table summarizes the models fitted to the performance data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

half). More precisely, we compared the performance in the first two layouts vs. the last two layouts participants completed. Table 6.5 summarizes the model specifications and results from model selection.

The analysis of the first vs. second half of the experiment revealed a main effect on task performance ($\chi^2(1) = 5.359, p < 0.01$). That means that listeners improved over time and worked better with the system, which resulted in faster task completion in the second half for both conditions (CONTRASTIVE ($M = 11.35\,sec, SD = 56.08\,sec$) vs. UNDERSPECIFIED ($M = 12.159\,sec, SD = 6.336\,sec$) than in the first half (CONTRASTIVE ($M = 13.34\,sec, SD = 72.16\,sec$) vs. UNDERSPECIFIED ($M = 13.13\,sec, SD = 7.30\,sec$)).

Listener gaze The gaze signal represents visual search and we analyzed the identification time span from instruction end until listeners inspect the target object for the first time. We fitted a model with *FeedbackSpecificity* as fixed effect and similarly to the performance results, the identification time did not reveal a significant difference ($\chi^2(1) = 0.0648, p = 0.799$) for Experiment 2, whereas this was the case in Experiment 1 (see Figure 6.11 for visualization).

In the second experiment, this point in time was a bit later given UNDERSPECIFIED ($M = 4.84\,sec, SD = 4.82\,sec$) than given CONTRASTIVE feedback ($M = 4.47\,sec, SD = 5.32\,sec$). The analysis of both experimental parts also did not reveal a significant effect.

Table 6.6 summarizes the mean response times for Experiment 2.

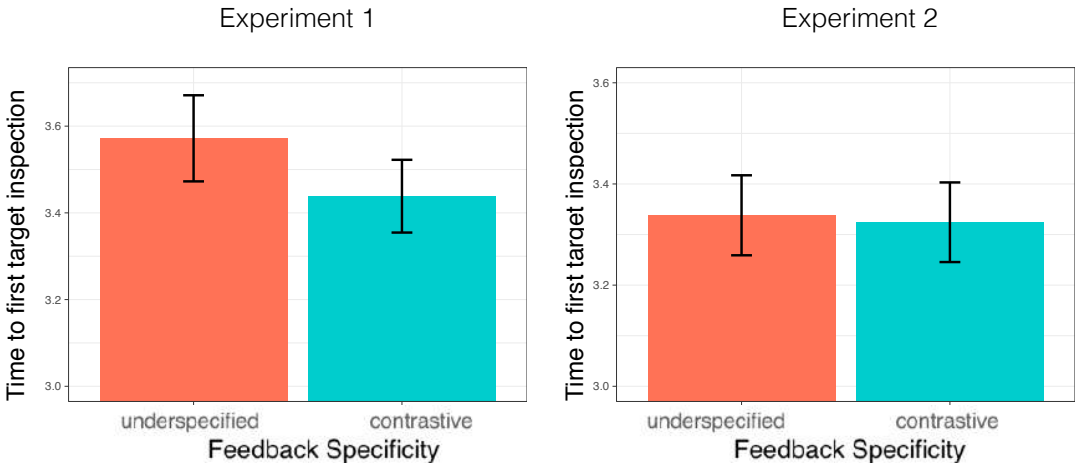

FIGURE 6.11: This plot depicts the time span from the instruction offset to the first target inspection in Experiment 1 (left plot) and in Experiment 2 (right plot).

Speech The type of instructions (*InteractionStrategy*) was not manipulated in this experiment, i.e. the system systematically generated AMBIGUOUS instructions. However, verbal feedback can be considered as a dependent variable since it is a direct consequence of participants' visual search behavior: Competitor inspections triggered negative feedback and target inspections triggered positive feedback. The negative feedback differed in specificity. There was a significant effect of *FeedbackSpecificity* ($\chi^2(1) = 5.169, p < 0.01$) on the number of feedback occurrences. Table 6.7 summarizes inferential statistics. That is, when listeners followed UNDERSPECIFIED feedback ($M = 2.19$ *inst* $SD = 1.56$ *inst*) their gaze triggered more negative instances, i.e. they considered more competitors before arriving at the target, in comparison to when they followed CONTRASTIVE feedback ($M = 1.74$ *inst*, $SD = 1.10$ *inst*). The analysis of the proportional feedback revealed a

FeedbackSpecificity	**Instruction**	**Identification**	**Grasp**	**Total Time**
UNDERSPECIFIED	2.81	4.84	4.98	12.63
CONTRASTIVE	2.79	4.47	5.07	12.33

TABLE 6.6: Mean durations in seconds of the three interaction phases and the total time for Experiment 2 as depicted in Fig. 6.2.

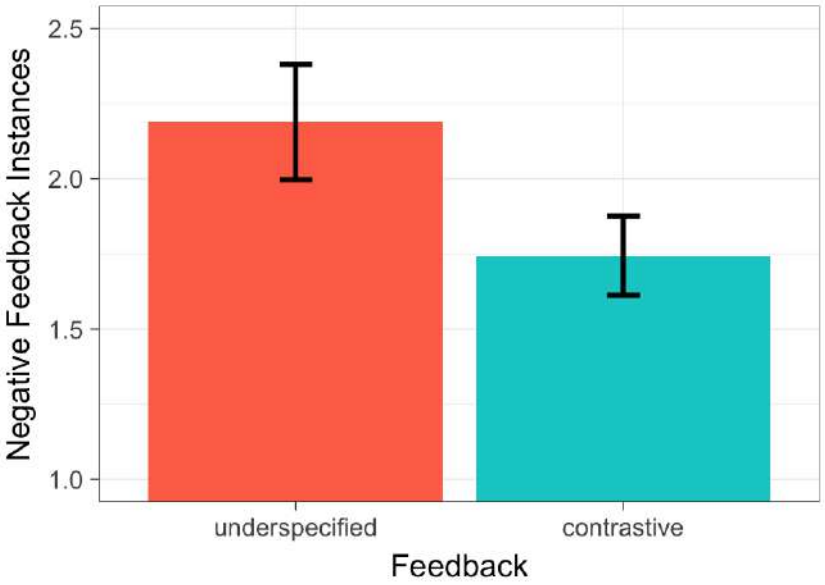

FIGURE 6.12: This plot depicts the number of negative feedback occurrences in Experiment 2.

marginal difference ($\chi^2(1) = 3.383, p = 0.065$).

Furthermore, we ran a sequential analysis on feedback occurrences to assess first relevant inspections. Typically, positive feedback occurred after negative feedback, revealed by a main effect ($\chi^2(1) = 373.146, p < 0.001$), but there is no significant difference with respect to our manipulation of *FeedbackSpecificity* ($\chi^2(1) = 0.100, p = 0.752$)

	Estimate	Std. Error	Wald z	p
(Intercept)	0.24	0.05	5.15	$< 0.001^{***}$
contrastiveFeedbackunderspec	0.17	0.07	2.40	0.02^{*}

Model: *negativeFeedbackInstances* $\sim$ *FeedbackSpecificity + (FeedbackSpecificity | Subject)*

TABLE 6.7: This table summarizes the models fitted to the listener gaze data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

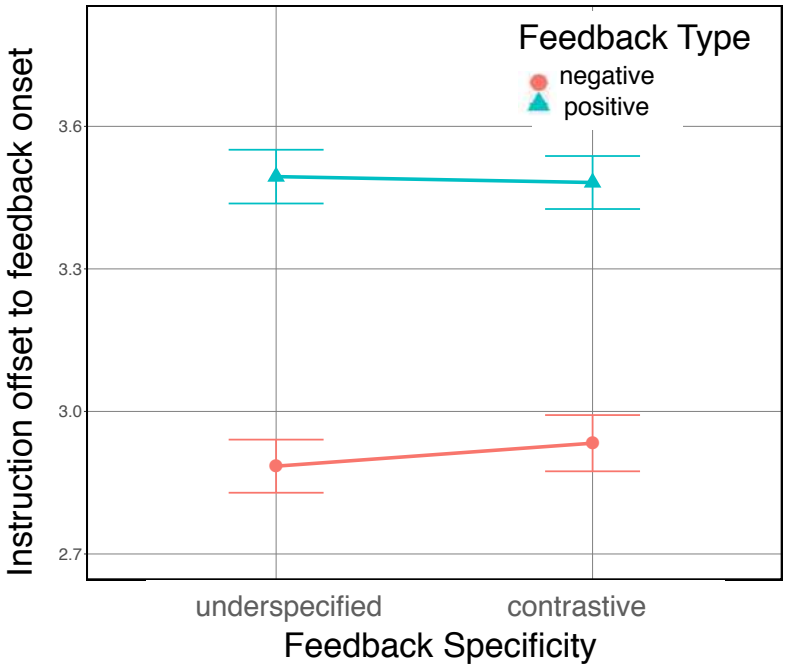

FIGURE 6.13: This plot depicts the time interval from the instruction offset to the onsets of the first negative and first positive feedback instances in Experiment 2.

(see Figure 6.13). Notably, although UNDERSPECIFIED feedback did not provide additional information, after hearing an instruction listeners quickly inspected a competitor object matching the description and triggered the first negative feedback instance ($M = 1.45\,sec, SD = 2.19\,sec$) and this happened similarly soon after getting CONTRASTIVE feedback ($M = 1.79\,sec, SD = 3.34\,sec$). There was no significant difference between the two conditions.

6.3.4 Discussion

The results of Experiment 2 suggest that the presence of more informative feedback (even when occurring only occasionally) influences the overall willingness to interact and cooperate with a system on solving a task together. Listeners seem to have greater expectations for the capabilities of the `GazInG` system, which is mirrored in their gaze behavior. Thus participants in the more difficult and rather unnatural condition (AMBIGUOUS instructions with UNDERSPECIFIED feedback) were now as efficient as those experiencing the more specific one (AMBIGUOUS instructions with CONTRASTIVE feedback), in contrast to

the results obtained in Experiment 1. These findings provide some evidence that participants are able to deal effectively with imperfect behavior of a system as long as they perceive it as helpful and efficient overall. In other words, it is not just the actual informativeness of the spoken output in a trial, but the confidence in the system's supportive behavior more generally, that determines how efficient information uptake is. In terms of the strategic use of gaze, it seems as if none of the participants spotted the manipulation. Thus, we assume that listeners adapted their engagement and behavior rather naturally and unconsciously instead of employing a tactic for where to look and which object to inspect next in a specific experimental condition.

It seems that splitting the description into subsequent chunks and providing these incrementally is beneficial and improves task performance. However, this interaction strategy does not exactly generate true installments, because the second piece of information is in the form of feedback, which is related to the current gaze position. Thus we developed a second NLG system to generate true installments and output them depending on the listener's gaze behavior, or concatenate all of them to output an exhaustive instruction.

Chapter 7

Human-Machine Interaction: Effects of Gaze-driven Installments and Information Order on Performance

Usually an NLG system plans and outputs a reference in a single noun phsase (Stoia et al., 2006; Garoufi & Koller, 2010). In highly interactive settings, however, it is common that speakers often start speaking before they have planned the entire utterance and so provide the information incrementally by presenting it not at all once, but in subsequent chunks to the listener, i.e. they refer to objects in installments (Striegnitz et al., 2012). Hence, speakers are better able to adapt to changes in the surroundings and to the listener's signals. Recently, Zarrieß and Schlangen (2016) applied installments to generate referring expressions and demonstrated that such a generation approach enhances identification of real objects depicted in static images. What remains unclear, and what we address in this chapter, is (1) whether installments can be applied for dynamic goal-oriented tasks and (2) whether the listener gaze can be utilized to trigger and inform such installments. Previous work by Fang et al. (2015) integrated gaze in a collaborative referring expression generation (REG) algorithm. They observed a performance drop when using gaze, but their work focused more on embodiment and robots gestures.

In this chapter, we present an experiment that was designed to investigate the effectiveness of referring to co-present objects in gaze-driven installments and the role of information order. Our findings presented in the previous chapter indicate that using listener gaze to

deliver a referring expression incrementally is an effective interaction strategy for an interactive instruction-giving system. Furthermore, the informativity of gaze-driven feedback determines efficiency, namely, participants solved the task faster when they received CONTRASTIVE feedback than when they received UNDERSPECIFIED feedback. The former even outperformed following an UNAMBIGUOUS description that gives all of the information at once. However, this is not quite a fair comparison because an UNAMBIGUOUS instruction contains more information not mentioned in the AMBIGUOUS variant, and the contrastive feedback directs listeners' attention to the target by specifying the position relative to the current gaze. Thus, we consider 'real' gaze-based installments in the following.

7.1 Experiment 3: Interaction with the NLG System *"Installments"*

In this experiment, we shed light on the comparison of referring in INSTALLMENTS (triggered by a listener's object inspections) vs. NOINSTALLMENTS. For this we implemented and used the NLG system "Installments" described in Section 5.2.3. Additionally, since the location of the target object (specified by a spatial descriptor) helps to resolve a referring expression, we investigate whether swapping the partial feature description of a target object, and the expression that specifies the location of the target object would be even more beneficial for the listener's understanding, that is, *SpatialDescriptor* is generated FIRST vs. SECOND, which occurred in an interleaved and randomized order on an item-by-item basis.

Here, the gaze signal is rather inconspicuously integrated in the incremental generation mechanism because the system provides an absolute viewer-centered and not a relative spatial expression as direct feedback to eye movements. Additionally, a fallback strategy if the target is still not identified after the first two installments is to deliver a third installment that specifies the remaining features of the searched-for object. We hypothesize that listeners will benefit from hearing the location specification FIRST because this information restricts the search space and with the next installment (feature description) the listener will more quickly (efficiently) identify the intended target (see example (2) in Section 5.2.3) as opposed to when it is the other way around (see example (3) in Section

5.2.3). If the interactions are equally efficient in both conditions, then this would indicate that the order of the information pieces does not play a role.

For the NoInstallments condition there are two corresponding versions with respect to the position of the *SpatialDescriptor* in the instruction (see examples (4) and (5) in Section 5.2.3).

If monitoring gaze to refer in installments is generally more suitable for such interactions, then following Installments would lead to better performance than following NoInstallments). If we obtain an effect in the opposite direction, this would mean that the form of the installments plays a crucial role, that is, relative vs. absolute spatial information determines efficiency. For the latter it could be further argued that more direct involvement of listener gaze increases listeners' attentiveness and willingness to collaborate with the instruction-giving system. If there is no difference, then again gaze-driven relative direction is essential to make an installment effective. Regarding visual search we expect to observe earlier target inspections when *SpatialDescriptor* is mentioned first than when it is mentioned second.

7.1.1 Method

In our third NLG experiment, we investigated the interaction with the NLG system *"Installments"* and whether monitoring listener gaze benefits performance when it is integrated in the generation method differently, namely to trigger the next information piece. Specifically, we examined if and how the *InformationDelivery* approach (NoInstallments vs. Installments) and *SpatialDescriptor* occurrence (first vs. second) affect the interaction with our system. We used the same setup and apparatus (see Section 6.1.1) but we upgraded to a successor model of the SMI Eye Tracking Glasses recording at 120 Hz.

The measures and analysis were almost the same as in the previous two experiments (see Section 6.1.2). Figure 6.2 depicts the interaction phases, which vary with respect to the *InformationDelivery* approach employed by the instruction-giving system. The only difference is that the identification time in Experiment 3 was measured from the instruction onset, i.e. the beginning of the first installment.

	List 1		List 2	
	layout	*InfoDelivery*	*layout*	*InfoDelivery*
Block 1	$\frac{1}{2}$	NoInstallments	$\frac{1}{2}$	Installments
Block 2	$\frac{3}{4}$	Installments	$\frac{3}{4}$	NoInstallments

Table 7.1: The design of Experiment 3.

Participants

Twenty-four German native speakers (18 female) took part in the experiment. The average age of the participants was 25 years with a range of 19–34. They reported normal or corrected-to-normal vision and no red-green color blindness, and their participation was compensated with €8.

Procedure

The task was the same as in Experiments 1 and 2 (see Chapter 6). The procedure was analogous to the one in Experiment 1 and included the two additional layouts tested in Experiment 2. The experimental part consisted of two blocks, one for each interaction

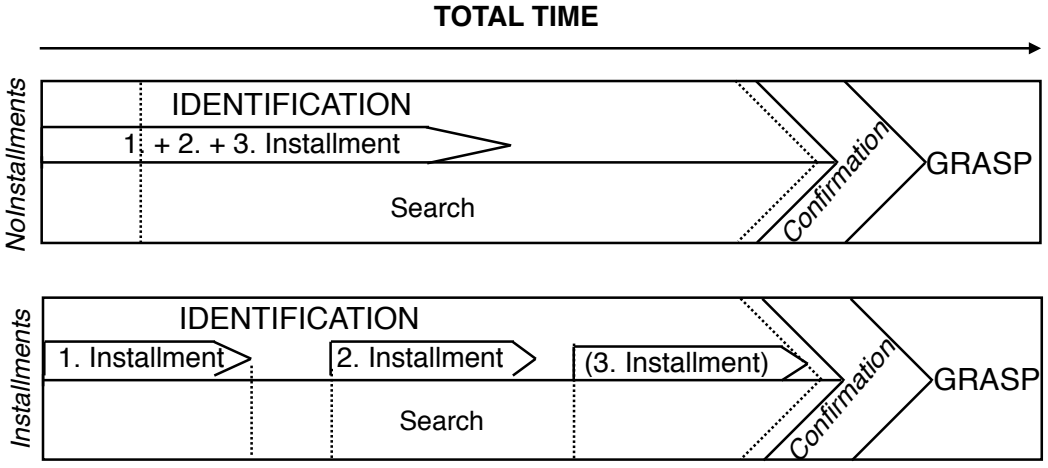

Figure 7.1: This diagram illustrates the interaction phases for both information delivery approaches NoInstallments and Installments

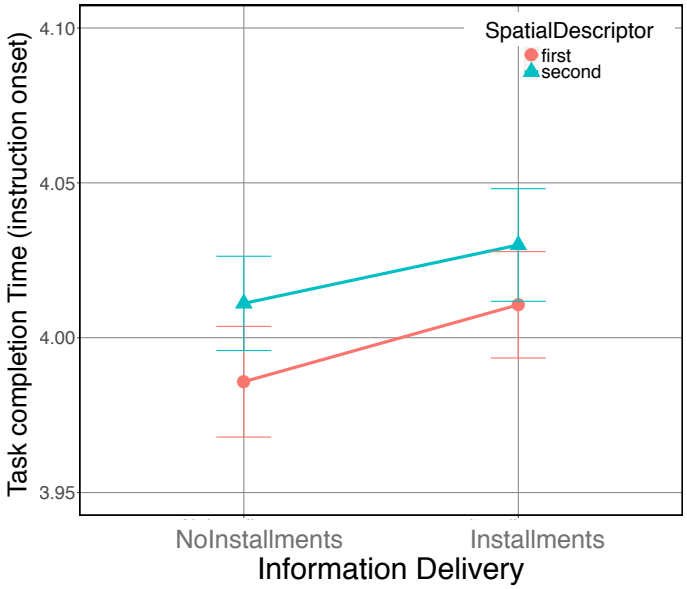

FIGURE 7.2: This plot depicts the task completion time in Experiment 3.

strategy and each block consisted of two layouts (see Table 7.1). Participants filled in a questionnaire after each block that assessed their perception of each information delivery approach, and at the end they answered questions about the comparison of the two. The experiment lasted around 45 minutes.

7.1.2 Results

Performance The overall performance measure is the total time needed to complete the task. In Figure 7.2 the mean task completion time in each condition is depicted: Participants completed the task faster when they followed NOINSTALLMENTS than when they were receiving the information incrementally in INSTALLMENTS. Furthermore, mentioning the *SpatialDescriptor* FIRST led to more efficient interactions compared to when the *SpatialDescriptor* appeared SECOND after a partial feature description, independent of the *InformationDelivery* approach. Interestingly, the performance in NOINSTALLMENTS with *SpatialDescriptor* SECOND and INSTALLMENTS with *SpatialDescriptor* FIRST is equally efficient and this validates the effectiveness of the piece-wise information delivery approach. Specifically, we fitted a linear mixed-effects model with random intercepts and random slope for *subject* to the dataset consisting of 713 trials in total.

Table 7.2 summarizes the models and results from the inferential analysis. Model comparison indicated that the *InformationDelivery* had a significant effect on task completion time ($\chi^2(1) = 4.63, p < 0.05$). We also found a significant main effect of *SpatialDescriptor* ($\chi^2(1) = 7.68, p < 0.01$). Listeners achieved the best performance when following NoInstallments and when the *SpatialDescriptor* was specified first ($M = 10.12\,sec, SD = 3.61\,sec$). Finding and collecting a specific object among several others took more time in the case of following NoInstallments with *SpatialDescriptor* second ($M = 10.59\,sec, SD = 3.08\,sec$) and similarly long in the case of Installments with *SpatialDescriptor* first ($M = 10.65\,sec, SD = 3.37\,sec$). The slowest task completion time was observed in Installments with *SpatialDescriptor* second ($M = 11.18\,sec, SD = 3.63\,sec$).

Listener gaze We analyzed visual search considering the listener's gaze signal and particularly the identification time, that is, the interval needed to inspect the intended target after instruction onset. The results show that listeners focused on the correct object earlier when its location was specified at the beginning of the instruction, which supports the hypothesis that the search space restriction is beneficial and this helps the instruction follower to speed up the search. This time span was longer when the *SpatialDescriptor* was uttered after the feature description, as visualized in Figure 7.3.

	Df	AIC	BIC	logLik	Deviance	χ^2	Chi Df	Pr(>Chisq)
Model0	12	-1122.69	-1067.86	573.34	-1146.69			
Model1	13	-1128.37	-1068.97	577.19	-1154.37	7.68	1	0.0056**
Model2	14	-1131.00	-1067.02	579.50	-1159.00	4.63	1	0.0315*
Model3	15	-1129.00	-1060.45	579.50	-1159.00	0.00	1	0.9990

Random Structure: *(InformationDelivery*SpatialDescriptor | Subject)*
Model 0: *totalTime* $\sim$ *1 + Random Structure*
Model 1: *totalTime* $\sim$ *SpatialDescriptor + Random Structure*
Model 2: *totalTime* $\sim$ *SpatialDescriptor+InformationDelivery + Random Structure*
Model 3: *totalTime* $\sim$ *SpatialDescriptor*InformationDelivery + Random Structure*

Table 7.2: This table summarizes the models fitted to the performance data and the model comparison results for Experiment 3. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

As was done for the performance measure, we fitted linear mixed-effects models and conducted model selection to assess statistical significance. Table F.5 summarizes the models and results from model selection. The analysis revealed main effects for both factors: *InformationDelivery* ($\chi^2(1) = 14.50, p < 0.001$) and *SpatialDescriptor* ($\chi^2(1) = 18.72, p < 0.001$). Specifically, listeners inspected the target sooner when they heard the *SpatialDescriptor* FIRST ($M = 5.56\,sec, SD = 1.91\,sec$) in NOINSTALLMENTS and ($M = 6.15\,sec, SD = 2.12\,sec$) in INSTALLMENTS . This time interval was longer when the *SpatialDescriptor* appeared in the SECOND position in NOINSTALLMENTS ($M = 6.16\,sec, SD = 1.65\,sec$) and even longer in INSTALLMENTS ($M = 7.09\,sec, SD = 2.53\,sec$).

Speech The *InformationDelivery* approach employed by the system to refer to specific objects was predefined and the NOINSTALLMENTS case did not allow for variation in the language modality (see Section 5.2.3). In the interactive version of the system, however, the number of installments that the system generated while instructing the human listener could differ. We evaluated this dependent variable by constructing a generalized linear mixed-effects model (with a logit link function) and observed a main effect of *SpatialDescriptor* ($\beta = -0.21452$, $SE = 0.0745$, $z = -2.881$, $p < .001$). There

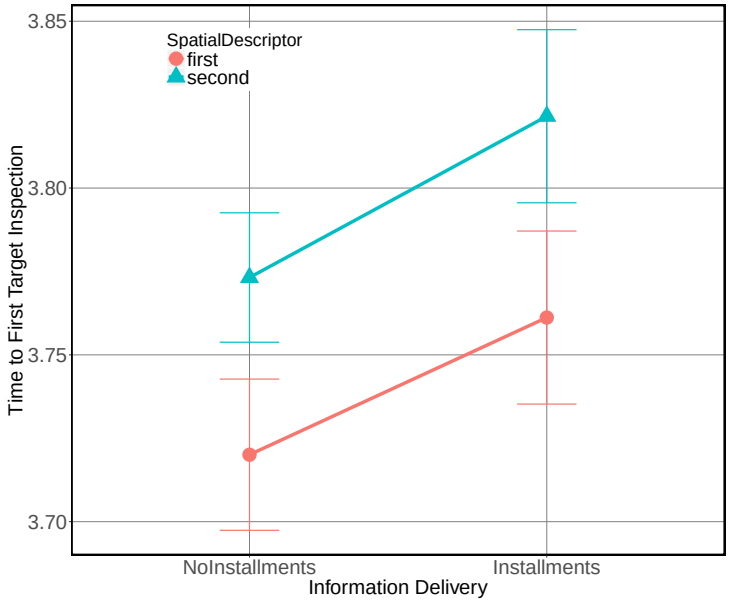

FIGURE 7.3: This plot depicts the time interval from instruction onset to first target inspection in Experiment 3.

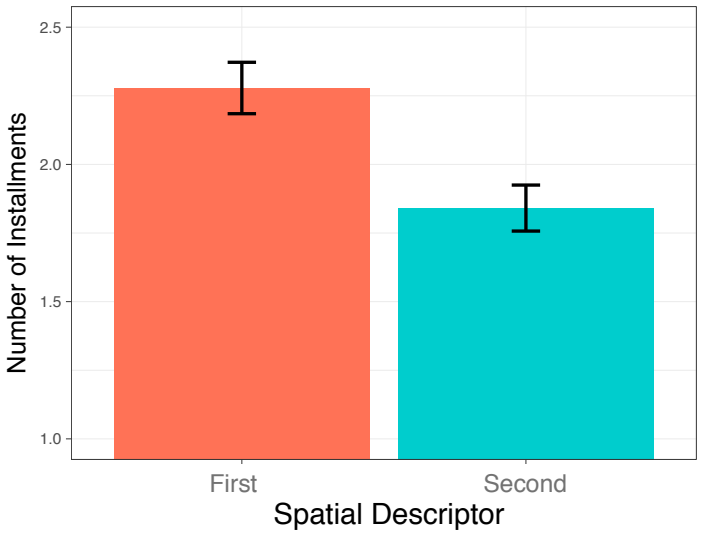

FIGURE 7.4: This plot depicts the time interval from instruction onset to first target inspection in Experiment 3. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

were more installments generated by the system when the *SpatialDescriptor* was mentioned FIRST ($M = 2.28inst, SD = 0.63inst$) than when it was mentioned SECOND ($M = 1.84inst, SD = 0.56inst$).

*SpatialDescriptor*s are very much related to referring expressions, but the main difference is that they can specify a location even when there is nothing there. This property may be the reason why participants hesitated to grasp a target when they received only this information. That is, even though *SpatialDescriptor* FIRST leads to earlier inspections on target objects, more installments are elicited before the participant finally grasps the object (compared to *SpatialDescriptor* SECOND). For example, if only the *SpatialDescriptor* and confirming feedback triggered by a target inspection is output by the system (e.g. "Pick the following building block! At the back toward the left ... < *target inspection* > ... Yes!") participants are likely to consider other objects. This means that in order to be confident enough to initiate an action listeners need to hear a feature description of the target object and thus when the *SpatialDescriptor* was planned to appear in the SECOND position, they would tend to grasp after hearing the first information bit and a confirmation of an inspection ("Pick the following building block! The big blue one ... < *target inspection* > ... Yes!"). During carefully inspecting of the video material of some trials in INSTALLMENTS FIRST condition, we found out that when the system mentioned the color of the

bottom building block, they inspected nearby competitors with the same color (which is an absolute feature), presumably to determine the meaning of the size modifier (which is a relative feature) in the current visual context.

Perception As was done for Experiment 1, we assessed the perception of the users with post-task questionnaires. Participants answered 7 questions to judge each information delivery approach. The questionnaires consisted of 4 questions using a five-point Likert scale (1 indicating a very good and 5 a poor score), e.g. "How good/precise did you find the spoken instructions?" or "How flexible did you find the interaction?". There were also 3 yes/no questions like "Was it clear at all time points during the interaction what you were supposed to do?" to assess if the interaction with the system felt natural or "Were the instructions exhaustive, i.e. were you able to identify a target upon hearing the instruction?" to check whether participants paid attention to the form of the instructions. In a final questionnaire, they answered 4 yes/no questions to assess which information delivery approach they preferred. Overall, the interaction with the system was well perceived and piece-wise information delivery was not rated as distracting. In order to assess whether

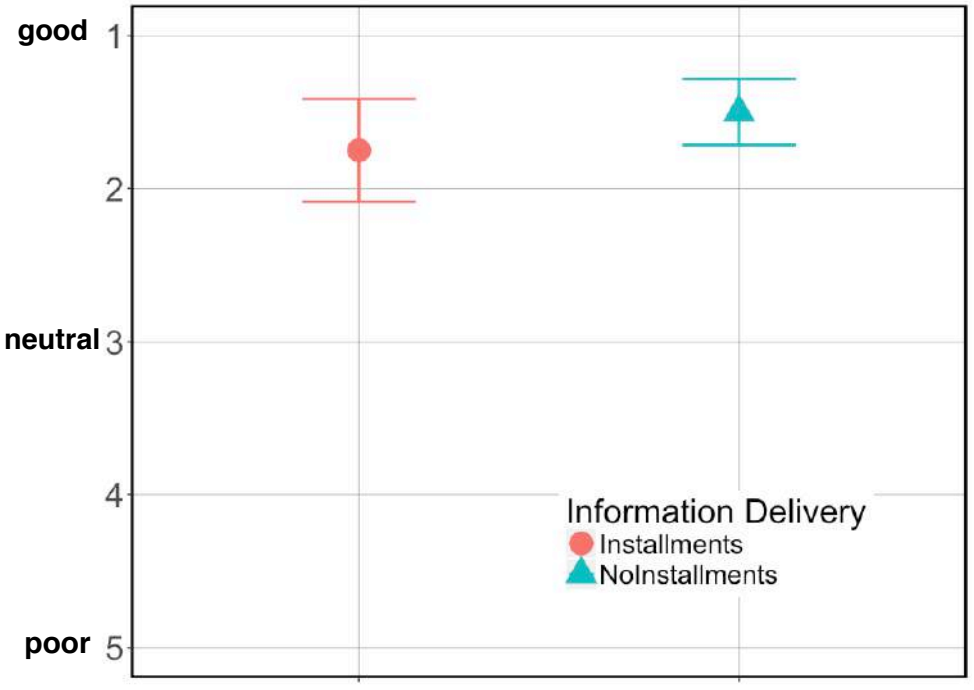

FIGURE 7.5: This plot depicts participants' perception and judgment of the interaction flow measured on a Likert scale for Experiment 3.

participants paid attention, they were asked if they noticed differences in the type of spoken instructions. In addition, we asked which one of the information delivery approaches they preferred. Most of the participants preferred to follow NoInstallments (70.8%) when they need to identify a specific building block. Figure 7.5 depicts the mean values and we analyzed the responses to the question "How good did you find the interaction flow?". Specifically, we ran a simple linear regression, but there was no effect of *InformationDelivery* ($\beta = -0.250,\ t(46) = -1.297,\ p = 0.201$). This means that the interaction flow was similarly good, as the scores we obtained were very high: NoInstallments ($M = 1.50, SD = 0.51$) and Installments ($M = 1.75, SD = 0.79$).

7.2 Discussion

Previous evidence suggests that referring in installments is common for situated task-oriented interactions (Striegnitz et al., 2012). Further, it has been shown that such an approach is beneficial for referring expression generation in static scenes (Zarrieß & Schlangen, 2016). Their approach focused more on the type of information being output and rephrasing of an expression when it was resolved incorrectly. We investigated whether listener gaze can also be used to trigger automatically generated Installments in a more dynamic setting, and we compare this approach with providing a full reference at once (NoInstallments).

Previous work by Fang et al. (2015) reports a significant performance drop when incorporating listener gaze into a generation algorithm and using it to trigger installments. Here we tested the usefulness of listener gaze to automatically generate installments in a different setup. Crucially, we implemented a different inspection detection method, namely to measure the duration of a fixation and not the number of fixations to relevant objects. Once the threshold was exceeded (200ms), our system output the next installment. This method was initially proposed by Garoufi et al. (2016) to generate proactive verbal feedback and we validate that it is applicable to piece-wise *InformationDelivery*.

Our results show that following Installments made it take longer to solve the task. A possible explanation is that listeners could have hesitated to grasp an object after receiving only a partial object description. The main caveat of our approach is probably that it takes some time until the listener looks at an object, our system detects the listener's

intention and only after that the next installment is output. On the other hand, in the collected interactions we obtained a very low error rate because our system interprets the listener's gaze cues, which are an early indicator of language understanding. When an object that was not the intended target was inspected, the system provides more information, crucially before performing an action, which reduces the number of wrong grasps.

We demonstrated that more interactive instruction generation using listener gaze is an effective strategy for goal-oriented interactions in the real world. Specifically, when the *SpatialDescriptor* was mentioned first, right at the beginning of the instruction, the INSTALLMENTS approach was as efficient as NOINSTALLMENTS with *SpatialDescriptor* mentioned second.

Chapter 8

Conclusion

This thesis investigated the role of listener gaze in situated spoken language interaction. We examined whether this non-verbal cue can facilitate human-human interaction and if it can improve human-machine collaboration. We addressed these research questions in various settings and studied them from different perspectives. Specifically, we evaluated the influence of listener gaze on automated prediction in virtual environments. Further, we investigated the role of listener gaze in an indoor guidance task, where two human interlocutors, a remote speaker and a listener walking inside a hall solved nine tasks together. The core of this thesis is the development of an artificial speaker, a multimodal assistance system, employed in a real environment to assist the user. The system exploits listener gaze to automatically generate an instruction that identifies a real-world object for assembly. We employed the system in three experiments to investigate how it interacts with real users.

In Section 8.1 we summarize our contributions and in Section 8.2 we discuss limitations and further research directions.

8.1 Summary

Lister gaze indicates language understanding and mirrors the listener's intentions (Tanenhaus et al., 1995). Importantly, interpreting gaze cues can improve the performance of an NLG

system (Garoufi et al., 2016). Our work replicated previous findings from virtual environments in a real setup. Furthermore, we extended previous work concerning referential success to also consider listener gaze.

Firstly, in Chapter 3, we investigated whether listener gaze can facilitate automatic prediction of reference resolution. As listener gaze is tightly linked to language and mirrors comprehension processes, we expected that augmenting a probabilistic model with such information would improve the accuracy of the model. Specifically, we extended an observational model proposed by Engonopoulos et al. (2013) with eye-tracking features and obtained a performance gain. Our results showed that encoding listener gaze awareness improves the accuracy particularly in hard referential scenes with many competitor objects.

Secondly, in Chapter 4, we investigated whether and how a human speaker would use listener gaze from an egocentric perspective during remote instruction giving. We designed an indoor guidance task to investigate the interplay of spontaneous speech and visual behavior. The tasks consisted of complex referential scenes such that it was not trivial to refer to an object. We varied the availability of listener gaze to the speaker by either not showing the gaze cursor, or showing the exact or a slightly shifted gaze position. Our results show that human speakers are very good at producing references, so additional information as to what the listener currently is fixating, did not have an impact on performance. However, we observed a tendency that speakers produced more negative feedback when they had access to the exact gaze position and that visual behavior differs in this condition just before and right after an utterance. These findings suggest that listener gaze can be seen as a *symptom*, i.e. an indicator of comprehension processes, but also as a *signal* that affects feedback type. Our investigation of more coarse-grained measures on the collected spoken material, such as utterance length (in words), did not reveal a significant effect. However, many words do not necessarily carry more information. Importantly, the salience threshold for the speech segmentation is a crucial parameter and can vary depending on the domain, the task and the setting, e.g. whether it is a uni- or bidirectional, free or goal-oriented conversation. Further, the word level may be too coarse to reveal qualitative differences in utterances as a function of listener gaze. Thus, we further annotated the type of referring expressions uttered during the recorded interactions. Surprisingly, our manipulation did not affect the type of referring expressions, although we expected to see more deictic expressions when gaze was visible. However, we observed

that speakers systematically used particular types of referring expressions independent from our manipulation and possibly characteristic for the task. Instructors generally produced rather *specific* expressions, like definite noun phrases (e.g. the blue pen), and interestingly, more *featural* expressions, which mentioned the object's attributes, than *spatial* expressions, which mentioned the location of the object.

Thirdly in Chapter 5, we provide a proof of concept that listener gaze can be used to augment NLG in real-time interactions taking place in a real setup. Our scenario is collaborative assembly, where a human listener follows system's instructions aiming at identifying specific objects to be collected and assembled. We used mobile eye tracking and augmented reality technology to realize the semantic mapping of object inspections (Pfeiffer & Renner, 2014). We proposed two NLG systems that use listener gaze either directly, to generate feedback, or more indirectly, to provide an instruction incrementally in installments. Our first system NLG system *"Feedback"* varied the interaction strategy by generating a short, ambiguous instruction or a long, unambiguous one. It further outputs verbal feedback in response to object inspections of different specificity: underspecified feedback (e.g. "No, not that one!") or more informative feedback that specifies the relative position of a target (e.g. "Further left!"). Our second system NLG system *"Installments"* implemented two information delivery approaches, either outputting the entire description at once or delivering it piece-wise in gaze-driven installments. Moreover, our system varied the order of mentioning the spatial expression — either first or second — in order to facilitate the search process, because this information restricts the search space. We conducted three experiments to test the effectiveness of these interaction approaches with users. In the first two experiments, we invited people to interact with the NLG system *"Feedback"* and investigated the role of gaze-driven feedback and its specificity using different experimental designs (see Chapter 6). In Experiment 1, we replicated previous findings from virtual environments, namely that gaze-driven feedback after an exhaustive, unambiguous instruction improves performance as opposed to when no feedback followed. Our novel contribution here is that the combination of an ambiguous instruction, i.e. a partial description with contrastive feedback, outperforms following an exhaustive, unambiguous instruction. Further, in Experiment 2 we observed that the presence of contrastive feedback influences a listener's engagement with the instruction-giving system and that this also has an impact on performance. That is, the expectation that the system will give additional information helps the listener to better perform even in the more difficult condition, when feedback was underspecified, only giving a warning but not providing

further information. The third experiment was designed to test the effectiveness of the gaze-driven installments and we employed the NLG system *"Installments"* for the interaction (see Chapter 7). Surprisingly, our findings from the third experiment revealed that long, exhaustive instructions were followed faster than gaze-driven installments. However, contrary to the results of Fang et al. (2015), who considered this question in their work, we showed that referring in gaze-driven installments can be as efficient as a long description when the spatial expression appears first, right at the beginning of an instruction.

8.2 Discussion

The findings described in this thesis provide evidence for the usefulness of listener gaze in various settings encompassing human-human and human-machine interactions in real environments. We discuss the implications and contributions of our work and address some open questions, limitations and possible future research directions.

Extending an Observational Probabilistic Model

In Chapter 3 we have shown that the listener's gaze is useful by showing that accuracy, improves over an observational model by including features from the visual context for predicting the resolution of a referring expression. In addition, we observed that our extended model turns out to be more robust than the basic model when the time interval between the prediction and the button press increases, i.e. gaze is especially beneficial in an early stage of an interaction. This approach shows significant accuracy improvement on hard referential scenes where more objects are visible. We have also established that gaze is particularly useful when combined with other simple features, as the features capturing a listener's visual behavior are not powerful enough to outperform even the simplest baseline. Gaze only benefits the model when it is added on top of features that capture the visual context, i.e. the current scene. This means that gaze alone is not sufficient to accurately predict reference resolution in such a dynamic navigational setting. Since Engonopoulos et al. (2013) showed that the combination of the basic observational model with the semantic model achieved the best performance, an immediate next step would be to combine the extended observational model with the semantic model. This was beyond

the scope of our study. Such a model could provide reliable predictions early enough and so give an accurate estimate before an action takes place. This aspect is particularly important for real-time interactions. That is, if the prediction model is embedded in an NLG system, it can improve the automatic language generation in such scenarios because it captures the focus of the listener's attention. Given that our work refers only to NLG systems, no analysis of a speaker's gaze is possible. However, it may be interesting to ask whether a human speaker could benefit from the predictions of the extended observational model. We could study whether predictions based on the gaze (mis-)match between both interlocutors are more effective than simply presenting the listener's gaze to the speaker and trusting the speaker to correctly interpret this signal. If such an approach is effective, it could point out misunderstandings to the speaker before either participant becomes aware of them and help optimize collaboration toward achieving a mutual goal.

Listener Gaze in Human-Human Interactions

In our exploratory study presented in Chapter 4, where we considered human-human interactions, we observed that the availability of listener gaze to the speaker did not affect the overall performance (the task completion time). We believe that this could be due to a ceiling effect; that is, speakers are very good at describing co-present objects even in complex referential scenes. This is contrary to findings in the joint attention literature, where mostly face-to-face social interactions are considered. Following gaze is useful and helps to better coordinate turn taking and predicting the intentions of the conversational partner (e.g. Raidt, Bailly, & Elisei, 2007; Foulsham et al., 2010). In our study, making listener gaze available to the speaker did not shorten interaction time. A possible explanation is the nature of the setup, where listener gaze is projected on the egocentric video. Moreover, it could have been too difficult to exploit this information while at the same time spontaneously planning and producing a unique description in the cluttered scene.

We observed different listeners' gaze behavior and particularly main effects of *GazeAvailability* before and after an utterance. This suggests that listeners used their gaze as a *signal* to communicate to the speaker. The lack of such an effect while listening to an utterance indicates that gaze is a *symptom* of language comprehension processes. Further,

the observation that more negative feedback was produced when listener gaze was available supports the claim that listener gaze was used as a signal to which speakers actively reacted. These feedback instances have the potential to quickly eliminate wrong beliefs by the listener about intended referents.

Our findings are in agreement with the results of Coco, Dale, and Keller (2018), who examined the role of feedback and alignment in a "spot the difference" task. Their study revealed that only if interlocutors could not exchange verbal feedback, their gaze aligned. Both studies indicate that exploiting a technical augmentation of the listener gaze (e.g. by visualizing a gaze cursor is not something that human speakers naturally do efficiently. The instructors were faced with the additional perception task of following gaze cursors, which might have increased the cognitive load too much. In contrast, an NLG system can easily exploit gaze. This is computationally inexpensive and can be used to generate verbal feedback as a response to eye movements. Depending on the task, parametrization could vary.

Initially we expected that in the collected corpus of interactions we would observe similar instructions and could use them as the basis for designing an instruction-giving system. However, we encountered very high variability of the lexical choices made by the speakers. That is, human speakers have an individual way of describing objects and use very diverse expressions. There was a systematic use of featural expressions most probably driven by the task. It may be interesting to investigate if the referring expressions emerging in such complex referential scenes are overspecified. This may be difficult to determine and could require a practice session and more annotators to resolve potential disagreement.

One caveat to this study is that the presence of hand movements and pointing gestures or hovering over objects probably added noise to the role of listener gaze as a feedback modality. The hand is much more prominent than the gaze cursor in the streamed scene video, such that it could also trigger a reaction. However, this is typical for such setups and it means that may be the speaker cannot easily separate both modalities and they also frequently overlap. It may be worth investigating if showing listener gaze to the speaker would have an effect when hand movements are restricted or not allowed at all while listening to an instruction. Moreover, perhaps it is difficult for a human speaker to constantly monitor and interpret the gaze signal. Or the mediation of gaze information by a gaze pointer overlaid on a scene camera video, as was used in that study, was creating an artificial situation that speakers could not exploit intuitively and efficiently.

Further, the experiment consisted of a micro and a macro scale task, the latter of which was originally intended to be more of a navigation task. The actual reduction in task complexity (and therefore the omission of the macro task from the analyses) was due the significant technical challenges of setting up a stable WLAN connection throughout a large building to transfer high-resolution video, audio, and gaze data in real time. Moreover, mono eye trackers do not handle eye movements as good as stereo systems that use two cameras, and thus the data quality is impacted. Further, during calibration the device was adjusted for the micro task. This makes the evaluation of the macro task with respect to the gaze availability uninterpretable.

Augmenting NLG with Listener Gaze in a Real-world Setup

Interactive systems which use natural language in situ to assist a user in solving a task can benefit from exploiting listener gaze. Although the gaze signal is continuous and rapid, there is evidence that it can effectively be exploited by an NLG system designed to give directions to a listener and to refer to objects in a virtual environment (Koller et al., 2012; Staudte et al., 2012). There the authors showed that using listener gaze led to higher success rates. Real-world interactions are noisier and the system's knowledge about the environment is usually far from perfect. Thus, it is more challenging to make use of listeners' eye movements in such a setting. We employed an artificial speaker, that is, a parametrized NLG system, which tracks users' eye gaze to real objects while simultaneously planning an utterance. The system has the advantage of generating instructions systematically and without the great variation that is typical for human production data. Such control over the (artificial) speaker allows us to integrate different modalities in the interaction without much additional effort, while avoiding recursive effects between independent and dependent variables (variation by the speaker would affect listener behavior, which in turn could affect the speaker). Importantly, providing gaze-driven feedback triggered by object inspections is computationally inexpensive for our system but enables it to be even more interactive and to better engage with the listener. Our experimental investigation with this system supports this view and suggest that exploiting listener gaze in real-world human-machine collaboration can indeed be beneficial. Our results extend previous research by looking at interactions with increased interactivity with an assistance system: Instead of generating long unambiguous instructions providing all required information, our system split the information and provided it

on demand, by giving partial instructions and requiring a non-verbal cue from the listener to progress the communication. While Experiment 1 showed that this might be considered more demanding, even exhausting, as listeners were more involved, the assessment of using such variants of installments to refer to co-present objects (ambiguous instruction with informative feedback) revealed that the interaction flow was perceived positively and rated as highly as following an unambiguous instruction. Moreover, an interaction strategy that refers to objects incrementally and reacts to listeners' gaze can be used to identify objects in the shared space faster. Experiment 2 then examined if the benefit of contrastive feedback is inherent to it or whether there is a learning effect specific to this system's behavior. Here, the system provided underspecified or contrastive, more informative feedback in an interleaved manner. Somewhat surprisingly, the results revealed that both conditions now led to equally high task performance: Participants were equally efficient in completing the task when listening to underspecified or contrastive feedback given the different study designs, although obtaining different results is not that unusual (Charness, Gneezy, & Kuhn, 2012). Specifically, we interpret the performance gain in Experiment 2 as a natural adaptation to the system's informative behavior which extends to and even absorbs the not-so-informative trials. Supportive evidence for this interpretation comes from the sequential feedback analysis, which shows that gaze was used more deliberately and this helps to quickly advance within a trial. Lastly, given that not only the specificity of gaze-driven feedback improves task performance, but that the listener's perception of an assistance system also influences it, an adaptation of the instructions' form could possibly contribute further to efficiency. In general, considering the form of automatically generated utterances when designing a system is important. Politeness is a key aspect of interaction design and can improve usability, making a system more user-friendly (Pemberton, 2011). The notion of human-computer interaction etiquette has been discussed by Hayes, Pande, and Miller (2002); this is a necessary but not sufficient criterion to establish effective interaction. Especially in urgent situations and under time pressure, social appropriateness is not as important as efficiency, as shown by Kellermann and Park (2001). A direction for future research could be to vary the syntactic structure and the lexicalization when generating an ambiguous instruction to examine the effect of the politeness aspect in this context.

Further, referring in installments is common in situated task-oriented interactions and has been shown to be beneficial for referring expression generation (Zarrieß & Schlangen, 2016). We investigated whether listener gaze can also be used to trigger automatically

generated installments and compared this approach to providing a full reference at once. Both information delivery approaches are realized in our NLG system *"Installments"*.

Although Fang et al. (2015) reported a performance drop when incorporating listener gaze into their generation algorithm, we test the usefulness of listener gaze in a different setup and crucially with a different inspection detection method proposed by Garoufi et al. (2016). We demonstrated that more interactive instruction generation which uses listener gaze is feasible for goal-oriented interactions in the real world, too. However, following installments made it take more time to solve the task. A possible explanation is that listeners might hesitate to grasp an object after receiving only a partial description. Additionally, they might need some time to get used to this style of interaction. Both interaction approaches were similarly perceived in terms of interaction flow. That validates the appropriateness of using listener gaze to deliver information piece-wise. Similar to the assessment in Experiment 1, most of the listeners preferred to follow an exhaustive description, probably because they felt more confident in their performance. However, nearly one third of the participants (29.2%) in Experiment 3 favored the incremental approach as opposed to Experiment 1: none of those participants preferred it when feedback was underspecified, and fewer of them when feedback was informative (12.5%). This suggests that the more indirect use of gaze cues with respect to the lexical realization of a spatial expression positively influences the perception of the interaction. These findings can be helpful when designing an interactive system, and depending on the goal of the application, the more appropriate approach could be used, i.e. optimizing for efficiency by using feedback or for better user perception by using gaze-driven installments.

A direction for future investigation could be to validate the effectiveness of gaze-sensitive instruction generation for another domain. Further, the perception of the interactive systems was not as good as that of the non-interactive ones. Thus there is some room for improvement and it may be that the form of the partial instruction should prepare the listener to expect that some information will follow, e.g. using indefinite noun phrases, such that a partial description does not feel disadvantageous.

In general it may be questionable why an assistance system should give verbal instead of non- verbal feedback like showing an arrow, playing a beep or even highlighting the relevant objects. These styles of interaction are effective and can be efficient; they are typically used for interactive collaborative assembly (e.g. Renner & Pfeiffer, 2017). Importantly, language is our usual mean of communication and there is no additional training

required to decipher its semantics because people already understand it. Thus it can be used for any domain and conceivable scenario. Perhaps a combination of both visual signals and verbal responses would be an optimal solution for an instruction-giving system that guides the user in how to complete a task. Sometimes the visualization of an action could be technically difficult (e.g. turning a large construction around) and so hard to comprehend, while at the same time, it could be expressed with a few words (e.g. *"Please turn it around"*) and so ensure correct understanding. On the other hand, instead of generating a long, exhaustive instruction that identifies an object, it may be easier to highlight it and use only a deictic expression like "Please take this". That would minimize misunderstandings and ensure an efficient and less error-prone human-machine interaction.

Final Remarks

In conclusion, we argue that gaze information can be used to automatically predict reference resolution. Further, showing listener gaze to a human speaker from an egocentric perspective does not affect performance because it is possibly too demanding to constantly interpret it while planning an identifying instruction in a complex visual scene.

Importantly, assistance systems that generate natural language to interact with the user can successfully integrate listener gaze into their generation mechanisms in real, noisier environments. Exploiting this information source minimizes error rate and optimizes speed; that is, shortens the interaction time when the response to gaze is in the form of verbal feedback, i.e. is directly connected to the gaze, but not when it is in form of installments. At the same time, we found that efficiency does not necessarily correlate with the perceived agreeableness. We provide a proof of concept that a system can use listener gaze in real environments to narrow down visual search, and validate the effectiveness of this interaction approach using a sample assembly scenario.

Appendix A

Micro Tasks used in the indoor guidance study

1. Brief schreiben, 35 Obj.
(a) Umschlag
(b) Kleber
(c) Stift
(d) Block

2. Kuchen backen, 28 Obj.
(a) Milch
(b) Dekor
(c) Löffel
(d) Eier

3. Sensor aufbauen, 28 Obj.
(a) Klebeband
(b) Box
(c) Sensor in der Box

4. Cocktail mixen, 26 Obj.
(a) Flasche
(b) Messbecher
(c) Glas
(d) Strohhalm

5. Notiz schreiben, 28 Obj.
(a) Schere
(b) Klebezettel
(c) Marker
(d) Block

6. PC aufbauen, 15 Obj.
(a) Kabel
(b) Maus
(c) Tastatur

7. Stift spitzen, 30 Obj.
(a) Spitzer
(b) Stift
(c) Radiergummi
(d) Klebezettel

8. Papiere heften, 32 Obj.
(a) Locher
(b) Papier
(c) Schnellhefter

9. Tee kochen, 26 Obj.
(a) Tasse
(b) Löffel
(c) Tee

Appendix B

GazInG: Preliminary Studies

In two preliminary studies, we tested the object density and if the latency of the gaze-driven feedback is acceptable. One requirement for the system is that the verbal feedback is generated in a timely way such that a delay will not cause misunderstandings. Hence, these studies we used predefined verbal instructions.

B.1 Object density

During the initial test runs, the positioning and density of the real-world objects were tested. There are 25 fiducial markers in total, on which target and competitor objects can be positioned on the marker field, 13 of which are large (located at the edges of the table) and 12 small (in the inner area). The corresponding bounding boxes in the 3D model also have different sizes, respectively. Figure B.1 shows the initial object positioning with the marker IDs (a). The calibration and fixation-to-marker mapping accuracy for the large markers is quite good, even for the last row in the back where the depth is increased. In contrast, fixation detection and the performance of the fixation-to-marker mapping algorithm for the smaller makers that are close to each other is rather sub-optimal when all of them are occupied (i.e. for marker IDs 66, 58, 89, 98, 112, 126, 195, 138).

The verbal feedback provided by the interactive system is triggered by an inspection of a particular object. An inspection is defined as a fixation that has a duration exceeding a certain threshold. The threshold for the inspection duration we currently experiment

with is set to 300ms. Accurate calibration is a prerequisite for implementing the correct behavior of the system. Thus it is crucial to ensure the best possible calibration accuracy in order not to output incorrect and inappropriate feedback, which could lead to incorrect actions or even interrupting an interaction. Thus, four objects that were originally located on small markers were removed from the scene (see Figure B.1 b). In this manner, we reduced the complexity and ensured an appropriate distance between the objects. Additionally, the size of the 3D object bounding boxes was adjusted by enlarging the smaller ones. Another advantage of removing four objects and having visible markers in each quadrant of the marker field is that it improves the stability of the 3D model, even if the listener is looking around and moving her head before any other objects are grasped.

FIGURE B.1: The GazInG setup: full (left picture) vs. reduced (right picture) referential scene.

B.2 Timing and Usefulness of Gaze-based Feedback

The main challenge of our mobile setup is the temporal synchronization of the two software components: the one that tracks listeners' eye movements and the one that outputs speech. To make the system's behavior believable and usable they need to be in synchronization, i.e. the system's feedback statements have to correspond to a listener's eye movements in order to be interpretable.

For this reason, we conducted a preliminary study to test the timing of the system's feedback. In particular we were interested in users' judgment of the gaze-speech synchronization. Importantly, the tracker necessarily receives the gaze data with some delay. Therefore it is very important to make sure that this does not disturb the interaction with the system. Participants followed human-authored ambiguous instructions describing eight predefined target objects. The instructions were also pre-synthesized such that only the playback happens on the fly.

After an interaction with the system, a post-questionnaire was filled in by each subject in order to assess if the feedback timing was appropriate and if the feedback statements were helpful. Eight subjects participated in the study (one male and six female). They answered eight questions, three of which used a 5-point rating scale (where 1 fits best) and the remaining five were of which yes-no questions. In the following, the evaluation of the questionnaire is presented:

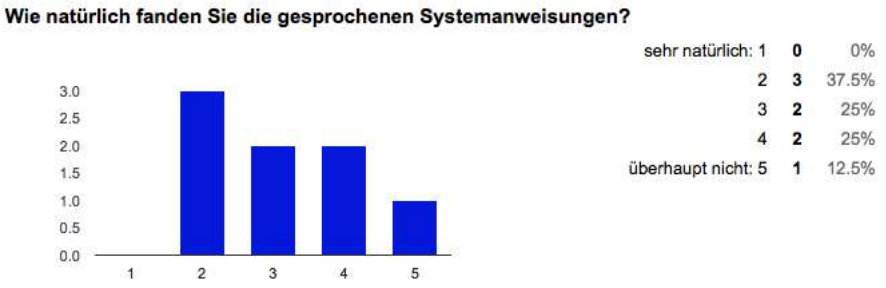

FIGURE B.2: How natural did you find the spoken system instructions?

Figure B.2 depicts how participants rated the naturalness of the system's instructions. The result is somewhat mixed: the instructions were mostly rated as rather natural or

neutral, but sometimes also unnatural. This may be partly because they are ambiguous, but also because of the synthesized speech.

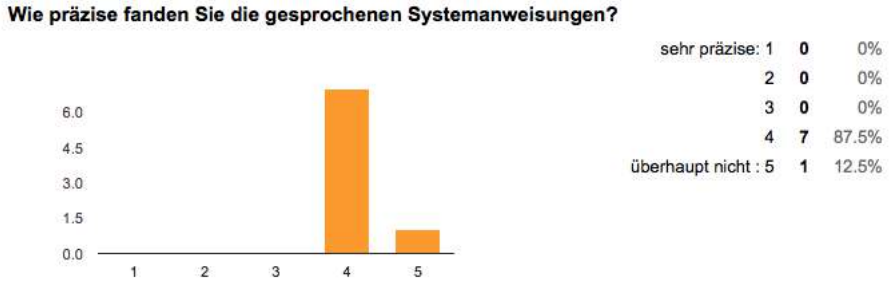

FIGURE B.3: How precise did you find the spoken system instructions?

Figure B.3 depicts how participants rated the precision of the system's instructions. The instructions were judged as rather imprecise, but as participants experienced only ambiguous descriptions, it was expected to see such ratings.

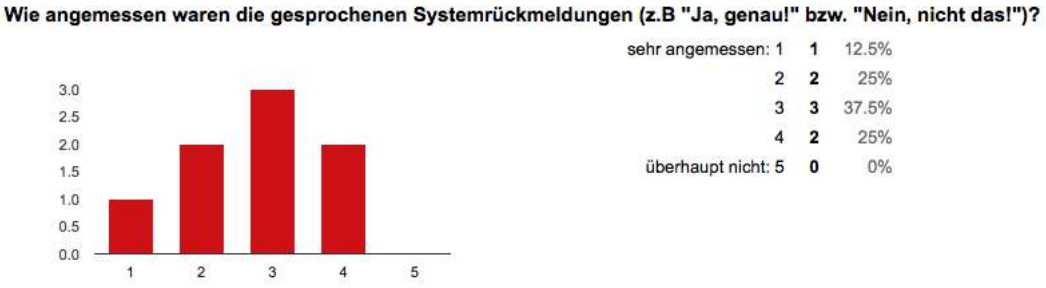

FIGURE B.4: How adequate was the system's feedback?

Figure B.4 depicts how participants rated the adequateness of the system's instructions. The ratings were again varied, ranging from very adequate to rather inadequate. The calibration mostly worked good but for three of the participants it was somewhat noisy.

Figure B.5 shows the results of how people answered a yes-no question about the timing of the utterance. Five out of eight people answered this question with yes. Sometimes participants expected the system to be quicker than it was.

The questions in Figures B.6, B.7 and B.8 were answered by seven subjects with "yes" while only one person answered "no". The questions aimed to assess the usefulness and importance of the feedback statements output by the system.

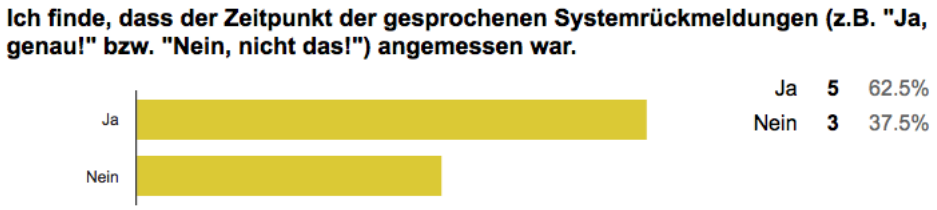

FIGURE B.5: I think that the timing of the system's feedback was appropriate.

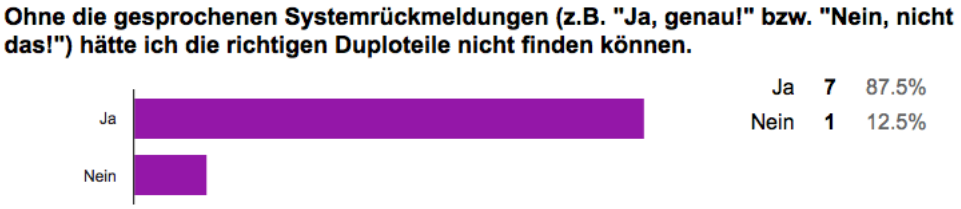

FIGURE B.6: Without the system's feedback I would not be able to find the right building blocks.

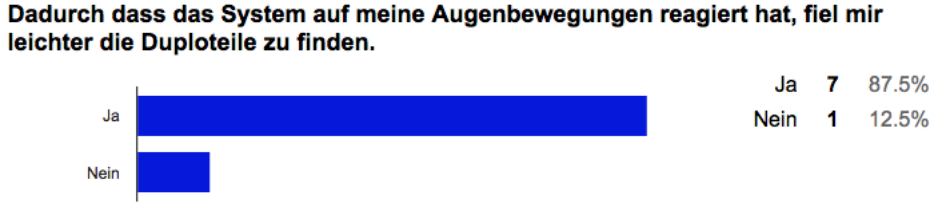

FIGURE B.7: Because the system reacted to my eye movements, it was easier for me to find the building blocks.

Finally, the participants were asked to list free-text comments to provide additional suggestions not covered by the questions they answered.

Some participants complained about the unnaturalness of the voice, but this concerns the synthesis system we used. However, we focused on the natural language generation and used an out-of-the-box synthesizer. The imprecision of the instructions was also criticized. However, we intended to give ambiguous instructions, which we later supplement with informative feedback.

In conclusion, the judgment of the feedback appropriateness was rather positive and confirmed that the flow of the interaction was smooth. This preliminary results validated

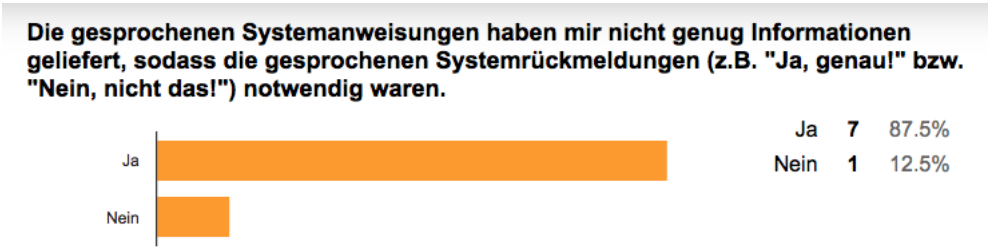

FIGURE B.8: The instructions did not contain enough information such that the system's feedback was crucial.

that interpreting listener gaze was reasonable to generate verbal feedback in real time and a real environment.

Appendix C

Scene Layouts for Assembly

C.1 Scene Layout 1 and Scene Layout 2 used in Experiment 1, 2 and 3

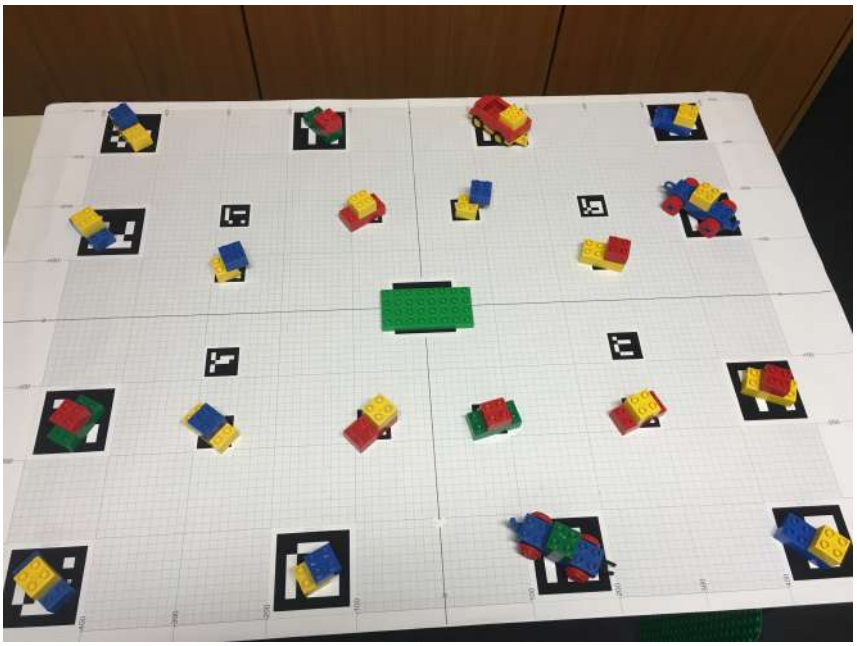

FIGURE C.1: First scene layout

FIGURE C.2: Second scene layout

C.2 Scene Layout 3 and Scene Layout 4 used in Experiment 2 and 3

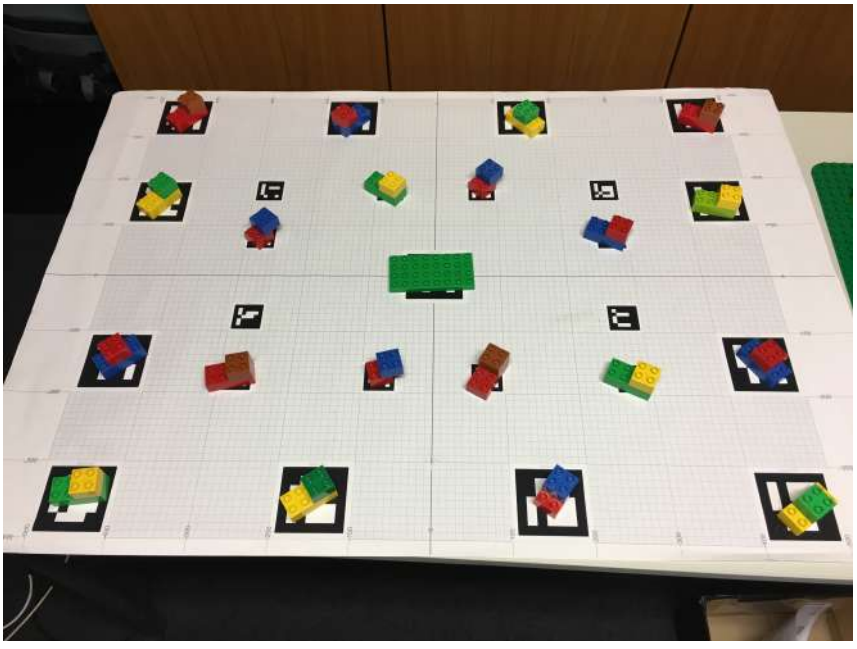

FIGURE C.3: Third scene layout

FIGURE C.4: Fourth scene layout

Appendix D

Questionnaires für Experiment 1

D.1 Assessment of the Interaction

1. Wie gut finden Sie den Ablauf der Interaktion?
2. Wie gut finden Sie die Art der Anweisungen?
3. Wie flexibel fanden Sie die Interaktion?
4. Wie präzise fanden Sie die gesprochenen Systemanweisungen?
5. Ich fand es notwendig eine Wiederholung der Anweisung zu verlanden.
6. Die Objektbeschreibung hatte andere Duploteile ausgeschlossen.
7. War Ihnen zu jedem Zeitpunkt klar, was Sie tun mussten?
8. Ich kann mir vorstellen ein solches System zu benutzen, wenn ich etwas zusammenbauen möchte, weil es die Suche der Teile erleichtert.
9. Ich denke, dass es einfacher sein wird, wenn man mit dem System sprechen kann.

D.2 Comparison of the Two Interaction Strategies

1. Die sprachliche Kommunikation der beiden Systeme war unterschiedlich.

2. Ich habe eine ausführliche Beschreibung des gesuchten Duploteils bevorzugt.

3. Eine ausführliche Beschreibung des gesuchten Duploteils war zu lang.

4. Dadurch dass ich in einer der Interaktionen Rückmeldungen auf meinem Blick erhielt (z.B. “Ja, genau!” bzw. “Nein, nicht das!”), fiel es mir leichter die Duploteile zu finden.

5. Die gesprochenen Systemruückmeldungen (z.B. “Ja, genau!” bzw. “Nein, nicht das!”) waren hilfreich.

6. Die gesprochenen Systemruückmeldungen (z.B. “Ja, genau!” bzw. “Nein, nicht das!”) waren verwirrend.

7. Ich fand es wichtig, dass ich dem System Signale mit meinem Blick geben konnte.

Appendix E

Questionnaire für Experiment 3

E.1 Assessment of the Interaction

1. Wie gut finden Sie den Ablauf der Interaktion?
2. Wie gut finden Sie die Art der Anweisungen?
3. Wie flexibel fanden Sie die Interaktion?
4. Wie präzise fanden Sie die gesprochenen Systemanweisungen?
5. War Ihnen zu jedem Zeitpunkt klar, was Sie tun mussten?
6. Ich kann mir vorstellen ein solches System zu benutzen, wenn ich etwas zusammenbauen möchte, weil es die Suche der Steine erleichtert.

E.2 Comparison of the Two Interaction Strategies

1. Die gesprochenen Anweisungen in beiden Hälften waren unterschiedlich.
2. Welche der beiden Interaktionen fanden Sie angenehmer?
3. Die ganze Beschreibung des gesuchten Steins war zu bevorzugen (auf einmal präsentiert).
4. Die schrittweise Präsentation der Beschreibung des gesuchten Steins war zu bevorzugen.

Appendix F

Model Selection Results for the Interactions with the GazInG System

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
model0	4	760.39	775.41	-376.19	752.39			
model1	5	702.13	720.91	-346.07	692.13	60.26	1	<0.001***

Group 1
Model 0: *identificationTime ~ 1 + (1 | Subject) + (1 | Item)*
Model 1: *identificationTime ~ InteractionStrategy + (1 | Subject) + (1 | Item)*

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
model0	4	845.16	860.52	-418.58	837.16			
model1	5	754.30	773.49	-372.15	744.30	92.87	1	<0.001***

Group 2
Model 0: *identificationTime ~ 1 + (1 | Subject) + (1 | Item)*
Model 1: *identificationTime ~ InteractionStrategy + (1 | Subject) + (1 | Item)*

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
model0	8	1610.55	1646.48	-797.28	1594.55			
model1	9	1606.65	1647.07	-794.33	1588.65	5.90	1	0.0151*

Group 1 and Group 2 AMBIGUOUS condition
Model 0: *identificationTime ~ 1 + (0+FeedbackSpecificity | Subject) + (0+FeedbackSpecificity | Item)*
Model 1: *identificationTime ~ FeedbackSpecificity + (0+FeedbackSpecificity | Subject) + (0+FeedbackSpecificity | Item)*

TABLE F.1: This table summarizes the models fitted to the time for identification data and the model comparison results of listener gaze behavior for Experiment 1. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Model : *negativeFeedbackInstances* ~ *Type + FeedbackSpecificity + (FeedbackSpecificity | Subject) + (FeedbackSpecificity | Item)*

	Estimate	Std. Error	Wald Z	p
(Intercept)	0.93	0.05	19.52	$< 0.001^{***}$
FeedbackType:neg	-0.09	0.05	-1.95	0.05^{*}
FeedbackSpecificity:underspecified	-0.01	0.05	-0.10	0.92

TABLE F.2: This table summarizes the model fitted to the feedback data and inferential statistics for Experiment 1.Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Df	AIC	BIC	logLik	deviance	χ^2	χ Df	Pr($>\chi^2$)
object	4	1635.67	1653.98	-813.84	1627.67			
..1	5	1637.61	1660.49	-813.80	1627.61	0.06	1	0.7991
..2	6	1637.78	1665.24	-812.89	1625.78	1.82	1	0.1771
..3	7	1639.77	1671.80	-812.88	1625.77	0.02	1	0.8987

Model 0: *identificationTime* ~ *1 + (1 | Subject) + (1 | Item)*
Model 1: *identificationTime* ~ *FeedbackSpecificity + (1 | Subject) + (1 | Item)*
Model 2: *identificationTime* ~ *FeedbackSpecificity + Half + (1 | Subject) + (1 | Item)*
Model 3: *identificationTime* ~ *FeedbackSpecificity * Half + (1 | Subject)+ (1 | Item)*

TABLE F.3: This table summarizes the models fitted to the listener gaze data and the model comparison results for Experiment 2. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Df	AIC	BIC	logLik	deviance	χ^2	Chi Df	p
Model0	5	-575.61	-553.32	292.80	-585.61			
Model1	6	-592.32	-565.58	302.16	-604.32	18.71	1	$< 0.001^{***}$
Model2	7	-604.82	-573.62	309.41	-618.82	14.50	1	$< 0.001^{***}$
Model3	8	-602.98	-567.32	309.49	-618.98	0.16	1	0.6903

Random Structure: *(InformationDelivery*SpatialDescriptor | Subject)*
Model 0: *identificationTime* ~ *1 + Random Structure*
Model 1: *identificationTime* ~ *SpatialDescriptor + Random Structure*
Model 2: *identificationTime* ~ *SpatialDescriptor + InformationDelivery + Random Structure*
Model 3: *identificationTime* ~ *SpatialDescriptor * InformationDelivery + Random Structure*

TABLE F.4: This table summarizes the models fitted to the performance data and the model comparison results. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Estimate	Std. Error	Wald Z	P
(Intercept)	0.83	0.05	16.59	$< 0.001^{***}$
SpatialDescriptor	-0.21	0.07	-2.88	$< 0.001^{***}$

Model: *numInstallments* $\sim$ *SpatialDescriptor + (1 | Subject), family = "poisson"*

TABLE F.5: This table summarizes analysis and the model fitted to the speech data. Differences are denoted to be significant at $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Bibliography

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Andrist, S., Gleicher, M., & Mutlu, B. (2017). Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2571–2582). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3025453.3026033` doi: 10.1145/3025453.3026033

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18 – 40. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0010027708001698` doi: https://doi.org/10.1016/j.cognition.2008.07.005

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015, 06). Parsimonious Mixed Models. , *1506*.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*(2), 237-266. Retrieved from `https://doi.org/10.1080/09541440143000050` doi: 10.1080/09541440143000050

Blattgerste, J., Strenge, B., Renner, P., Pfeiffer, T., & Essig, K. (2017). Comparing conventional and augmented reality instructions for manual assembly tasks. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 75–82). ACM. doi: 10.1145/3056540.3056547

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005, 08). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 708–713). doi: 10.1109/IROS.2005.1545011

Brennan, S. E., Schuhmann, K. S., & Batres, K. M. (2013). Entrainment on the move and in the lab: The walking around corpus. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Berlin, Germany.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0749596X0900045X` doi: https://doi.org/10.1016/j.jml.2009.04.003

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89. Retrieved from `https://doi.org/10.1080/01690965.2010.543363` doi: 10.1080/01690965.2010.543363

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science*, *32*(4), 643-684. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210802066816` doi: 10.1080/03640210802066816

Bulling, A., & Roggen, D. (2011). Recognition of visual memory recall processes using eye movement analysis. In *Proc. UbiComp* (pp. 455–464).

Bulling, A., Ward, J. A., Gellersen, H., & Tröster, G. (2011). Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, *33*(4), 741-753.

Campana, E., Baldridge, J., Dowding, J., Hockey, B. A., Remington, R. W., & Stone, L. S. (2001). Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces* (pp. 1–5). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/971478.971489` doi: 10.1145/971478.971489

Carter, M., Newn, J., Velloso, E., & Vetere, F. (2015). Remote gaze and gesture tracking on the Microsoft Kinect: Investigating the role of feedback. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction* (pp. 167–176). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2838739.2838778` doi: 10.1145/2838739.2838778

Channarukul, S. (1999). *YAG: a template-based natural language generator for real-time systems* (Tech. Rep.).

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. Retrieved from `www.sciencedirect.com/science/article/pii/S0167268111002289` doi: https://doi.org/10.1016/j.jebo.2011.08.009

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

Clark, H. H., & Krych, M. A. (2004, January). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81. doi: 10.1016/j.jml.2003.08.004

Coco, M. I., Dale, R., & Keller, F. (2018). Performance in a collaborative search task: The role of feedback and alignment. *topiCS*, *10*(1), 55–79. Retrieved from `https://doi.org/10.1111/tops.12300` doi: 10.1111/tops.12300

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 84–107.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233-263. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1902_3` doi: 10.1207/s15516709cog1902_3

DeVault, D., Traum, D., & Artstein, R. (2008, June). Practical grammar-based NLG from examples. In *The Fifth International Natural Language Generation Conference (INLG 2008).* Salt Fork, OH. Retrieved from `http://ict.usc.edu/pubs/Practical%20Grammar-Based%20NLG%20from%20Examples%20.pdf`

Eaddy, M., Blasko, G., Babcock, J., & Feiner, S. (2004). My own private kiosk: Privacy-preserving public displays. In *Eighth International Symposium on Wearable Computers, 2004. ISWC 2004.* (Vol. 1, pp. 132–135).

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409-436.

Engelhardt, E. P., Bailey, K., & Ferreira, F. (2006, 05). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, *54*, 554-573. doi: 10.1016/j.jml.2005.12.009

Engonopoulos, N., Villalba, M., Titov, I., & Koller, A. (2013). Predicting the resolution

of referring expressions from user behavior. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Seattle.

Fang, R., Doering, M., & Chai, J. Y. (2015). Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 271–278). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2696454.2696467` doi: 10.1145/2696454.2696467

Fischer, K., Jensen, L. C., Kirstein, F., Stabinger, S., Erkent, Ö., Shukla, D., & Piater, J. (2015). The effects of social gaze in human-robot collaborative assembly. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings* (pp. 204–213). Cham: Springer International Publishing. Retrieved from `https://doi.org/10.1007/978-3-319-25554-5_21` doi: 10.1007/978-3-319-25554-5_21

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., & Kingstone, A. (2010). Gaze allocation in a dynamic situation: Effects of social status and speaking. *Cognition*, *117*(3), 319–331. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0010027710002167` doi: https://doi.org/10.1016/j.cognition.2010.09.003

Galley, M., Fosler-Lussier, E., & Potamianos, A. (2001, September). Hybrid natural language generation for spoken dialogue systems. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH–01)* (p. 1735–1738). Aalborg, Denmark. Retrieved from `https://www.microsoft.com/en-us/research/publication/hybrid-natural-language-generation-for-spoken-dialogue-systems/`

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010, may). The give-2 corpus of giving instructions in virtual environments. In N. C. C. Chair) et al. (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Garkavijs, V., Okamoto, R., Ishikawa, T., Toshima, M., & Kando, N. (2014). GLASE-IRUKA: Gaze feedback improves satisfaction in exploratory image search. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 273–274). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2567948.2577313` doi: 10.1145/2567948.2577313

Garoufi, K., & Koller, A. (2010). Automated planning for situated natural language generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL).* Uppsala.

Garoufi, K., Staudte, M., Koller, A., & Crocker, M. W. (2016). Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*, *40*(7), 1671–1703. Retrieved from `https://doi.org/10.1111/cogs.12298` doi: 10.1111/cogs.12298

Gatt, A., & Reiter, E. (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 90–93). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1610195.1610208`

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics: Vol. 3: Speech Acts* (p. 41-58). New York: Academic Press. Retrieved from `http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf`

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274–279.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*(1), 105–115.

Hayes, C., Pande, A., & Miller, C. (2002). Etiquette in human computer interactions: What does it mean for a computer to be polite? or Who needs polite computers anyway? In *Proceedings of the Workshop on Etiquette for Human-Computer Work, held at the AAAI Fall Symposium (AAAI Press, California* (pp. 15–17).

Henderson, J. M., & Smith, T. J. (2007). How are eye fixation durations controlled during scene viewing? Further evidence from a scene onset delay paradigm. *Visual Cognition*, *17*(6-7), 1055–1082. doi: 10.1080/13506280802685552

Imai, M., Ono, T., & Ishiguro, H. (2003, August). Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics*, *50*(4), 636–643. doi: 10.1109/TIE.2003.814769

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. UbiComp* (pp. 1151–1160). Retrieved from `http://pupil-labs.com/pupil/`

Kelleher, J. D., & Kruijff, G.-J. M. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference*

on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (pp. 1041–1048). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1220175.1220306 doi: 10.3115/1220175.1220306

Kellermann, K., & Park, H. S. (2001). Situational urgency and conversational retreat: When politeness and efficiency matter. *Communication Research*, *28*(1), 3–47.

Kennington, C., & Schlangen, D. (2014). *Comparing listener gaze with predictions of an incremental reference resolution model. RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production.* Retrieved from http://www.macs.hw.ac.uk/InteractionLab/refnet/abstracts/refnet2014_submission_35.pdf

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Kirk, D., Rodden, T., & Fraser, D. S. (2007). Turn it this way: Grounding collaborative action with remote gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1039–1048). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1240624.1240782 doi: 10.1145/1240624.1240782

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: the use case WebMAUS. In *Proceedings Digital Humanities 2012, Hamburg, Germany* (pp. 30–34). Hamburg.

Klarner, M., & Ludwig, B. (2004). Hybrid natural language generation in a spoken language dialog system. In S. Biundo, T. Frühwirth, & G. Palm (Eds.), *KI 2004: Advances in Artificial Intelligence* (pp. 97–111). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koleva, N., Hoppe, S., Moniri, M. M., Staudte, M., & Bulling, A. (2015). On the interplay between spontaneous spoken instructions and human visual behaviour in an indoor guidance task. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015.* Retrieved from https://mindmodeling.org/cogsci2015/papers/0204/index.html

Koleva, N., Villalba, M., Staudte, M., & Koller, A. (2015). The impact of listener gaze on predicting reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference*

on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers (pp. 812–817). Retrieved from http://aclweb.org/anthology/P/P15/P15-2133.pdf

Koller, A., Staudte, M., Garoufi, K., & Crocker, M. (2012). Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 30–39). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2392800.2392806

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., & Oberlander, J. (2010). The first challenge on generating instructions in virtual environments. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation* (pp. 328–352). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-15573-4_16 doi: 10.1007/978-3-642-15573-4_16

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., ... Oberlander, J. (2010). Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231 - 3250. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378216611001731 doi: https://doi.org/10.1016/j.pragma.2011.06.008

Kopp, S., Jung, B., Leßmann, N., & Wachsmuth, I. (2003). Max - A multimodal assistant in virtual reality construction. *KI*, *17*(4), 11.

Kosunen, I., Jylha, A., Ahmed, I., An, C., Chech, L., Gamberini, L., ... Jacucci, G. (2013). Comparing eye and gesture pointing to drag items on large screens. In *Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces* (pp. 425–428). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2512349.2514920 doi: 10.1145/2512349.2514920

Krahmer, E., van Erk, S., & Verleg, A. (2003, March). Graph-based generation of referring expressions. *Comput. Linguist.*, *29*(1), 53–72. Retrieved from http://dx.doi.org/10.1162/089120103321337430 doi: 10.1162/089120103321337430

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces. In *Proceedings of the Third International*

Conference on Advances in Multimodal Interfaces (pp. 1–7). London, UK: Springer-Verlag. Retrieved from http://dl.acm.org/citation.cfm?id=645524.656806

Mitchell, M., van Deemter, K., & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1174–1184). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N13-1137

Mitev, N., Renner, P., Pfeiffer, T., & Staudte, M. (2018). Using listener gaze to refer in installments benefits understanding. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018, Madison, Wisconsin, July 25-28, 2018.*

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*.

Pemberton, L. (2011). Politeness in interaction design. *Romanian Journal of Human-Computer Interaction*, 1.

Peng, H. (2007). *mRMR Feature Selection Toolbox for MATLAB, http://penglab.janelia.org/proj/mrmr/.* Retrieved from http://penglab.janelia.org/proj/mRMR/

Pfeiffer, T. (2012). Using virtual reality technology in linguistic research. In S. Coquillart, S. Feiner, & K. Kiyokawa (Eds.), *Virtual Reality Short Papers and Posters (VRW)* (pp. 83–84). Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/VR.2012.6180893

Pfeiffer, T. (2013). Gaze-based assistive technologies. In G. Kouroupetroglou (Ed.), *Assistive Technologies and Computer Access for Motor Disabilities* (pp. 90–109). IGI Global. doi: 10.4018/978-1-4666-4438-0.ch004

Pfeiffer, T., Feiner, S. K., & Mayol-Cuevas, W. W. (2016). Eyewear computing for skill augmentation and task guidance. In A. Bulling, O. Cakmakci, K. Kunze, & J. M. Regh (Eds.), *Eyewear Computing–Augmenting the Human with Head-Mounted Wearable Assistants* (Vol. 23, p. 199). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany. doi: 10.4230/DagRep.6.1.160

Pfeiffer, T., & Latoschik, M. E. (2004). Resolving object references in multimodal dialogues for immersive virtual environments. In Y. Ikei, M. Göbel, & J. Chen (Eds.), *Proceedings of the IEEE Virtual Reality 2004* (pp. 35–42).

Pfeiffer, T., & Renner, P. (2014). EyeSee3D: A low-cost approach for analyzing mobile 3D eye tracking data using computer vision and augmented reality technology. In

Proceedings of the Symposium on Eye Tracking Research and Applications (pp. 369–376).

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raidt, S., Bailly, G., & Elisei, F. (2007). Gaze patterns during face-to-face interaction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops* (pp. 338–341). Washington, DC, USA: IEEE Computer Society. Retrieved from `http://dl.acm.org/citation.cfm?id=1339264.1339721`

Reiter, E. (1995). NLG vs. templates. In *Proceedings of the 4th European Workshop on Natural Language Generation (EWNLG 1995).*

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems.* New York, NY, USA: Cambridge UniversityA Press.

Renner, P., & Pfeiffer, T. (2017). Attention guiding techniques using peripheral vision and eye tracking for feedback in augmented-reality-based assistance systems. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)* (pp. 186–194). IEEE. doi: 10.1109/3DUI.2017.7893338

Reynal, M., Colineaux, Y., Vernay, A., & Dehais, F. (2016). Pilot flying vs. pilot monitoring during the approach phase: An eye-tracking study. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace* (pp. 7:1–7:7). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2950112.2964583` doi: 10.1145/2950112.2964583

Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., & Ikeuchi, K. (2004). Flexible cooperation between human and robot by interpreting human intention from gaze information. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, *1*, 846–851.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proc. ETRA* (pp. 71–78).

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *In Proceedings of the ACL SIGDAT-Workshop* (pp. 47–50).

Schröder, M., Charfuelan, M., Pammi, S., & Steiner, I. (2011, 8). Open source voice creation toolkit for the MARY TTS platform. In *Proceedings of Interspeech 2011.* ISCA.

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: A study of

human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (pp. 78–84). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/964442.964458 doi: 10.1145/964442.964458

Staudte, M., Koller, A., Garoufi, K., & Crocker, M. W. (2012). Using listener gaze to augment speech generation in a virtual 3D environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Sapporo, Japan.

Stent, A. (2001, 01). Content planning and generation in continuous-speech spoken dialog systems.

Stent, A., & Bangalore, S. (2014). *Natural Language Generation in Interactive Systems.* New York, NY, USA: Cambridge University Press.

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation* (pp. 81–88).

Striegnitz, K., Buschmeier, H., & Kopp, S. (2012). Referring in installments: A corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the 7th International Natural Language Generation Conference* (pp. 12–16).

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tessendorf, B., Bulling, A., Roggen, D., Stiefmeier, T., Feilner, M., Derleth, P., & Tröster, G. (2011). Recognition of hearing needs from body and eye movements to improve hearing instruments. In *Proc. Pervasive* (pp. 314–331).

Torrey, C., Fussell, S., & Kiesler, S. (2013). How a robot should give advice. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction* (pp. 275–282). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2447556.2447666

Turk, M., & Robertson, G. (2000, March). Perceptual user interfaces (introduction). *Commun. ACM*, *43*(3), 32–34. Retrieved from http://doi.acm.org/10.1145/330534.330535 doi: 10.1145/330534.330535

Van Deemter, K., Krahmer, E., & Theune, M. (2005, March). Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, *31*(1), 15–24. Retrieved from http://dx.doi.org/10.1162/0891201053630291 doi: 10.1162/0891201053630291

Viethen, J., Dale, R., Krahmer, E., Theune, M., & Touset, P. (2008, May). Controlling redundancy in referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco: European Language Resources Association (ELRA). (http://www.lrec-conf.org/proceedings/lrec2008/)

Villalba, M., Teichmann, C., & Koller, A. (2017). Generating contrastive referring expressions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 678–687). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/P17-1063` doi: 10.18653/v1/P17-1063

Yantis, S., & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 121–134.

Zarrieß, S., & Schlangen, D. (2016). Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016).*