Form and Meaning in Dialog-Based Computer-Assisted Language Learning

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

vorgelegt von
Sabrina Wilske
aus Belzig

Saarbrücken, 2015

Dekan der Philosophischen Fakultät II:	Prof. Dr. Dietrich Klakow
Berichterstatter:	Prof. Dr. Manfred Pinkal Prof. Dr. Detmar Meurers Prof. Dr. Stefanie Haberzettl
Tag der letzten Prüfungsleistung:	12. Juni 2015

Für meine Tante – Dr. Helga Winkler (1939-2011)

Abstract

The goal of this thesis is to explore how foreign language learning can be facilitated through the use of intelligent computer-assisted language learning (ICALL) based on natural language processing (NLP) methods. ICALL was provided in the form of a task-based dialog system that gives corrective feedback. We investigated how different parameters of the interaction affect the learning progress. Based on a comprehensive review of existing comparable ICALL applications and the underlying methods and technology, we selected parameters linked to the sophistication and effort required to implement a particular form of interaction and related them to parameters that are based on two much debated issues from the field of second language acquisition (SLA). One is the debate that pits *form* against *meaning* and leads to a discussion of the extent to which language instruction should focus on linguistic forms and formal correctness as opposed to emphasizing communicative skills and the ability to use the language to make meaning in the real world. Related to that is the second controversial issue which concerns the dichotomy between implicit and explicit knowledge, learning and instruction: How explicit or implicit should instruction be, how does the degree of explicitness affect the development of explicit and implicit knowledge, and how do these two types of knowledge contribute to language skills?

These two general issues are condensed into three different experimental conditions, that differ with regard to how much they constrain the learner input and how explicit the feedback is. More precisely, we compare strictly form-focused activities where the learner input is constrained to supply a grammatical target form with generally unconstrained participation in a meaning-oriented task-based dialog. For the latter, we further compare recast and metalinguistic feedback as implicit and explicit types of feedback respectively. The findings of this study indicate that there are small differences in the language skill development afforded by different types of computer-provided instruction. We find that constrained, explicit form-oriented instruction yields in general greater immediate learning gains, while the free, more implicit and meaning-oriented instruction yields more delayed effects. Similarly, comparing implicit recast feedback with explicit metalinguistic feedback we find that the immediate effects are on par but recast feedback leads to greater delayed effects. These differences interact considerably with other parameters of the experimental setting, in particular with the selected target structures. This suggests that the effectiveness of certain types of instruction is highly dependent on the particular content and goal of the instruction.

By using current SLA issues as motivation and guide to develop an ICALL system and an experimental framework this work contributes to the yet small field of existing research and development which integrates ICALL and SLA perspectives.

Kurzzusammenfassung

Das Ziel dieser Arbeit ist es zu untersuchen, wie Anwendungen für intelligentes computer-unterstütztes Sprachenlernen (intelligent computer-assisted language learning – ICALL), welche auch Techniken der natürliche Sprachverarbeitung benutzen, das Erlernen von Fremdsprachen unterstützen können. ICALL wird in dieser Arbeit als aufgaben-basiertes Dialogsystem realisiert, welches korrigierendes Feedback gibt.

Ausgehend von einer eingehenden Analyse bestehender vergleichbarer ICALL-Systeme und den Methoden und Technologien, die ihnen zugrunde liegen, sowie aktuellen Fragestellungen in der Zweitspracherwerbsforschung (second language acquisition – SLA), untersuchen wir, wie sich verschiedene Interaktionsparameter auf die Lernergebnisse auswirken. Dazu wählen wir einerseits Parameter, die verbunden sind mit dem Aufwand, der für die Realisierung einer bestimmten Interaktionsform nötig ist. Diese setzen wir in Beziehung mit Parametern, die sich aus zwei umstrittenen Fragen in der Spracherwerbsforschung ergeben.

In der ersten dieser Fragen geht es um die jeweilige Rolle von Form und Bedeutung von Sprache und ob Sprachunterricht eher auf die korrekte Beherrschung von sprachlichen Strukturen oder eher auf kommunikative Fähigkeiten Wert legen sollte. Im Zusammenhang dazu steht die zweite Streitfrage, in der es um den Gegensatz zwischen impliziten und explizitem Wissen bzw. Lernen geht. Hier wird diskutiert, wie explizit oder implizit Unterricht sein soll, wie der Grad an Explizitheit sich auf explizites und implizites Wissen auswirkt und wie welches Wissen zu sprachlichen Fähigkeiten beiträgt.

Diese beiden generellen Fragestellungen sind in drei verschiedenen Experimentbedingungen zusammengefasst, die sich unterscheiden darin wie sehr sie die Eingaben der Lernenden einschränken und wie explizit das Feedback ist. Genauer gesagt vergleichen wir strikt form-fokussierte Übungen, in denen die Lernenden lediglich eine grammatische Zielform eingeben sollen mit offenen Konversationsübungen, in denen die Lernenden alle sprachlichen Mittel frei stehen, um eine praktische Aufgabe zu lösen. Die Verwendung der Zielform wird soll hierbei von der Aufgabe provoziert werden. Für die offene Bedingung vergleichen wir ferner Recast und metalinguistisches Feedback als implizite beziehungsweise explizite Formen von Feedback.

Die Ergebnisse der Untersuchung zeigen, dass die unterschiedlichen Übungsbedingungen zu kleinen Unterschieden in der Entwicklung von Sprachkenntnissen führen. Wir stellen fest, dass eingeschränkte, explizite form-orientierte Übungen zu größeren kurzfristigen Lernfortschritten führen, während freie, implizitere und bedeutungsorientierte Übungen zu größeren langfristigen Fortschritten führen. Im Vergleich von implizitem Recast-Feedback und explizitem metalinguistischem Feedback finden wir, dass die kurzfristigen Lernerfolge ähnlich sind, aber Recasts zu länger anhaltenden

Fortschritten führen. Diese Unterschiede interagieren allerdings mit anderen Parametern des Experiments, insbesondere mit den zu erlernenden grammatischen Zielstrukturen. Daraus folgern wir, dass die Wirksamkeit bestimmter Instruktionen stark vom spezifischem Inhalt und Ziel der Vermittlung abhängt.

Diese Arbeit trägt dazu bei, die bisher nur schwach vertretenden Bezüge zwischen SLA und ICALL zu stärken, indem sie aktuelle SLA-Forschungsfragen als Motivation und Richtlinie für die Entwicklung eines ICALL-Systems und eines dazugehörigen Rahmens für die experimentelle Untersuchung verwendet.

Zusammenfassung

Diese Dissertation untersucht wie Anwendungen für intelligentes computer-unterstütztes Sprachenlernen (intelligent computer-assisted language learning – ICALL), welche auch Techniken der natürliche Sprachverarbeitung benutzen, das Erlernen von Fremdsprachen unterstützen können. ICALL wird in dieser Arbeit als aufgaben-basiertes Dialogsystem realisiert, welches korrigierendes Feedback gibt. Der genaue Untersuchungsgegenstand und das Vorgehen ergeben sich aus konkreten Fragen im Bereich der Zweitspracherwerbsforschung (second language acquisition – SLA) mit Berücksichtigung des derzeitigen Standes der Technik und damit verbundenen Parametern für Interaktionsmöglichkeiten.

Die Untersuchungsparameter für den sprach-pädagogischen Bereich sind einerseits der Gegensatz zwischen Form und Bedeutung und andererseits zwischen implizitem und explizitem Lehren, Lernen und Wissen. Abbildung 1 illustriert den daraus resultierenden Parameterraum. Auf der vertikalen Achse ist der Gegensatz abgebil-

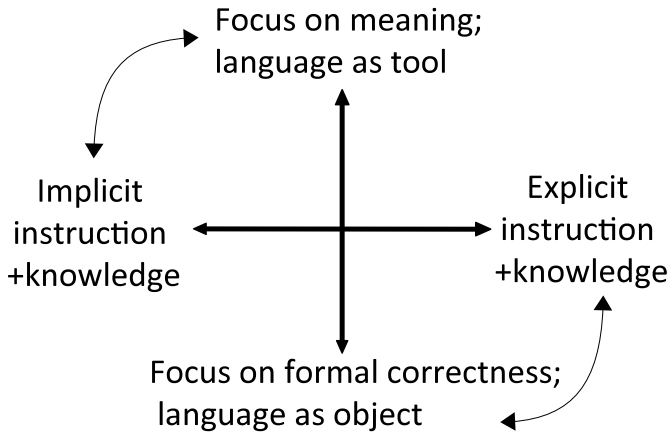


Abbildung 1 – Parameter aus dem Gesichtspunkt der Spracherwerbsforschung

det zwischen Unterricht, der auf Bedeutung und sprachliche Handlungsfähigkeit Wert legt und Unterricht, der auf formale Korrektheit abzielt. Während ersterer Sprache als Werkzeug betrachtet, welches dazu dient, Ziele in der realen Welt zu verwirklichen, betrachtet letzterer Sprache als Objekt, das gelernt werden soll.

Die horizontale Achse stellt den Bereich dar zwischen impliziten und expliziten Formen von Unterricht und den damit korrespondierenden Kenntnissen, die aus dieser Vermittlung entstehen. Zwischen den beiden Achsen bestehen Zusammenhänge. Form-basierte Vermittlungsansätze rücken die Formen und grammatischen Regeln oft auf eine explizite Art und Weise in den Vordergrund. Bedeutungsorientierter Unter-

richt auf der anderen Seite ist of impliziter, weil bestimmte Merkmale der Sprache nicht explizit in den Fokus gerückt werden. In der Abbildung ist diese Nähe durch die beiden gekrümmten Pfeile gekennzeichnet.

Im Zusammenhang mit den pädagogischen Dimensionen betrachten wir auch zwei Aspekte, die mit der Entwicklung und Realisierung von Systemen zur Mensch-Computer-Interaktion zu tun haben. Im Allgemeinen lässt sich feststellen, dass es mit dem derzeitigen Stand der Technik in der Computerlinguistik nicht möglich ist, unbeschränkte Äußerungen zuverlässig korrekt und vollständig zu analysieren. Dies führt dazu, dass ICALL-Anwendungen meist abwägen zwischen einerseits möglichst freier Eingabe für die Lernenden und andererseits genauer Analyse der Eingaben, um möglichst informatives und genaues Feedback geben zu können. Diesen beiden Parametern beschreiben nun einen Raum, der sich weiterhin danach charakterisieren lässt, mit wie viel Aufwand die Entwicklung verbunden ist. Abbildung 2 stellt diesen Parameterraum dar.

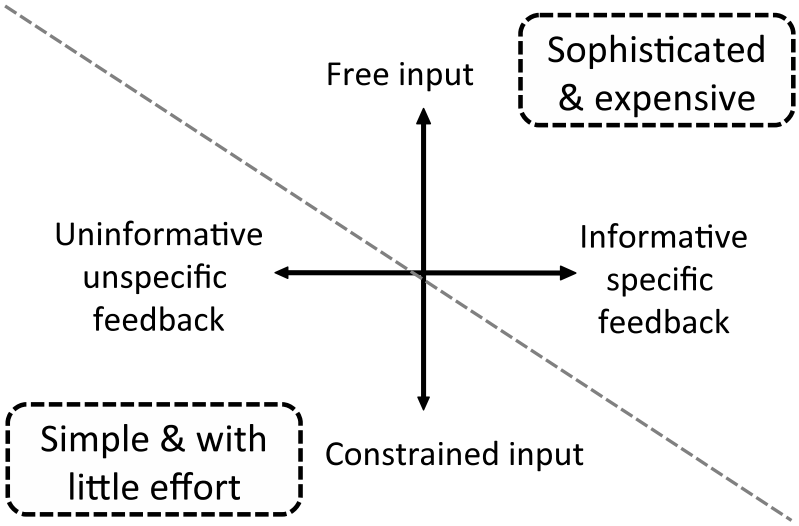


Abbildung 2 – Parameter aus dem Gesichtspunkt des Implementierungsaufwandes

Im Allgemeinen lässt sich sagen, dass Systeme, die eine größere Breite und Vielfalt an möglichen Lerner-Eingaben zulassen und entsprechend angemessen reagieren sollen, größeren Entwicklungsaufwand erfordern. Je informativer das Feedback vom System ist, desto mehr entsprechende Informationen müssen für das System modelliert werden. Der Aufwand für ein System, welches nur sehr eingeschränkte Eingaben zulässt und wenig informatives Feedback gibt ist dementsprechend geringer. Ausgehend von diesen Parameterräumen untersucht die vorliegende Arbeit die Wirksamkeit von bestimmten Interaktionsformen im Bereich des computergestützten Sprachlernens.

Im folgenden geben wir einen Überblick über die Inhalte der einzelnen Kapitel dieser Arbeit. Die Arbeit gliedert sich in elf Kapitel. Nach dem einleitenden Kapitel, welches wir gerade zusammengefasst haben, geben die Kapitel 2 bis 5 einen Einblick in die für diese Arbeit relevanten Disziplinen und deren Fragestellungen und bisherige

Forschungs- und Entwicklungsergebnisse. Dabei bietet Kapitel 2 einen Einblick in das Gebiet computer-gestütztes Sprachenlernen und seine generellen Herausforderungen. Kapitel 3 beschreibt die Grundlagen der Dialogmodellierung in natürlichsprachlichen Systemen und fokussiert sich damit auf ein Teilgebiet der existierenden Sprachlernanwendungen. Kapitel 4 beschreibt die zugrundeliegenden Forschungsergebnisse und Fragen aus dem Gebiet der Fremdspracherwerbsforschung. Kapitel 5 verbindet die beiden vorher beschriebenen Gebiete durch die Konzentration auf Feedback als integralen Bestandteil von Dialog und wichtiges Mittel für Sprachlernprozesse. Kapitel 6 bis 8 beschreiben dann die Einzelheiten unserer Untersuchung, wobei Kapitel 6 die gewählte Vorgehensweise im Allgemeinen darlegt und begründet und Kapitel 7 dann die Details des Experiments beschreibt. Kapitel 8 beschreibt Einzelheiten des implementierten Systems. Kapitel 9 stellt die Ergebnisse dar, die dann in Kapitel 10 diskutiert werden. Kapitel 11 zieht ein Fazit.

Kapitel 2 stellt die Problemstellungen von computer-gestütztem Sprachenlernen (computer-assisted language learning CALL) vor. Zu den Funktionen von CALL gehört es, Gelegenheit für Kommunikation zu bieten und dabei auch Rückmeldung an die Lernenden zu geben (Zhao, 2003). Ein Unterbereich von CALL ist intelligentes CALL (ICALL), welches benannt ist nach den Methoden, die zur Realisierung verwendet werden. Obwohl der Begriff nicht eindeutig abgegrenzt ist und unter *intelligent* manchmal alles verstanden wird, was mit erhöhtem Aufwand verbunden ist oder Methoden aus dem Bereich der künstlichen Intelligenz verwendet, ist ein weithin angenommenes Merkmal von ICALL, dass Methoden der natürlichen Sprachverarbeitung (NLP) und Computerlinguistik benutzt werden. Ein Beweggrund dafür, auf linguistische Repräsentationen und Modelle zurückzugreifen ist, dass sie effizientere Mittel bieten, um größere Mengen von erwarteten Lerner-Äußerungen darzustellen als dies eine extensive, aufzählende Darstellungsweise vermag (Nagata, 2009; Meurers, 2012). Eine der größten Herausforderungen von natürlicher Sprachverarbeitung ist die weitverbreitete Ambiguität (Mehrdeutigkeit) von sprachlichen Äußerungen. In fehlerhafter Lernersprache kommt zusätzlich die Unsicherheit über Fehlerursachen hinzu und damit die Schwierigkeit zu ermitteln, welches die beabsichtigte Äußerung war und welche Fehlvorstellung oder Wissenslücke zum Fehler führte. Eine Methode, Lernersprache angesichts der existierenden Schwierigkeiten handhabbar zu machen, ist es, die Eingabe der Lernenden auf verschiedene Arten einzuschränken. Hierbei ist es erstrebenswert Einschränkungen möglichst so zu realisieren, dass die Lernenden sich nicht übermäßig eingeengt fühlen und die Übungen trotzdem noch zur Förderung von möglichst freien Ausdrucksfähigkeit beitragen. Kapitel 2 stellt weiterhin eine Taxonomie für die verschiedenen Ansätze zur Fehler-Diagnose von Lernersprache vor, wobei nach Meurers (2012) auf oberster Ebene zwischen Lizenzierung und Musterabgleich unterschieden wird. Weiterhin lassen sich solche Ansätze danach charakterisieren, ob sie auf einer Erwartung basieren, auf welcher Ausschnittsgröße der Äußerung sie operieren und ob sie eine Korrektur bzw. Erklärung des Fehlers anbieten können. Kapitel 2 schließt ab mit einer Diskussion der Eigenschaften und Nutzen von computer-vermittelter Kommunikation in geschriebenen Sofortnachrichten (synchronous text chat).

Kapitel 3 gibt zum einen eine Einführung in die Grundlagen der Modellierung von natürlichem Dialog und präsentiert zum anderen die relevantesten Beispiele für inter-

aktive ICALL-Systeme, die korrektives Feedback an die Lernenden geben. Im ersten Teil beschreiben wir die wichtigsten Phänomene von natürlichsprachlichen Unterhaltungen. Wir legen dar, dass es für die Analyse von Dialog essentiell ist, diesen als kollaboratives Handeln zu betrachten. Damit lässt sich erklären, wie sich Sprecher mit ihren Redebeiträgen abwechseln und dabei ständig bemüht sind, das gegenseitige Verständnis zu sichern (Clark, 1996). Es erklärt auch, wie man Redebeiträge als Sprachhandlungen des Sprechers charakterisieren kann. Der erste Teil gibt weiterhin eine Einführung in die Grundlagen von Dialogsystemen, welche Anwendungskontexte und wesentliche Architekturmerkmale umfassen. Weiterhin enthält er eine Beschreibung der wesentlichen Komponenten eines Dialogsystems und ihrer Funktionen. Der Teil endet mit einer Diskussion der gebräuchlichsten Modellierungsansätze für Dialoge, welche endliche Automaten, Frames (bzw. Attribut-Merkmals-Strukturen), Informationszustände (information state), agenten- bzw. plan-basierte und statistische Ansätze umfasst.

Im zweiten Teil des dritten Kapitels werden existierende Beispiele für ICALL-Systeme vorgestellt und grundsätzlich unterschieden zwischen solchen, die vor allem versuchen, grammatische Formen zu vermitteln und solchen, die hauptsächlich für kommunikativen Austausch dienen sollen. Die vorgestellten Beispiele werden charakterisiert nach der erwarteten Eingabe von den Lernenden und wie diese eingeschränkt wird, nach der Fehlerdiagnose und den Rückmeldungen dazu, nach der Art und Weise, wie diese Ansätze evaluiert wurden und auf welchen pädagogischen Theorien die Systeme fußen. Es wird klar, dass nur wenige der vorgestellten Systeme bezüglich ihrer Lernfortschritte evaluiert wurden, was im Kontrast zu dem in dieser Studie verfolgten Ansatz steht.

Kapitel 4 erläutert die grundlegenden Konzepte aus der Zweitspracherwerbsforschung, die für die vorliegende Studie relevant sind. Zum einen erläutern wir verschiedene pädagogische Ansätze, die sich darin unterscheiden, wie viel Bedeutung sie grammatischer Korrektheit (Formen) einerseits und kommunikativer Handelsfähigkeit (Bedeutung) andererseits beimessen. Hierbei unterscheidet man zwischen FOCUS-ON-FORMS, FOCUS-ON-MEANING und FOCUS-ON-FORM, wobei das letztere eingeführt wurde, um die Vorteile der beiden ersten zu vereinen und ihre Unzulänglichkeiten auszugleichen. Der FOCUS-ON-FORM-Ansatz zeichnet sich dadurch aus, dass er Formen erst dann in den Vordergrund rückt, wenn sie aus einem bedeutungsorientierten Kontext heraus relevant werden. Die Vagheit dieser Definition führte allerdings zu verschiedenen praktischen Implementierungen, die sich u.a. darin unterscheiden, inwieweit sie geplant und proaktiv oder spontan und reaktiv sind. Weitere Unterschiede bestehen darin wie genau die Verbindung zwischen Bedeutung und Formen gestaltet wird.

Im zweiten Teil des vierten Kapitels diskutieren wir die Dichotomie von explizitem und implizitem Lernen, Lehren und Wissen. Wir stellen bisherige Erkenntnisse über explizite und implizite Lernvorgänge und die Wirksamkeit von beiden Arten von Unterricht vor. Weiterhin definieren wir implizites und explizites Wissen und geben die verschiedenen Positionen der Interface-Debatte wieder, in der diskutiert wird, inwiefern beiden Arten von Wissen zusammenhängen und ob sie ineinander übergehen können. Dieser Teil endet mit einer Präsentation von verschiedenen Mitteln zur Messung beider Wissensarten, auf die wir in unserem Experiment zurückgreifen werden.

Im dritten Teil des vierten Kapitels diskutieren wir wichtige Merkmale von linguistischen Formen, die das Lernziel darstellen (sogenannten target forms oder Zielformen). So werden in der Literatur Salienz, Frequenz, Regularität, funktionaler Wert und Verarbeitbarkeit von Formen als wichtige Einflussfaktoren für ihre Lernbarkeit diskutiert. In diesem Zusammenhang diskutieren wir auch das Konzept von Entwicklungsstufen für bestimmte Formen, für die gezeigt wurde, dass Lernende sie regelmäßig durchlaufen. Dies wirkt sich auf die Reihenfolge aus, in der Formen sinnvollerweise unterrichtet werden sollten.

Das Kapitel endet mit einer Diskussion von *conversational interaction* (Interaktion im Gespräch) und aufgaben-basiertem Unterricht, welche beide im engen Zusammenhang mit dem FOCUS-ON-FORM-Ansatz stehen. Wir stellen dar, durch welche Prozesse Lernende von Kommunikation profitieren. Aufgaben sind ein Mittel des FOCUS-ON-FORM-Ansatzes, um bedeutungsvollen Kontext zu schaffen und Gelegenheiten zum Sprechenüben in Situationen zu schaffen, die dem späteren realen Anwendungskontext ähnlich sind. Wir diskutieren weiterhin das Konzept von zielgerichteten Aufgaben (*focused tasks*), die darauf ausgerichtet sind, bestimmte sprachliche Strukturen auf einem möglichst natürlichen und ungezwungenen Weg zu elizitieren.

Ein zentrales Element für den pädagogischen Nutzen von Kommunikation und Interaktion ist die Rückmeldung (Feedback), welche die Lernenden erhalten. Im **Kapitel 5** diskutieren wir die Bedeutung und Wirkungsweisen von Feedback näher. Dazu fassen wir Diskussionen über die Notwendigkeit, die Wirksamkeit und möglichen Nachteile von Feedback zusammen. Weiterhin verbinden wir dann die beiden Bereiche Fremdspracherwerb und ICALL indem wir die verschiedenen Arten von Feedback, die im Unterricht und in der Kommunikation mit nicht muttersprachlichen Sprechern vorkommen, klassifizieren und sie in Beziehung setzen zu den technologischen Anforderungen, die zur Bereitstellung solcher Feedback-Arten in einem ICALL-System nötig sind. Kriterien, anhand derer wir Feedback klassifizieren sind der Grad der Explizitheit, ob Lernende zu einer Berichtigung aufgefordert werden und der Informationsgehalt des Feedbacks. Das Kapitel endet mit einer detaillierten Besprechung von existierenden Arbeiten über zwei bestimmte Typen von Feedback, die wir auch in unserer Arbeit näher vergleichen: Recasts und metalinguistisches Feedback.

Kapitel 6 erklärt den Ansatz, den wir mit unserer Studie verfolgen um das Potenzial von NLP-basierten ICALL-Anwendungen zu ermessen. Zentral in unserem Ansatz ist es, uns auf eine kleine Anzahl von Bedingungen zu fokussieren und diese mit Hilfe von in der Zweispracherwerbsforschung üblichen Methoden zu untersuchen. Die Auswahl der untersuchten Bedingungen basiert auf zwei Blickwinkeln - einem technologischen und einem pädagogischen. In jedem Blickwinkel kommen zwei Parameterräume zum Tragen. In der pädagogischen, von der Zweitspracherwerbsforschung geprägten Perspektive spielen (a) das Kontinuum zwischen impliziter und expliziter Unterweisung und (b) die Bandbreite zwischen Form und Bedeutung eine Rolle. In der technologischen Perspektive ergeben sich die Parameter aus (c) dem Informationsgehalt von Feedback und (d) der Freiheit für die Lernenden sich zu äußern. Diese vier Parameter spannen einen mehrdimensionalen Raum auf. Wir begründen die Wahl der Parameter und diskutieren auch mögliche Alternativen. Die Auswahl der Parameter ist vom derzeitig verfügbaren Stand der Technik bestimmt. Da es im Moment

noch nicht möglich ist, eine weitgehend freie Eingabe bei gleichzeitig zuverlässiger und tiefgehender Analyse und Interpretation dieser Eingabe zum Zweck von aussagekräftigen Feedbacks anzubieten, müssen Dialogsysteme in mindestens einer dieser Dimensionen Einschränkungen haben. Bei den untersuchten Systemen zeigt sich demzufolge auch meist ein Abwägen zwischen den beiden Zielen mit dem Ergebnis, dass eines als wichtiger erachtet wird.

Unsere Studie ist an der Schnittstelle zwischen den drei Disziplinen Computerlinguistik/NLP, Fremdspracherwerb (SLA) und ICALL positioniert. Wir entwickeln ein neues ICALL-System, welches dazu dient SLA-Forschungsfragen zu beantworten und damit auch anwendungsorientierte Erkenntnisse schafft über die praktische Nutzung von CL/NLP als Forschungswerkzeug. Die SLA-Forschungsfragen, die wir stellen sind folgende:

> Gibt es einen Unterschied zwischen den Lern-Effekten von computer-basierter FOCUS-ON-FORM and FOCUS-ON-FORMS Übungen?
>
> Gibt es einen Unterschied zwischen der Wirksamkeit von Recasts und metalinguistischem Feedback, welche von einer ICALL-Anwendung dargeboten werden?

Am Ende von Kapitel 6 präsentieren und begründen wir die grundlegenden Parameter des Forschungsdesigns.

Kapitel 7 beschreibt die Einzelheiten des Experiments. Dazu legen wir zuerst die Auswahl der grammatischen Zielstrukturen dar, die sich vor allem auf drei Kriterien begründet: (a) ihre Eignung sich in einer bedeutungs-orientierten Aufgabe elizitieren zu lassen, (b) Anhaltspunkte, dass die Beherrschung den Lernenden Schwierigkeiten bereitet und (c) ihre Testbarkeit. Wir beschreiben dann die beiden Strukturen – Dativ-Präpositionalphrasen und Nebensätze und diskutieren dabei auch Merkmale, die ihre Lernbarkeit bestimmen. Im zweiten Teil dieses Kapitels beschreiben wir die zielgerichteten Aufgaben innerhalb derer die Zielstrukturen verwendet werden sollen. Eine Wegbeschreibungsaufgabe anhand einer vereinfachten Landkarte soll die Verwendung von Dativ-Präpositionalphrasen anregen, mit denen auf Orientierungspunkte verwiesen wird, die Teilziele sind oder an denen die Richtung geändert werden soll. Einen Termin zu vereinbaren ist die Aufgabe, bei der Nebensätze verwendet werden sollen, und zwar hauptsächlich kausale, mit denen Absagen bzw. Verhinderungen begründet werden. Hier beschreiben wir auch die Interaktion zwischen den Lernenden und dem System und die Strategien des Systems, Feedback zu geben und den Lernenden die Zielstrukturen zu entlocken. Für jede der beiden Aufgaben beschreiben wir das zugrunde liegende Dialogmodell bzw. die möglichen Eingaben der Lernenden und die Ausgaben des Systems für jede der drei unterschiedlichen Bedingungen, unter denen die Versuchspersonen interagieren. Diese drei Bedingungen sind ausgehend von den Forschungsfragen (a) eingeschränkte Eingabe mit Orientierung auf sprachliche Formen, (b) freie Eingabe mit implizitem Recast-Feedback und (c) freie Eingabe mit metalinguistischem Feedback.

Im dritten Teil des siebten Kapitels präsentieren wir die Tests mit denen wir die Lernfortschritte messen. Wir legen dar, dass es für eine umfassende Messung wichtig

ist, sowohl implizite als auch explizite Kenntnisse zu erfassen. Dann begründen wir die Wahl eines Grammatikalitätsurteilstest mit Zeitlimit als Maß für implizites Wissen damit, dass es Hinweise darauf gibt, dass der Zeitdruck den langsameren Zugriff auf metalinguistisches explizites Wissen behindert und die Lernenden dazu zwingt, auf ihr schneller erreichbares implizites Wissen zurückzugreifen. Für die Messung von explizitem Wissen benutzen wir eine Satzbildungsaufgabe, weil diese die Aufmerksamkeit auf die sprachlichen Formen lenkt und den Lernenden genügend Zeit lässt, auf ihre expliziten Wissensstrukturen zuzugreifen. Für beide Tests präsentieren wir die einzelnen Testfragen. Zusätzlich zu diesen Form-orientierten Tests möchten wir auch die Entwicklung der spontan-sprachlichen Fähigkeiten messen, da aufgaben-basierter FOCUS-ON-FORM-Unterricht gerade für die Förderung solch Kontext-gebundener Anwendung von Sprachkenntnissen in Echtzeit gepriesen wird. Dazu lassen wir die Lernenden paarweise mündliche Konversationsaufgaben erfüllen, die den obengenannten Aufgaben sehr ähnlich sind und ähnliche Materialien benutzen. Die daraus entstandenen Gespräche werden aufgezeichnet und ihre Flüssigkeit untersucht. Dazu verwenden wir zwei komplementäre Messinstrumente – einerseits bewerten Lehrkräfte für Deutsch als Fremdsprache die Flüssigkeit der einzelnen Sprecher, andererseits extrahieren wir eine Reihe von messbaren temporalen Eigenschaften durch Annotation der Sprachdaten. Bedingt durch den deutlich erhöhten Aufwand für diese Art von Datenaufbereitung, wurde hierfür nur eine Teilmenge der Daten in Betracht gezogen, insbesondere wurden die Daten für die Gruppe mit metalinguistischem Feedback nicht berücksichtigt. Wir beenden das Kapitel mit einer Beschreibung des zeitlichen Ablaufs und der Details der Datensammlung sowie der Lernenden, die als Versuchspersonen für diese Studie dienten. Die Studie umfasste für jeden Teilnehmer drei Termine, wobei die ersten beiden im Abstand von einer Woche stattfanden und jeweils eine Interaktion mit dem System sowie Vor- und Nachtests beinhalteten. Der dritte Termin, der im Normalfall fünf Wochen nach dem ersten stattfand, diente allein der Durchführung der nachgelagerten Nachtests. Der Großteil der Lernenden nahm im Rahmen eines semesterbegleitenden Deutsch-Kurses für Austauschstudierende an der Studie teil, wobei die Lernenden die Aufgaben und Tests individuell am Computer absolvierten.

Kapitel 8 beschreibt weitere Implementierungsdetails des ICALL-Dialogsystems, mit dem wir die Studie durchführen. Dabei gehen wir insbesondere auf die Techniken für die Fehleranalyse ein und beschreiben außerdem die Leistung und Schwachpunkte des Systems. Weiterhin fassen wir die Bewertungen der Lernenden zusammen und zeigen, dass die Interaktion und das Arbeiten mit dem System als positiv bewertet wurde.

Kapitel 9 beschreibt die Ergebnisse unserer Studie im Detail. Generell stellen wir fest, dass über alle Experiment-Bedingungen hinweg mit den Übungen für Dativ-Präpositionalphrasen deutlichere Lernfortschritte erzielt wurden als für Nebensätze. Für die Entwicklung der spontanen Sprachfertigkeiten zeigt sich eine komplementäre Entwicklung – die Gruppe, die in freier Interaktion Recast-Feedback bekommt, verbessert ihre Flüssigkeit in der Wegbeschreibungsaufgabe aber nicht in der Terminvereinbarungsaufgabe, während die Gruppe mit eingeschränkter Eingabe eine Steigerung für das Verabredungsszenario zeigt, aber nur eine sehr kleine Entwicklung für

die Wegbeschreibung. Weiterhin zeigt sich im Grammatikalitätsbewertungstest, dass grammatisch korrekte Testaufgaben insgesamt richtiger beurteilt werden als inkorrekte. Bei näherer Betrachtung der Kenntnisentwicklung für Nebensätze fallen zwei Tendenzen auf. Die Recast-Gruppe zeigt eine Verbesserung ihrer Kenntnisse erst im spätesten Nachtest fünf Wochen nach der ersten Sitzung. Für diesen Testzeitpunkt ist sie auch der Gruppe mit eingeschränkter Eingabe signifikant überlegen. Diese Gruppe zeigt nur eine Verbesserung zwischen dem Vortest und den beiden ersten Nachtests und dies auch nur für die inkorrekten Testaufgaben im Beurteilungstest. Für die spontansprachlichen Fähigkeiten für das Verabredungsszenario lässt sich feststellen, dass die Gruppe mit eingeschränkter Eingabe Verbesserungen für einige wenige Messpunkte zeigt, während die Recast-Gruppe keine Steigerung zeigt.

Für die Entwicklung von Kenntnissen über Dativ-Präpositionalphrasen gibt es drei Beobachtungen. Erstens schneidet die Gruppe mit eingeschränkter Eingabe in den unmittelbaren Nachtests besser ab als die Gruppe mit metalinguistischem Feedback bei freier Eingabe. Zweitens zeigt die Gruppe mit eingeschränkter Eingabe die meisten unmittelbaren Steigerungen im Vergleich zu den beiden Gruppen mit freier Eingabe. Drittens lässt sich feststellen, dass die Recast-Gruppe im Vergleich zu den beiden anderen Gruppen die größten langfristigen Verbesserungen zeigt, so wie sie mit dem letzten Nachtest fünf Wochen nach der ersten Sitzung gemessen wurden. Die Entwicklung der spontansprachlichen Fähigkeiten für die Wegbeschreibungsaufgabe ist durch zwei Merkmale gekennzeichnet. Zum einen zeigt die Recast-Gruppe mit freier Eingabe hier deutliche Verbesserungen im nachgelagerten Nachtest. Die Gruppe mit eingeschränkter Eingabe hingegen zeigt nur minimale Verbesserungen.

Fazit Kapitel 10 diskutiert die Ergebnisse und Kapitel 11 zieht ein Fazit, welches hier in Auszügen wiedergegeben wird. Die Erkenntnisse aus unserer Studie zeigen, dass verschiedene Arten von computer-basierten Übungen zu unterschiedlichen Lernfortschritten führen können. Zusammenfassend lässt sich feststellen, dass explizite FOCUS-ON-FORMS-Übungen generell zu größeren kurzfristigen Verbesserungen führen, während FOCUS-ON-FORM Übungen mit freier Eingabe und eher längerfristige Verbesserungen erzielen. Im direkten Vergleich von implizitem Recast-Feedback mit explizitem metalinguistischem Feedback finden wir, dass die unmittelbaren Fortschritte vergleichbar sind, aber Recasts zu größeren längerfristigen Effekten führen.

Diese Unterschiede sind allerdings stark abhängig von anderen Experimentparametern, insbesondere von den Zielstrukturen. Grammatische Formen unterscheiden sich dahingehend wie leicht sie in einer natürlichen, realitätsrelevanten Aufgabe zu elizitieren sind. Daraus folgt, dass die Wirksamkeit von bestimmten Unterrichts- bzw. Übungsansätzen stark vom jeweiligen Ziel der Übung abhängen kann. Damit bestätigen wir auch die Feststellung, dass die Wirksamkeit zielgerichteter Aufgaben ihre Grenzen findet in der Eigenschaft der Zielstrukturen für die Erfüllung der Aufgabe natürlich, nützlich oder unerlässlich zu sein. Darüber hinaus ist die Gestaltung von Aufgaben selbst für unerlässliche Zielstrukturen keineswegs ein simpler oder klar definierter Prozess, sondern erfordert gewisse Fähigkeiten und Erfahrung. Aus dieser Einschränkung folgern wir, dass es notwendig ist, den aufgaben-basierten Unterrichtsansatz mit anderen Methoden zu kombinieren.

Unsere Ergebnisse stimmen größtenteils überein mit den Ergebnissen von bisherigen Forschungsarbeiten, die sich auf die Mensch-Mensch-Interaktion im Klassenraum beziehen und den Unterschied zwischen impliziten und expliziten Lehrmethoden untersuchen oder spezieller auch die Unterschiede zwischen Recast und metalinguistischem Feedback. Dies stimmt auch überein mit den Befunden von Petersen (2010), der feststellte, dass Recasts in schriftlicher unmittelbarer Kommunikation zwischen Computer und Mensch genauso effektiv waren wie Recasts, die in mündlicher Lehrer-Schüler-Interaktion auftraten. Beide diese Übereinstimmungen weisen darauf hin, dass die Unterschiede zwischen Mensch-Computer-Interaktion und rein zwischenmenschlicher Interaktion zumindest unter einigen Bedingungen nicht so groß sind, dass sie zu grundlegend anderen Lernbedingungen und Lernergebnissen führen. Dies gibt Anlass zu der Annahme, dass wir auch andere hinreichend ähnliche Erkenntnisse aus der Mensch-Mensch-Kommunikation in Mensch-Computer-Kommunikation überführen können und mit ähnlichen Lernergebnissen rechnen können.

In Anbetracht der Tatsache, dass die Konversationsfähigkeiten eines künstlichen Systems in der Regel den menschlichen noch unterlegen, sind müssen wir eine solche Übertragbarkeit jedoch einschränken auf jene Aufgaben und Kommunikationsbedingungen, die ein künstliches System realistischerweise emulieren kann. Es ist daher wichtig und erstrebenswert, eben jene Bedingungen und Grenzen zu finden, innerhalb derer Sprachlernen durch ein intelligentes und unterhaltsames, wenn auch nicht dem Menschen ebenbürtiges System erleichtert werden kann.

Die besseren langfristigen Effekte von bedeutungsorientierten impliziten Aufgaben mit freier Eingabe können die erhöhten Entwicklungsausgaben für die Realisierung von Systemen, die solche Aufgaben und entsprechende Interaktionsformen anbieten, rechtfertigen. Während unsere Ergebnisse zeigen, dass simplere Ansätze, die nur beschränkte Eingabemöglichkeiten bieten, zwar kurzfristig zu deutlicheren Verbesserungen führen, zeigt sich auch, dass diese kurzfristigen Effekte nicht länger als einige Wochen anhalten und die Effekte daher nicht nachhaltig sind.

Nichtsdestotrotz sollte man aus unseren Ergebnissen nicht ableiten, dass simple drillartige Grammatikübungen überhaupt keinen Wert haben. Diese in einen übergeordneten bedeutungsorientierten Zusammenhang einzubetten, anstatt sie als Batterie von dekontextualisierten Aufgaben, die abzuarbeiten sind, den Lernenden zu präsentieren, kann solche Übungen unterhaltsamer und attraktiver machen. Dafür sprechen auch Bewertungen der Lernenden in unserer Studie. Dort wurde die Systemvariante mit beschränkter Eingabe, in der man eine Lücke füllen oder Wörter in die richtige Reihenfolge für einen Satz bringen musste und diese Sätze dann Teil eines vorgefertigten, sich sukzessive entfaltenden Dialogs wurden, nicht schlechter beurteilt als die Systemvarianten mit freier Eingabe, in der die Lernenden ihren Beitrag zum Dialog allein gestalten mussten. Bewertungskriterien waren die Freude mit der das System benutzt wurde, die empfundene Nützlichkeit und die Wahrscheinlichkeit mit der die Lernenden das System in Zukunft noch einmal benutzen würden, wenn sie dazu Gelegenheit hätten. Mögliche Ursachen für die Gleichheit der Bewertungen können darin liegen, dass wir den Kontext bewusst identisch gehalten haben und dass das Systems mit freier Eingabe noch einige Fehler produzierte, welche möglicherweise zu Unzufriedenheit führten.

Die positiven Ergebnisse für alle drei unterschiedlichen Bedingungen von ICALL-Übungen stimmen überein mit den Beobachtungen von Grgurović et al. (2013), die in einer Meta-Analyse herausfanden, dass CALL-Anwendungen (die sowohl simple als auch intelligente Ansätze umfassten) immer mindestens genauso wirksam waren wie Unterweisungen ohne Computer und sogar effektiver in Studien, die durch sehr streng kontrollierte Designs gekennzeichnet waren.

Daraus schließen wir, dass sowohl simple als auch avancierte CALL-Anwendungen ihre Berechtigung haben und wirksame Unterstützung zum Sprachenlernen bieten können. Während aufwändige Systeme den menschlichen Fähigkeiten näher kommen und dadurch möglicherweise unterhaltsamer sein können und auch zu nachhaltigeren Lernfortschritten führen können, ist es notwendig, dass sie möglichst fehlerfrei funktionieren, welches beträchtlichen Entwicklungsaufwand erfordert. Daher behalten weniger aufwändige, mit weniger Entwicklungskosten verbundene Anwendungen durchaus ihre Berechtigung wenn man eine Kosten-Nutzen-Analyse vornimmt.

Auch im Gebiet der aufwändigeren Ansätze, die wie in unserem Beispiel eine freie Eingabe erlauben und Feedback geben, gibt es verschiedene Abstufungen von Ausgereiftheit, welche bewusst eingesetzt werden sollten. Für die Recasts in unserem System ist eine perfekte Sicherheit in der Fehlererkennung nicht unbedingt nötig, weil Recasts als Reaktion auf eine fehlerfreie Äußerung des Lerners keine nachteiligen Effekte haben müssen. Das begründet sich damit, dass sie als auch als einfache Bestätigungen oder Umformulierungen interpretiert werden können ohne zu fälschlich zu behaupten, dass die Äußerung des Lerners fehlerhaft war und diesen damit zu verwirren oder falsche Informationen zu geben. Metalinguistisches Feedback oder andere explizite Arten von Feedback hingegen können schädlich sein, wenn sie fälschlicherweise als Reaktion auf korrekte Lerneräußerungen ausgegeben werden. Daher sollte die Art und Weise der Interaktion, in diesem Fall die Art des Feedback, abhängig von der Zuverlässigkeit der Fehler-Diagnose gewählt werden, um negative Effekte für die Lernenden zu vermeiden.

Danksagung

Viele Menschen haben mich bei der Erstellung dieser Arbeit unterstützt. Ich möchte mich hiermit bei ihnen bedanken. Manfred Pinkal und Detmar Meurers haben die Arbeit betreut und mir hilfreichen Rat, Zuspruch, aber auch Freiheiten bei der Umsetzung gewährt. Magdalena Wolska gilt mein besonderer Dank, ohne ihr tiefgründiges und anhaltendes Interesse an meiner Arbeit und die daraus resultierende konkrete Zusammenarbeit und die gemeinsamen Publikationen wäre diese Arbeit nicht entstanden.

Ich danke den Koordinatorinnen des DaF-Programms an der Universität des Saarlandes Kristin Stezano und Meike van Hoorn sowie den anderen Lehrkräften, die mir Zugang zu meinen Versuchspersonen organisiert und gewährt haben. Ich danke den Lernenden der DaF-Kurse, die an meinen Versuchen teilgenommen haben und den Deutschlernern, die sich im Vorfeld bei der Entwicklung der Prototypen als Tester zur Verfügung stellten.

Die Mitglieder des Graduiertenkollegs Language Technology and Cognitive Systems haben mir in vielfältiger Weise geholfen, insbesondere Michael, Barbara, Rui, Mark und Verena. Christoph Clodo und die anderen Mitarbeiter der Systemgruppe haben alle mögliche technische Unterstützung geleistet, insbesondere verloren gegangene Daten wieder gefunden.

Ich danke meinen Kolleginnen und Kollegen von dimeb in Bremen, die Anteil genommen haben und mir Raum gaben, meine Arbeit fertig zu stellen.

Ich danke meiner Familie und meinen Freundinnen und Freunden für ihre vielfältige Hilfe und moralische Unterstützung, insbesondere John, Friederike und Eva. Sonja und Shaku haben mich über die Jahre am engsten begleitet und gestützt und immer wieder daran erinnert, worauf es ankommt, ich danke euch dafür.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Outline of the thesis . . . 5

2 Computer-Assisted Language Learning **9**
- 2.1 Introduction . . . 9
- 2.2 Natural language processing in ICALL . . . 11
  - 2.2.1 Motivation . . . 11
  - 2.2.2 Expectations and challenges . . . 11
  - 2.2.3 Constraining input as a strategy to deal with limits . . . 13
- 2.3 Error diagnosis . . . 14
  - 2.3.1 Language licensing . . . 15
  - 2.3.2 Pattern matching . . . 17
  - 2.3.3 Summary . . . 17
- 2.4 Computer-mediated communication . . . 18
  - 2.4.1 Properties of text chat language . . . 19
  - 2.4.2 Benefits of participating in text chat . . . 20
- 2.5 Summary . . . 21

3 Dialog for Language Learning **23**
- 3.1 Dialog . . . 23
  - 3.1.1 Dialog phenomena in human interaction . . . 24
  - 3.1.2 Dialog systems . . . 28
  - 3.1.3 Components . . . 31
  - 3.1.4 Approaches to dialog modeling and management . . . 36
- 3.2 State of the art in existing ICALL systems . . . 41
  - 3.2.1 Introduction . . . 41
  - 3.2.2 Systems with a focus on grammar . . . 43
  - 3.2.3 Systems with a focus on communication . . . 45
- 3.3 Summary . . . 52

4 Second Language Acquisition **55**
- 4.1 Introduction . . . 55
- 4.2 Form and meaning in language instruction . . . 56
  - 4.2.1 Focus on forms . . . 57
  - 4.2.2 Focus on meaning . . . 58
  - 4.2.3 Combining both approaches - focus on form . . . 59

4.2.4 Parameters of focus on form . . . 61
4.3 Implicit and explicit learning, instruction, and knowledge . . . 63
4.3.1 Implicit and explicit learning . . . 64
4.3.2 Implicit and explicit instruction . . . 65
4.3.3 Empirical evidence . . . 66
4.3.4 Implicit and explicit knowledge . . . 68
4.3.5 Interface debate . . . 70
4.3.6 Measures of explicit and implicit knowledge . . . 71
4.4 Properties of target structures . . . 73
4.4.1 Salience . . . 74
4.4.2 Frequency and regularity . . . 74
4.4.3 Functional value . . . 75
4.4.4 Developmental readiness and processability . . . 75
4.5 Conversational interaction and task-based instruction . . . 76
4.5.1 Conversational interaction . . . 76
4.5.2 Tasks . . . 79
4.5.3 Evaluating tasks and communicative interaction . . . 80
4.6 Summary . . . 82

5 Feedback 83
5.1 Introduction . . . 83
5.2 Necessity and benefit of feedback . . . 84
5.2.1 Necessity of corrective feedback . . . 84
5.2.2 Effectiveness, benefit and potential harm of corrective feedback . 85
5.2.3 Further issues in feedback research . . . 87
5.3 Classification of feedback . . . 87
5.3.1 Types of feedback . . . 88
5.3.2 Parameters of feedback . . . 90
5.4 Feedback in the ICALL context . . . 93
5.4.1 General issues in ICALL feedback . . . 93
5.4.2 Information requirements for different feedback types . . . 95
5.5 Recast and metalinguistic feedback . . . 97
5.5.1 Recasts . . . 98
5.5.2 Metalinguistic feedback . . . 102
5.5.3 Recasts versus metalinguistic feedback . . . 105
5.6 Summary . . . 110

6 The Approach 111
6.1 Introduction . . . 111
6.2 Implementing ICALL dialog and feedback . . . 112
6.2.1 Informativity of feedback . . . 114
6.2.2 Freedom of input . . . 116
6.3 Relating to the pedagogic perspective . . . 117
6.3.1 Explicit and implicit feedback . . . 118
6.3.2 Meaning, form, and freedom of input . . . 119
6.3.3 Relations between pedagogic and implementational parameters . 120

6.4 Alternative parameters . . . 121
6.4.1 Parameters related to learning . . . 121
6.4.2 Parameters related to dialog . . . 123
6.5 The context of this thesis . . . 124
6.6 Research design . . . 127
6.6.1 Research questions . . . 127
6.6.2 Methodological choices . . . 128
6.6.3 Parameters of the experimental treatment . . . 131
6.7 Summary . . . 133

7 The Experiment **135**
7.1 The target structures . . . 135
7.1.1 Dative case in prepositional phrases . . . 136
7.1.2 Word order in subordinate clauses . . . 142
7.2 Tasks and interaction . . . 145
7.2.1 Giving directions task . . . 146
7.2.2 Appointment task . . . 155
7.3 Assessment of linguistic development . . . 160
7.3.1 Implicit knowledge: timed grammaticality judgment test . . . 161
7.3.2 Explicit knowledge: sentence construction test . . . 163
7.3.3 Communicative skills . . . 165
7.4 Procedures . . . 168
7.5 Summary . . . 171

8 The System **173**
8.1 Design and implementation . . . 173
8.2 System performance . . . 179
8.2.1 Dative prepositional phrases . . . 179
8.2.2 Recasts for dative prepositional phrases . . . 180
8.2.3 Metalinguistic feedback for dative prepositional phrases . . . 182
8.2.4 Subordinate clauses . . . 183
8.3 Learners' perception and rating of the system . . . 185
8.4 Summary . . . 189

9 Findings **191**
9.1 Development of grammatical accuracy . . . 192
9.1.1 Word order in subordinate clauses . . . 193
9.1.2 Dative case in prepositional phrases . . . 197
9.2 Development of oral communicative skills . . . 207
9.2.1 Holistic rating of perceived fluency . . . 208
9.2.2 Temporal measures . . . 211
9.2.3 Summary of oral skills development . . . 220

10 Discussion **223**
10.1 Summary of findings . . . 223
10.2 Discussion of results . . . 228
10.2.1 Constrained instruction versus free input instruction . . . 229
10.2.2 Recasts versus metalinguistic feedback . . . 231
10.2.3 Differences between grammatical and ungrammatical items . . . 233
10.2.4 Differences between development for the two target structures . . . 234
10.2.5 Communicative skill development . . . 236
10.3 Limitations . . . 237

11 Concluding Remarks **243**
11.1 Summary of contributions . . . 243
11.2 Outlook . . . 245

Bibliography **247**

List of Figures

1 Parameter aus dem Gesichtspunkt der Spracherwerbsforschung IX
2 Parameter aus dem Gesichtspunkt des Implementierungsaufwandes . . X

1.1 Pedagogical aspects: Parameters in language instruction and learning . . 3
1.2 Implementational aspects: Dimensions for sophistication and computational effort of ICALL applications 4

2.1 Taxonomy for error detection and diagnosis 14

3.1 Architecture for dialog systems 29
3.2 Finite state automaton dialog model 36
3.3 Information state update dialog model 39

6.1 Implementational aspects: Dimensions for sophistication and computational effort of CALL applications 114
6.2 Pedagogical aspects: Parameters in language instruction and learning . . 118
6.3 Instances for experimental comparison 120

7.1 System Interface 145
7.2 Task material from giving directions: the map 147
7.3 Dialog model for directions scenario 152
7.4 System interface for directions, constrained interaction 153
7.5 Task Material for Making Appointments: The Agenda 154
7.6 Dialog model for appointments scenario 159
7.7 System interface for appointments, constrained interaction 160
7.8 Experiment timeline 168
7.9 Experiment sample groups 170

8.1 The system architecture 174
8.2 Simplified fragment of the interpretation grammar 177
8.3 Simplified fragment of the generation grammar 178
8.4 Ratings for enjoyment of system interaction 186
8.5 Ratings for perceived usefulness 186
8.6 Ratings for likelihood of future usage 186
8.7 Ratings for naturalness of the system interaction 187
8.8 Ratings for coherence and appropriateness of system's utterances 187
8.9 Ratings for comprehension skills of system 187

9.1 Box plots SubC SC test 194

9.2 Box plots SubC TGJT . . . 195
9.3 Box plots SubC TGJT, grammatical items . . . 196
9.4 Box plots SubC TGJT, ungrammatical items . . . 196
9.5 Box plots DatPP SC test . . . 199
9.6 Box plots DatPP TGJT . . . 201
9.7 Box plots DatPP TGJT, grammatical items . . . 201
9.8 Box plots DatPP TGJT, ungrammatical items . . . 202
9.9 Ratings appointment task, free-recast group . . . 209
9.10 Ratings appointment task, constrained group . . . 209
9.11 Ratings directions task, free-recast group . . . 210
9.12 Ratings directions task, constrained group . . . 211
9.13 Temporal measures appointment task . . . 214
9.14 Temporal measures directions task . . . 218

10.1 Significance Between Test times, grammatical items . . . 224

List of Tables

2.1 Error diagnosis techniques and their properties 18

3.1 Example for slots, questions, and response instances in a frame-based dialog manager 38

4.1 Implicit and explicit instruction 66

5.1 Feedback strategies and their properties 90

6.1 Information content of different types of feedback 115
6.2 Freedom of input and constraints, a rough characterization 117

7.1 Determiners and Cases in German NP 137
7.2 Participants breakdown 171

8.1 System performance directions task 180
8.2 Distribution of successful and failed recasts 181
8.3 Distribution of successful and failed metalinguistic feedback 182
8.4 System performance appointments task 183
8.5 Breakdown of interpretation failures 184
8.6 Learners' performance on weil-clauses and distribution of system recasts 184
8.7 Questionnaire results 188

9.1 Test results for SubC 194
9.2 Test results for DatPP 198
9.3 Between group differences, DatPP SC test 199
9.4 Test results for DatPP SC test 200
9.5 Between-group differences DatPP TGJT 203
9.6 Test results DatPP TGJT 204
9.7 Test results DatPP TGJT, grammatical items 205
9.8 Test results DatPP TGJT, ungrammatical items 206
9.9 Significant differences between test results for all groups 207
9.10 Breakdown of existing participant data 208
9.11 Summary temporal measures for appointment task 215
9.12 Summary temporal measures for directions task 219
9.13 Summary of oral skill development 220

10.1 Summary of between-test differences, temporal measures 225
10.2 Summary of between-group differences across all tests 226

1 Introduction

One of the most impressive stories of my childhood was that of a young girl who unwittingly falls asleep with a book under her pillow written in a foreign language. She wakes up the next morning, speaking that language fluently. Part of the ensuing adventure is to find out which language she speaks and she goes on to leave other books under her pillow, this time deliberately, to pick up huge amounts of new knowledge in the process. Even decades after first reading this story, I never get tired to tell others about it if the conversation comes to the difficulties of learning a new language. I think I am not mistaken to believe that such an effortless acquisition of foreign languages is a great dream shared by many. If books under pillows seem a bit magical, a current, more contemporary version of that fantasy may be to have a machine or a software application that helps you learn a new language at a much faster rate than the traditional ways, preferably without much effort on your part. Unfortunately, despite all the recent revolutionary advances in technology, the acquisition of new languages seems to remain a considerable hurdle for all but the most talented adults. We are not much closer to an effortless automatic upload of language knowledge to the storage device that is our brain.

1.1 Motivation

The work in this thesis is concerned with the efforts spent by researchers of second language acquisition, computational linguists, computer scientists, and engineers to find ways to help us learn foreign languages more efficiently and with less effort. It explores how current methods and technology from the field of natural language processing (NLP) and computational linguistics (CL) can be employed to provide opportunities for foreign language learning and practice. Our work is thus driven and informed by two fields of research. On the one hand, we consider pedagogical issues grounded in existing research in the field of second language acquisition (SLA)

and foreign language learning (FLL). On the other hand, we take into account existing work in the field of NLP and CL in general, and its application for the field of intelligent computer-assisted language learning (ICALL) in particular.

Problems

Despite the fact that SLA and ICALL seem very relevant to each other, as the former is trying to understand the cognitive processes that govern the learning of a new language while the latter is concerned with developing tools to support the learning processes, the two fields are relatively distinct and only a few researchers have sought to integrate both perspectives. Currently, relatively little is known about the effectiveness of NLP-informed ICALL because, despite the plethora of applications that exist, they only rarely get evaluated in terms of the learning gains they enable. Related to that, ICALL developers often fail to take into account pertinent findings from SLA research or if they do, they relate to them mostly in superficial ways. Part of the reason for that gap may be the fact that SLA research tends to be conducted in settings that focus on the interaction between humans, mostly the learner and the teacher, and often in a classroom-like situation. It is not obvious how these contexts can be transfered to the computer-centered contexts that dominate in the ICALL sphere. How can findings from human-human communication be applied to human-computer communication? Is it sufficient to just attempt to replace the teacher with the computer or should we, considering the particular strengths and limits of computers, attempt to find ways in which the computer complements and augments conventional ways of learning and teaching languages?

Objectives

The goal of this thesis is essentially to address these questions and problems by developing an exemplary ICALL application with a profound consideration of relevant SLA research. We attempt to employ the current state of research and its open issues to inspire the design of ICALL interaction. On the other hand, we make use of the current state of the art in NLP and CL for implementing an ICALL application which then serves to generate new insights into the conditions of learning with ICALL support and thus contributes to the body of SLA research. Thus, through mutual inspiration and support we add more links to connect SLA and ICALL which are still so unconnected. Considering the current limitations of the conversational skills of computers and the expenses that are required to implement such skills, we also hope to gain insights into the question how worthwhile it is to put forth such effort in relation to the expected effect for learning we can yield. In summary, the goal of this thesis is to examine how current methods and technology from the field of NLP and CL can be employed to provide opportunities for foreign language learning and practice. More specifically, we focus on approaches that engage the learner in a dialog and provide feedback about the learner's utterances.

Approach

The first step to approach our goal is an extensive review of current state of the art and research in the disciplines SLA and NLP, ICALL, and NLP and CL. This review allows us to identify parameters and issues that are relevant for both the pedagogical point of view (SLA/FLL) and the implementational perspective (ICALL and NLP/CL). The result are two pedagogical and two implementational parameters that we describe in the following.

This study relates to two widely discussed issues within the discipline of SLA. One is the debate that pits *form* against *meaning* and leads to a discussion of the extent to which language instruction should focus on linguistic forms and formal correctness as opposed to emphasizing communicative skills and the ability to use the language to make meaning in the real world. Related to that is the second controversial issue which concerns the dichotomy between implicit and explicit knowledge and learning: How explicit or implicit should instruction be, how does the degree of explicitness affect the development of explicit and implicit knowledge, and how do these two types of knowledge contribute to language skills? While some argue that language proficiency can only evolve from implicit instruction, others make a case for the effectiveness of explicit instruction.

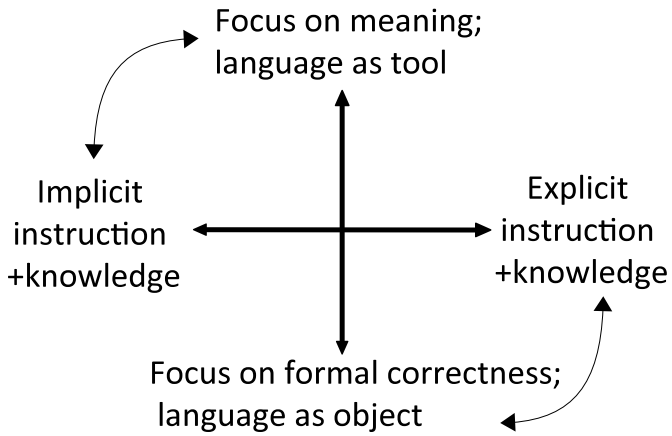


Figure 1.1 – Pedagogical aspects: Parameters in language instruction and learning

These two oppositions can be illustrated as two dimensions that span the space of parameters for possible objectives in language instruction. Figure 1.1 illustrates such a space. The vertical axis describes the opposition between instruction that focuses on meaning and instruction that emphasizes formal correctness. While the first considers language as a tool to accomplish goals arising in real life, the latter considers language as an object that is to be learned. The horizontal axis indicates the range between implicit and explicit forms of instruction and the corresponding type of knowledge that may result from this instruction. As will be shown later in more detail, the two dimensions are not entirely independent. The form-oriented approach is related to explicit instruction in that form-oriented instruction often draws explicit attention to formal features of the language. The meaning-focused instruction is usually more implicit

because specific language features are usually not explicitly mentioned. This relation is indicated by the two bent arrows.

These pedagogical dimensions are of particular interest and relevance from an NLP and ICALL engineering perspective, as they entail considerably different efforts for developing applications that can provide the according types of instructions. More specifically, we consider two aspects that regard the engineering perspective and the developmental effort for different kinds of interaction. One aspect is the degree of freedom the learner is given for producing utterances in the language to be learned. The second aspect is the informativeness and specificity of feedback that learners receive in response to problematic productions. In general, the more freedom and flexibility a learner has to form utterances, the more sophisticated a conversational agent needs to be in order to react appropriately to this unrestricted learner input. Similarly, the more informative a certain type of corrective feedback is, the more knowledge needs to be implemented within a system that can provide such feedback. The effort for handling constrained input and providing uninformative feedback on the other hand is comparatively low. Figure 1.2 illustrates the two dimensions and their relation to the developmental effort.

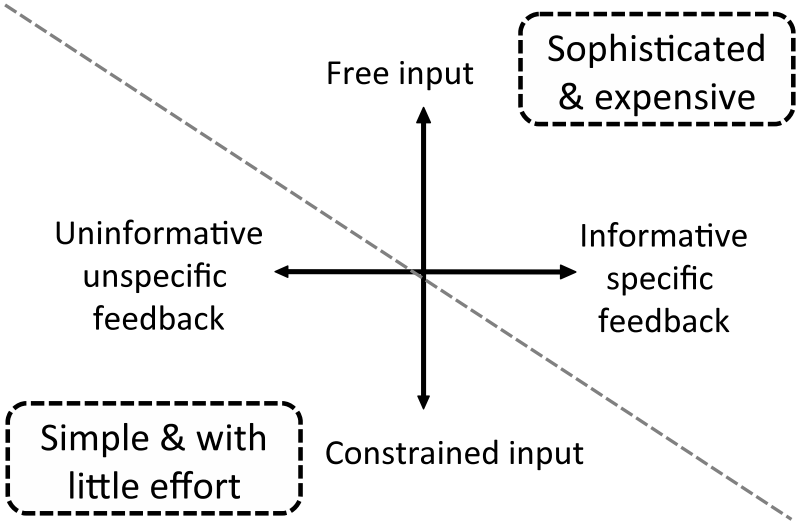


Figure 1.2 – Implementational aspects: Dimensions for sophistication and computational effort of ICALL applications

The pedagogical aspects and the implementational aspects in this model relate to each other in different ways. First, meaning-based instruction usually allows the learner to produce relatively freely, whereas form-based instruction is often realized by exercise types that are rather constrained. However, as we will show later in this thesis, this correlation is more a tendency than a firm rule, as there are meaning-based types of instruction that can be constrained as well as form-focused types of instruction that allow for relatively unconstrained input. Second, there is a relation between feedback specificity and explicitness – more specific feedback is usually more explicit, while less informative feedback tends to be more implicit.

The focus on this parameter space is one crucial aspect of our approach. A second central condition of our approach is the objective to gain insights that are valu-

able for SLA. This requires us to evaluate the ICALL instruction we develop in terms of the learning gains it enables. Such an in-depth evaluation requires a considerable amount of human resources, in shape of learners willing to participate. Limited access to participants then prevents us from attempting a broader exploration covering a more extensive set of instances from the parameter space. Instead, we constrain ourselves to only a few instances, but compare those in more depth regarding their effect for language acquisition, in an experiment following the rigor of SLA studies.

Contributions

Based on our approach we hope to be able to make the following contributions. We provide an example of how to integrate the goals of SLA research and ICALL development in one unified approach to examine and compare the effect of instructional parameter on language learning and put them in relation to the implementational efforts required to realize them. Through that we evaluate the benefit of using NLP-informed ICALL. We further show how human-centered ways of instruction can be transfered and implemented in a human-computer setting and examine if the effects of instructions in both contexts are comparable. The experimental results allow us to make a concrete contribution to the existing research in SLA. The contribution to ICALL lies in the development of an application that supports both learning and learning about learning. Related to that, the contribution to the field of NLP/CL lies in investigating how different levels of effort and sophistication in NLP afford different instructional parameters of ICALL that in turn lead to different experiences and effects in language learning.

1.2 Outline of the thesis

We describe the structure of the thesis chapter by chapter below.

Chapter 2: Computer-Assisted Language Learning

This chapter provides the first part of the technological background to this thesis. It starts by discussing the motivation for the use of NLP methods for developing ICALL applications and describing the pertinent challenges. It then focuses on one of the challenges by presenting an overview of the different approaches for error diagnosis. The chapter finishes with a review of relevant research in the field of computer-mediated communication between humans. Even though this is an area that does not depend on NLP, it provides insights to the non face-to-face mode of synchronous written interaction that resembles the communicative setting between humans and computers that we will explore in the present study.

Chapter 3: Dialog for Language Learning

This chapter provides the second part of the technological background by focusing on the computational treatment of human-computer dialog. In the first part, we will discuss every aspect of dialog modeling, comprising a characterization of the features

and structures of natural dialog, a description of the general architecture of dialog systems and the role of the components, and finally a review of the approaches to dialog modeling and management. The second part presents the current state-of-the-art by providing a comprehensive survey of existing interactive ICALL systems.

Chapter 4: Second Language Acquisition

This chapter presents the pedagogical background of this thesis by discussing the relevant concepts, theories and empirical findings related to SLA/FLL. We discuss the different approaches to language instruction that differ in respect to (a) how much emphasis they put on meaning versus form and (b) how explicit or implicit they are. The chapter further discusses properties of linguistic structures that influence how easily they can be learned. The chapter finishes with a presentation of conversational interaction and task-based instruction.

Chapter 5: Feedback

This chapter completes the background discussion of this thesis by reviewing the relevance of feedback. The chapter integrates the two angles SLA and ICALL by classifying feedback that is provided in SLA contexts and relating it to the technological conditions to provide such feedback through an ICALL application. The chapter finishes with an in-depth inspection of recasts and metalinguistic feedback and the existing empirical evidence for their effectiveness, since these are the feedback types we further examine in this study.

Chapter 6: The Approach

Based on the background of the theoretical and empirical material expounded in the preceding chapters, this chapter discusses the approach we used for exploring how language technology can support language learning. It identifies important parameters from both the pedagogical and implementational perspective that serve as a means to constrain the general goal of this work and render it into a more focused study with concrete experimental research questions that seek to compare the effect of different instructional parameters. The chapter then introduces the research design we adopt and justifies the choices we make.

Chapter 7: The Experiment

This chapter describes the details of the experimental setting that we employ to answer the concrete questions and compare the parameters. This comprises the selection of linguistic target structures and the specification of the tasks and interaction that constitutes the experimental instruction. Furthermore it also includes a discussion of the measures that we use to assess the development of linguistic knowledge and skills. Finally, we describe the procedures and conditions of the practical implementation of the experiment.

Chapter 8: The System

This chapter describes further details about the design and implementation of the ICALL dialog system that we use for providing the experimental instruction. Furthermore, it provides a detailed analysis of the performance of the system during the experiment. The chapter concludes by presenting the results of the learner survey regarding their perception and enjoyment of the system.

Chapter 9: Findings

This chapter presents the results of the experiment in detail. First, this includes the development of grammatical accuracy for the target structures in terms of test scores along the four test times during the experiment. Second, we describe the development of the communicative spoken language skills as measured by holistic ratings and temporal analysis of fluency in speech samples.

Chapter 10: Discussion

This chapter discusses our findings in the light of the concrete and general questions we sought to answer with the experiment. We will also discuss the limitations and suggestions to address them in future work.

Chapter 11: Concluding Remarks

The final chapter concludes this dissertation by summarizing the contributions and giving a brief outlook on possible continuations and extensions of the presented work.

2

Computer-Assisted Language Learning

2.1 Introduction

This chapter discusses computer-assisted language learning (CALL) and the role of natural language processing (NLP) for CALL. In general, CALL refers to technology and software applications that support people in learning foreign languages. CALL applications can be used as supplement to traditional teacher-dependent language instruction or as a substitute, in case teachers are unavailable or unaffordable (Nerbonne, 2003). Both as substitute and as additional resource, one key motivation for CALL is that it fosters the autonomy of learners (Benson, 2001).

The use of computers for language instruction dates back to the 1960s, even before the advent of personal computers. The PLATO system (Programmed Logic/Learning for Automated Teaching Operations) (Curtin et al., 1972), which provided grammar drills on a mainframe computer, is often cited as one of the first CALL efforts (Levy, 1997). Since the days of these first approaches, many other systems and tools have been developed, and today CALL is a broad discipline which covers a variety of activities and applications.

These applications can be classified according to the language areas and skills that they target. Levy (2009), for instance, in his review of CALL technology, distinguishes the following target areas: grammar, vocabulary, reading, writing, pronunciation, listening, speaking, and culture. A different classification has been proposed by Zhao (2003), who distinguishes three functions of CALL applications:

1. Providing access to linguistic and cultural material
2. Providing opportunities for communication
3. Providing feedback on learner responses

Access to linguistic and cultural materials refers to the context-dependent provision of additional lexical, morphosyntactic and cultural information, which supports the learner

in understanding authentic material that originated from native contexts but was not specifically targeted at learners (see, for instance, Lyman-Hager (2000) and Nerbonne and Dokter (1999)). Other examples of features that enhance comprehensibility are captions for videos or the option to slow down the speech rate of audio material (Shea, 2000; Zhao, 1997). Authentic material can also be automatically enhanced to emphasize linguistic forms and make learners more aware of them (Meurers et al., 2010).

With regard to *opportunities for communication*, Zhao (2003) distinguishes two areas. One is concerned with technology that enables learners to communicate remotely with other learners or native speakers – this field is known as computer-mediated communication (CMC). The other area refers to technology that allows learners to conduct near natural conversations with a computer program, we know these as dialog systems and conversational agents.

Finally, the *provision of feedback* on learner utterances comprises corrections on errors in pronunciation, orthography, morphology, syntax, semantics, and even pragmatics. It also includes the development of learner models based on a record of previous errors.

In this and the next chapter we will see examples of CALL applications that serve one or more of these three functions. Many of these CALL applications require rather sophisticated, i.e., *intelligent* techniques. Indeed, with the beginning of the 1990s, when artificial intelligence (AI) technologies had reached a sufficient state of maturity, they brought forth a subdiscipline of CALL - *Intelligent* CALL (ICALL). In a general sense, intelligent CALL comprises the use of techniques such as knowledge representation, expert systems, intelligent tutoring systems (ITS), user modeling, natural language processing (NLP), automatic speech recognition and speech synthesis, and machine translation (for reviews see Gamper and Knapp (2002); Levy (2009); Schulze (2008)). Most often, however, ICALL is used in a narrower sense in which *intelligent* refers particularly to the automated analysis and generation of natural language. To eschew this ambiguity, some prefer to call it *parser-based* CALL, referring to the process of parsing, which describes the syntactic analysis of natural language (Schulze, 2008). We will use the term ICALL to refer to the NLP-supported CALL in this thesis.

This chapter has three parts which present different aspects of CALL and ICALL. In Section 2.2, we will discuss the reasons to use NLP techniques in ICALL, describe the challenges and introduce a general strategy to deal with the challenges. As we will see, one of the challenges is the frequent occurrence of errors in learner language which an ICALL application has to account for in some way. In Section 2.3 then we will review the range of approaches to error diagnosis which is a prerequisite to provide feedback. In the last section of this chapter (2.4) we will take a step back to non-intelligent CALL by giving an account of computer-mediated communication between humans. This area does not rely on using NLP methods but it is relevant for this thesis since the remote, non face-to-face mode of interaction resembles the communicative setting between humans and computers that we will employ for the current study.

2.2 Natural language processing in ICALL

2.2.1 Motivation

While a large proportion of today's CALL applications make no use of NLP techniques, the need and value of such an enhancement has been widely recognized (Meurers, 2012; Nagata, 2009; Heift and Schulze, 2007). The main argument for employing NLP is its ability to cope with relatively free and unconstrained learner input. Meurers explains the advantages of NLP in the following way: Traditional language activities such as, for instance, multiple choice questions or gap-filling involve only a small set of predefined learner responses and an equally small set of system responses. In such a context, learner responses and the corresponding feedback of the system can be enumerated explicitly. Comparing the actual learner response with the set of expected responses is a matter of simple string comparison. However, this approach becomes unfeasible if the goal is to allow the learner to produce language freely, as in communicative, meaning-based tasks. Also for more constrained activities such as summarizations or sentence translations, the number of possible correct answers is too large to be listed extensionally. This is because natural languages are rich and one meaning can be expressed by many different realizations. Nagata (2009) illustrates this problem by showing how a seven word target response for a translation task from English to Japanese can result in more than 6000 correct responses and almost a million possible incorrect responses through the combinatorial explosion of lexical, orthographical and word order variants. Enumerating these variants and the corresponding feedback extensionally is obviously not feasible. Therefore, a more concise, intensional representation of possible learner responses and the mapping to feedback is needed, if one wants to treat relatively free learner input. This can be realized using recursive structures or linguistically informed grammar formalisms instead of extensional list of strings (Meurers, 2012; Nagata, 2009; Heift and Schulze, 2007).

2.2.2 Expectations and challenges

Although the benefit of NLP techniques in CALL is commonly acknowledged, instances of NLP-enhanced ICALL are still rather rare within the greater field of CALL today. In a review of CALL literature, Stockwell (2007) mentions NLP only as a side note, the vast majority of the technology he reviewed does not use NLP techniques. Further, if NLP is used, it is often not very sophisticated: "most grammar programs are still very basic in the ways they process learner input, diagnose errors, and provide feedback" (Levy, 2009, page 770). One reason for this may be the considerable cost and effort that is required to develop such NLP tools and resources (Schulze, 2008). Apart from that, there is some skepticism regarding the capability of NLP to support automated language learning. Salaberry (1996), for instance, argues that NLP cannot deal with the complexity of natural language. However, Nerbonne (2003) surmises that Salaberry's skepticism is probably grounded in inflated expectations on the part of learners and teachers. Obviously, we are still a long way from perfectly imitating human-like language abilities in artificial systems. For instance, until today, no computer program has managed to pass the Turing test, i.e., make its behavior indis-

tinguishable from human behavior as judged by humans (Saygin et al., 2000; Shieber, 2004). Furthermore, despite the long history of machine translation (MT), current MT systems are still far from achieving the skills of a human translator. As Feigenbaum (2003) notes, for an artificial system to *understand* as well as a human is still an open challenge, despite the relative success in analyzing the syntactic structure of natural language. Gamper and Knapp (2002) summarize: "A full-fledged analysis of written text in all its complexity is a very difficult task, which exceeds current state of the art technology in NLP" (page 334).

However, while there are certainly some aspects of language complexity that are still hard to process, the usefulness of NLP in focused and controlled, if less ambitious, approaches has been convincingly demonstrated in a wide range of applications. For instance, the technology for morphological processing, that is, the analysis of the inner structure of words and how they are constructed of smaller meaningful units, known as morphemes, is sufficiently advanced. For many languages, it is mature and reliable enough to provide almost error-free lemmatization – deriving the canonical form of inflected word forms – as needed, for instance, in dictionary lookup tools (Nerbonne et al., 1998; Nerbonne, 2003).

Ambiguity

Contrary to that, syntactic and semantic analysis are much more challenging due to the inherent ambiguity of many sentences. There are two types of ambiguity, lexical and structural. Lexical ambiguity refers to the fact that a word can have several meanings. Often, contextual information helps to disambiguate the word and arrive at the appropriate meaning. Structural ambiguity describes the fact that a sentence can have more than one possible syntactic structure, and consequently also more than one meaning. Consider as an example the sentence *"I shot an elephant in my pajamas."* If its ambiguity is not apparent to the reader at first sight, it becomes evident when followed by the addendum *how he got in my pajamas, I'll never know*[1]. The ambiguity is based on the prepositional phrase *"in my pajamas"*, which can specify either the object of the sentence (the elephant is wearing the pajamas) or the subject (the one who shoots is wearing pajamas).

Ambiguity and the analysis of learner errors

While ambiguity is already problematic within the domain of native and correct language, it is even more difficult for learner language, which is often incorrect. Erroneous language is parsed with more difficulty, because the potential for ambiguity is increased. Amaral and Meurers (2011) explain that for the analysis of native and correct language, the search space is constrained by lexical and syntactic rules. However, since learners are likely to violate these rules, the rules need to be extended to account for potentially ill-formed learner input. The expansion of rules increases the search space and thereby the number of possible ambiguities. Consider, for example, the learner production *"The man eat cheese."*[2] The sentence is incorrect according to stan-

[1] The joke is attributed to Groucho Marx in the film *Animal Crackers*.

[2] Thanks to Detmar Meurers for this example.

dard English, but the source of the error and the intended meaning is unclear. The verb form *eat* cannot be used with third person singular subjects - so the learner might have used an incorrect verb form – it should be: *The man **eats** cheese.* But it is also possible that the learner intended to make a statement about several men and failed in producing the correct plural form – the correct sentence would be: *The **men** eat cheese.* Yet another possible source of an error in this sentence is the use of a wrong tense. If the learner wanted to express that the event has already taken place in the past, they might have failed in producing the past tense form *ate*. The correct sentence would be *The man **ate** cheese.* In summary, there are at least three possible sources of errors for this example.

Analyzing ill-formed language is hard because the deviations from correct input increase the space of possible analyses. These difficulties, however, have not deterred ICALL researchers and engineers from attempting to implement natural language processing facilities in their systems. Tokenizers, morphological analyzers, part-of-speech taggers, chunkers, tools for concordancing and text alignment, parsers, and semantic analysis tools have been successfully put to use (Amaral and Meurers, 2011; Nerbonne, 2003). We will describe some of these use cases in Section 3.2. However, developing tools for deeper analysis is a complex and costly endeavor (as noted, among others by Schulze (2008)). Developers have therefore sought for another approach to compensate for the increased difficulty of higher level analysis.

2.2.3 Constraining input as a strategy to deal with limits

A common strategy for dealing with the difficulties in analyzing learner language and for making processing tractable is to constrain the possible input that the learner can give to the system (Amaral and Meurers, 2011). The key is to do this in such a way that the learner does not feel too constrained and that the activity is still effective for fostering language skills. One very restrictive way to constrain the input is to let the learner choose from a set of pre-fabricated utterances, an approach taken, for instance, in the interactive systems described by Pollard and Yazdani (1993) or Stewart and File (2007). However, such restrictions eliminate the need for using NLP techniques altogether. A less constrained approach is taken for instance by Nagata (2009) or Heift (2003), who constrain learner input through the choice of task type, e.g., by prompting for a translation, dictating sentences, or providing a list of words that is to be used for the response. Another approach is to a priori constrain the input language to a sublanguage covered, for example, by a first-year textbook (Schwind, 1995; Levin and Evans, 1995). Schwind argues that a system should work on a sublanguage which is entirely known to the system, and the system "has to ensure that the student does only form sentences which can be analyzed, i.e., does not form well-formed sentences outside the competence of the system" (page 296). Schwind further explains that "[t]his requirement is fulfilled by formulating the exercises so as to suggest a restricted language to the student". Although she remains unspecific about how exactly to achieve this, she seems to imply the usage of more implicit ways of constraining the learner language. This is in accord with the desire to provide more freedom to the learner and "more space for negotiation of meaning as needed for meaning-based activities" (Amaral and Meurers, 2011, page 9). Amaral and Meurers propose to use pictures, lists

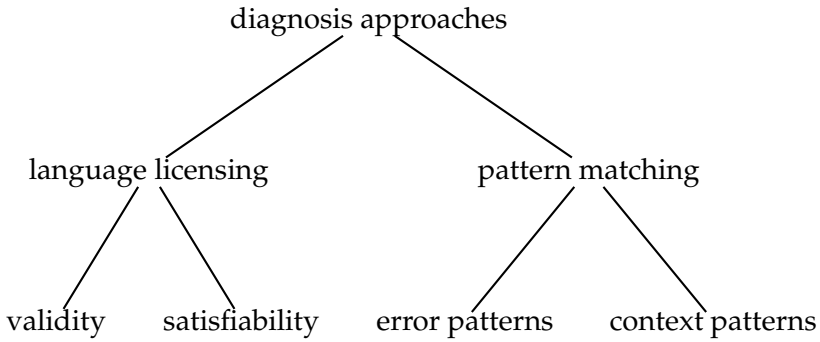


Figure 2.1 – Taxonomy for error detection and diagnosis according to Meurers (2012)

of L2 words given as prompts or written cues in L2 to implicitly constrain the learner input to the system. Price et al. (1999) consider the task scenario of their dialog system (ordering meals in a restaurant) to be a sufficient restriction to the possible learner production.

The fact that adequate processing of largely unconstrained learner input is still beyond the current state of the art is confirmed by failures of projects that aimed at just that. Amaral and Meurers (2011) cite El Corrector (Klein, 1998) and FreeText (L'Haire, 2004) as examples of such disappointed expectations.

Constraining the learner input is related to the difference between emphasizing either meaning or formal correctness and explicit and implicit instruction. The more constrained the learner input is, the more likely this results in an explicit way of instruction and one more focused on forms. The less constrained input is more likely to provide a more implicit instruction and a focus on meaning. We will discuss these pedagogic parameters in more detail in Chapter 4. For the study that we conduct in the scope of this thesis, we will come back to the issue of constraining learner input by using it as one of the implementational parameters that determine the developmental cost of an ICALL system.

After this general characterization of the state of the art in NLP for ICALL and its challenges, we will now discuss a particular relevant challenge for treating learner language.

2.3 Error diagnosis

Since learners of a foreign language are likely to produce utterances that are divergent from the target language, any ICALL application should be able to handle such divergences. There are basically two different ways to treat ill-formed and unexpected language – the deviation can be ignored or it can be diagnosed. In the context of language learning, errors are usually ignored when the primary goal is to communicate with learners and to provide meaning-based conversation. On the other hand, errors are diagnosed in systems that are built to provide corrective feedback. ICALL systems that follow the communicative approach are built to be robust regarding ill-formed input and therefore to ignore most errors (Jehle, 1987; DeSmedt, 1995; Sanders and Sanders, 1995). Similarly, many chat bots – artificial agents that engage in small talk

with humans primarily for entertainment – are another example for robust approaches that tend to gloss over errors for the purpose of maintaining a coherent conversation (we will discuss chat bots in more detail in Section 3.2.3).

Robust systems usually take a shallow approach to language processing – instead of attempting a complete grammatical analysis of the entire input they pick features of the input, for instance, keywords, key phrases, or patterns and process them. Working with features is also a characteristic of statistical, data-driven approaches to language processing – they attempt to solve the task by deriving and applying probabilistic models that estimate the likelihood of a specific analysis based on features of the input. Such statistical approaches are in general more robust in respect of unexpected input.

However, there are two important types of applications for which it is essential that errors be detected, diagnosed, and corrected instead of ignored. One type of application is spell and grammar checking for native speakers, the other type are pedagogical applications for L2 learners that provides corrective feedback for errors. Although a subset of learner errors can be covered by traditional spell and grammar checkers that were originally targeted at native speakers, there are many errors in non-native language that differ quite substantially from those in native language (Rimrott and Heift, 2008). As a consequence, tools that were developed for native speakers are in general not well-suited to handle learner language. Consider, for instance, a learner who produces the non-word "goed" for which a conventional spell checker would suggest "god" or "goes" as corrections. These corrections are based on string similarity metrics, some of which, for instance, consider to the number of edit operations required to transform one string into another (Gusfield, 1997). However, using such metrics, the spell checker is unable to propose "went" as an alternative because it cannot guess that the learner intended to produce the past tense of the verb "go".

For errors in learner language, "the goal is to understand what the student wanted to do, where he went wrong and what grammar rules he misunderstood or was unaware of" (Schwind, 1995, page 295). In addition to analyzing the language, this goal requires the location of the error and the provision of a correction and/or explanation if necessary. We will now briefly sketch current approaches to error detection and diagnosis, following in large part the classification proposed by Meurers (2012). Figure 2.1 illustrates the taxonomy of such strategies, according to Meurers. On the top level, approaches fall into two categories, those based on **language licensing** and those based on **pattern-matching**. Language licensing approaches usually attempt to analyze the complete learner utterance, while pattern-matching approaches focus on parts of the utterance that fit a pattern and ignore the rest.

2.3.1 Language licensing

Language licensing refers to the way formal grammars are used to describe the well-formed and acceptable utterances of a language. Errors are detected based on the fact that they cannot be licensed by the formal description of the language. In other words, erroneous utterances are not covered by the grammar and consequently are not licensed. There are basically two different approaches to describe a language by formal grammars, one is based on *validity*, the other is based on *satisfiability* (Johnson, 1994; Meurers, 2012). In validity-based grammars, the grammar is a set of axioms, usually

called *rules*. An utterance is licensed by the grammar if it is possible to derive the utterance by expanding the rules. Satisfiability-based grammars on the other hand, are construed as a set of constraints. An utterance is licensed by the grammar if it satisfies all constraints.

For each of the two kinds of grammar models there is a corresponding approach for diagnosing learner errors. In validity-based grammars, errors are handled by adding rules that cover the erroneous input – this is known as the **mal-rules** approach. For satisfiability-based grammars, errors are handled by relaxing certain constraints, for instance agreement constraints, with the result that more utterances than before are accepted by the grammar. This is known as **constraint relaxation** and goes back to work by Kwasny and Sondheimer (1981). The constraint relaxation technique can be used in strictly satisfiability-based grammars, but it can also be applied to rule-based grammars that are augmented with constraints for features of the components of the rules.

As an example for the mal-rules approach, consider a grammar rule that covers agreement errors as in (1)

(1) He drive.

A simple context-free grammar that does not license this string would contain the following rules (2), which state that a sentence (S) consists of a noun phrase (NP) and a verb phrase (VP) and that the NP and VP agree on their person and number feature – they are either both 3rd person singular or they are both not 3rd person singular.

(2) $S \rightarrow NP_{3sg}\ VP_{3sg}$
$S \rightarrow NP_{\neg 3sg} VP_{\neg 3sg}$

A mal-rule that would license this string could be like (3):

(3) $S \rightarrow NP_{3sg} VP_{\neg 3sg}$

In a constraint-based approach, agreement could have been modeled through a constraint like (4-b) for the rule (4-a), in which *agr* is a feature structure that contains information about number and person.

(4) a. $S \rightarrow NP\ VP$
b. $\langle NP.agr = VP.agr \rangle$

(1) violates this constraint and by relaxing it, we are able to license the string and keep a record of which constraints were violated. The advantage of the mal-rules approach is that the feedback can be fairly specific because each mal-rule can be annotated with an explanation. The disadvantage is that it requires that learner errors be anticipated (Heift and Schulze, 2007). The constraint-relaxation approach is more flexible in that errors do not have to be explicitly anticipated (Menzel and Schröder, 1999). However, an error can only be diagnosed if it corresponds to a specific constraint (Meurers, 2012). There are also approaches that combine the two techniques in order to compensate the disadvantage of each (Reuer, 2003; Schwind, 1995).

2.3.2 Pattern matching

Pattern matching approaches are the second general class of diagnosis approaches besides the licensing-based approaches. Pattern matching approaches are based on detecting the divergence of the learner input from some correct model (De Felice, 2008). They are thus focusing on specific parts of the utterance. Like the mal-rule approach, they are also derived from anticipations of errors. Meurers (2012) distinguishes further between **error patterns** and **context patterns**. Error patterns are usually restricted to very specific well-known errors in a small specific context, such as "suppose to be" instead of "supposed to be" or two consecutive articles: "the a". Such error rules have been implemented in the open source LanguageTool (Naber, 2003) and commercial grammar checkers. Although they are originally intended for native speakers, they can be adapted to cover typical learner errors – if these errors are known.

Learner errors that are less specific and contingent on a larger context, can be treated by context patterns (De Felice, 2008). The contexts of problematic items, such as determiners or prepositions in English are modeled based on the properties of a corpus of correct usage. The context is described by lexical, syntactic, and semantic features. In order to check the correctness of a learner language sample, the context features of problematic items are compared to the context features in the correct model. As a simplified example, consider a learner sentence in which a preposition p_x is used in the context of a feature vector $f_v = \langle f_1, f_2, ..., f_n \rangle$. The correct model contains for each feature vector $f_v = \langle f_1, f_2, ..., f_n \rangle$ and a preposition p the probability that p occurs. If, according to the correct model, the most likely preposition in the context of f_v is p_x, then the learner is probably correct. If, however, the most likely preposition in the context of f_v is another preposition p_y with $p_y \neq p_x$, p_y will be proposed as a correction.

In this way, rules for correct usage are not modeled explicitly, because this would be hard or impossible, but the underlying regularities, implicitly contained in a native speaker corpus, are used to notice the errors in the learner language and to predict the correction (De Felice and Pulman, 2008). As such, the model cannot give a grammatical explanation. The quality and power of context patterns depends on the features they use. Context patterns have been used successfully for problematic items that are correlated to features that are within short distance, as is the case for determiners and prepositions. However, local features are not able to adequately model long-distance dependency relations. For instance, checking agreement between distant subjects and verbs would require features that model dependency structures (Levin et al., 1991).

2.3.3 Summary

Table 2.1 summarizes the properties of the different approaches to error diagnosis that we discussed above. For each of the four approaches it indicates whether it requires to anticipate the error, whether the entire input is used, and whether it can provide a correction or an explanation.

A method that falls outside of the two general classes of error diagnosis is presented by Vlugter et al. (2006). In their approach the original input is modified by permutating character and word sequences based on error hypotheses. These variants, that are potential corrections of the original input are then parsed with the non-

Approach	Anticipation	Entire Input	Correction	Explanation
Language Licensing				
Mal-rules	+	+	+	+
Constraint relaxation	-	+	◯	◯
Pattern matching				
Error patterns	+	-	+	+
Contextual patterns	+	-	+	-

Table 2.1 – Error diagnosis techniques and their properties. The symbol ◯ indicates that whether or not the feature is existent depends on the variant that is used.

expanded grammar if parsing of the original input failed. Thus, error hypotheses are expressed by creating transforming rules for corrections. A correct variant that is successfully parsed can then be used as a suggestion for correction. Since the hypotheses are parsed, this can be considered as an example of the licensing approach. On the other hand, the rules for creating permutations are based on specific known error patterns.

The ability to correct and explain errors makes feedback more informative and arguably more useful. However, under certain circumstances less information may be advantageous too. We will discuss this issue in more detail in Section 5.4.

Aside from using NLP to detect and diagnose errors, a number of ICALL applications have also tried to reproduce certain characteristics of an expert teacher by trying to assess the importance of an error and to develop a learner model based on past learner errors in order to adapt the feedback and remediation (Levy, 2009).

In Section 3.2 we will present examples of ICALL systems that have implemented different error diagnosis techniques and provided feedback accordingly. We will discuss the benefits of feedback in more detail in Chapter 5. For the present study, we will select a certain approach for diagnosing errors based on the requirements of the instruction and the desired information content of the feedback, as discussed in detail in Chapter 6 and 8.

2.4 Computer-mediated communication

Computer-mediated communication (CMC) refers to communication between learners on the one hand and teachers, native speakers, or other learners on the other hand via communication tools such as text, voice, or video chat, bulletin boards, or e-mail. CMC belongs to the larger field of CALL because it serves one of its functions, namely, to provide opportunities for communication. In particular, text-based chat has received a significant amount of attention among the language learning community lately. This form of communication is situated somewhere between planned, formal writing and spoken, spontaneous language (Abrams, 2003). As Abrams further characterizes, text chat allows more time for processing and planning than oral interaction, but less time than ordinary writing since the interaction is intermediate, and

responses are expected within a short time window. Compared to oral conversation, it is easier for the interlocutors to converse about different topics at the same time, which can result in interleaving and overlapping strands of discourse. Finally, while text chat releases learners from the demands of adequate pronunciation, it requires additional effort for orthographic encoding and decoding.

Regarding this thesis, CMC is relevant not in terms of the underlying technology - which does not involve any linguistic processing or artificial intelligence - but in terms of the interaction that it entails. In text chat CMC, as well as in text-based human-computer interaction, which we examine in this thesis, the learner produces utterances and receives immediate feedback via a text-channel. The characteristics of this mode of interaction can be advantageous for language learners. The visual salience and the fact that learners can re-read all utterances during the conversation enables learners to better attend to formal aspects while still maintaining the flow of communication (Abrams, 2003; Smith, 2004). Further, since writing usually takes more time than speaking, turn taking is slower, which provides more time for processing and planning. It has been suggested that the visual support and the reduced time pressure can focus the attention on target language forms in the input as well as in the learner production (Sauro, 2009) and consequently increase comprehension and accuracy (Smith, 2004).

These theoretical claims have been partially supported by empirical research, which can be divided into work that examines properties of actual text chat discourse and work that examines the ensuing effects of participation in such discourse on language skills. As we will describe in more detail below, it has been shown that text chat induces more self-correction and more complex language than face-to-face oral conversation. It also leads to a greater amount of learner production and more balanced participation. Regarding the effects of participation in chat, research findings concern (a) the amount of contributions in subsequent oral face-to-face discussions, (b) general oral language skills, and (c) the acquisition of pragmatic competence.

2.4.1 Properties of text chat language

Lai and Zhao (2006) examined the one-on-one interaction of English language learners of different levels and compared online text chat with face-to-face interaction. They found that the chat interaction elicited significantly more self-correction than face-to-face interaction. This finding suggests that the written mode allows learners to notice forms and problems with them better. This hypothesis is supported by the self-reporting of the learners: 8 out of 11 participants reported that they paid more attention to their own productions in the text chat than in oral interaction. Related to that, there are three studies that show that learner language in text chat is more complex than in face-to-face conversation. This is probably a consequence of the amount of time learners can use on processing. Warschauer (1996) found that in group discussions of four English learners the language used in text chat discussions was more complex in terms of vocabulary and syntax than the language in oral discussions. Lexical complexity was evidenced by a high type-token ratio (total number of different words divided by the total number of words). Syntactic complexity was manifested by a high proportion of subordination. Similarly, Fitze (2006) showed that the lan-

guage produced in text chat exhibited a greater lexical range than the language in oral face-to-face conversation for learners of English in group discussions with 13 or 14 students. Related to that are the findings of Kern (1995): He examined discussions with a group size of 14 to 18 second-semester students of French as a second language. In text chat, learners used a greater range of morphosyntactic features and expressed a greater variety of discourse functions than in oral interaction.

Another advantage of the text chat mode in comparison with oral face-to-face conversation is the increased amount of learner production, which was shown by Kern (1995) and Bump (1990). This was ascribed to the more student-centered nature of chat discourse, which reduces the contributions of the instructor and induces students to interact with each other instead of solely with the teacher (Chun, 1994). Another possible reason for the increase in production is the fact that in chat, contributions can overlap (Kern, 1995). However, evidence by Fitze (2006) and Abrams (2003) contradicts Kern's and Bump's results. Fitze and Abrams found no significant difference in the amount of learner production between oral and text chat mode. Another significant feature is the distribution of productions between different participants. Warschauer (1996) and Kern (1995) showed that the learners' contributions are distributed more evenly in chat group discussions. In contrast, face-to-face group discussions were less balanced, due to the dominance of one or two speakers in each group.

2.4.2 Benefits of participating in text chat

After the summary of research that examined the properties of text chat discourse, we will now give a short account of research that explores if any of the immediate benefits of text chat transfer to subsequent performance. Abrams (2003) found that the amount of production carries over to subsequent oral conversations: Learners who had participated in a text chat group discussion produced more speech during face-to-face discussions than learners who had taken part in asynchronous bulletin-board discussions before.

With respect to language skills, the few findings are mixed. Payne and Whitney (2002) compared the effects of text chat with face-to-face interaction in terms of subsequent oral performance. They found that participants in text chat outperformed participants in face-to-face interaction with regards to general oral proficiency. Proficiency was rated by two human raters, based on a monologic speech sample of five minutes, according to five different criteria: fluency, comprehensibility, vocabulary, grammar, and pronunciation. Note that both types of interaction were conducted in groups of four to six, therefore the results may not be transferable to one-on-one interaction, which ensures a higher rate of involvement for the individual learner per se. Abrams (2003), on the other hand, could find no significant difference between the oral production of students that took part in a chat group discussion compared to students that communicated asynchronously via a bulletin board and a control group who did not communicate at all but worked on regular classroom exercises. She measured the quality of the oral output by means of lexical richness and diversity and syntactic complexity. The group size for discussions of 18-22 students was rather large. Finally, Sykes (2005) found positive effects for the acquisition of speech acts in Spanish as a second language: participating in chat conversations was more beneficial than partic-

ipating in oral face-to-face conversations for small groups of three. The studies that we summarized above differ in the size of the groups whose interaction they examine. Although it is likely that the number of participants has a considerable effect on the properties of the communication and the relative effectiveness of different interaction modes, to our knowledge, this variable has not yet been specifically addressed.

Research about general human-human communication in text chat modus is relevant for this thesis, because text chat shares important characteristics with the interaction mode examined in this thesis. We will examine the effects of human-computer interaction in text mode, and we will evaluate the linguistic development of learners, in terms of accuracy as well as in terms of oral skills.

2.5 Summary

This chapter provided the first portion of technological background that is relevant for this thesis. It started with an introduction to the disciplines of CALL and ICALL and a brief overview of their goals, among which are the provision of opportunities for communication and feedback on learner productions.

Section 2.2 presented an introductory overview of the the use of NLP and CL for the development of ICALL applications. It started with a motivation and illustrated the expectations and challenges related to processing natural language. It then went on to explain how the inherent ambiguity of language is further increased through erroneous learner language and it characterized attempts to constrain the learner input to remedy that problem. Section 2.3 introduced approaches to error diagnosis and presented a taxonomy that distinguishes between language licensing and pattern matching at the top level. Diagnostic approaches can further be classified according to whether or not they rely on an explicit anticipation of the errors, whether or not they process the complete utterance or only parts, and whether or not they provide a correction or an explanation of the error. The chapter concluded with a discussion of computer-mediated communication and described the properties and benefits of engaging in text chat communication in Section 2.4.

In order to round out the background on ICALL and NLP, the next chapter will provide the second part in form of (a) a detailed account on how dialog for language learning can be modeled and treated computationally and (b) a survey of existing ICALL applications.

3 Dialog for Language Learning

When people learn a foreign language, usually, one of their goals is to be able to have conversations in that language. But verbal communication is not only the final goal of learners, it can also facilitate the learning process, by providing comprehensible input and urging the learner to modify their output, a process that we will explain in more detail further down in Chapter 4. One of the purposes of CALL, as we stated in the previous chapter, is to provide opportunities for communication. There, we also discussed how communication between humans can support language learning. In this chapter we now focus on communication between humans and computers.

The chapter is divided into two parts. The first part (3.1) is concerned with every aspect of dialog modeling. It starts with an explanation of phenomena in natural dialog that need to be modeled. It then continues with a general description of dialog systems including architectures and significant features that characterize them. Further on, it goes into more detail by describing the essential components of dialog systems and their functions. This first part concludes with a review of approaches to dialog modeling and management. Based on that, the second part (3.2) presents the current state-of-the-art by providing a comprehensive survey of existing interactive ICALL systems. This presentation is further divided into systems that focus on grammar and systems that focus on dialog and communicative interaction.

3.1 Dialog

This section takes a step back by describing the fundamentals of human dialog and the foundations for building human-computer dialog systems.

Dialog fulfills many different purposes. The goals of communication can range from the requesting and passing of information, negotiating, to asking or commanding others to do certain tasks and to coordinate the accomplishment of shared goals. Through that, dialogic communication serves practical goals as well as social goals

like maintaining relationships. Each participant in a conversation is guided by their individual goals, beliefs, preferences and expectations and a consideration of these is helpful for analyzing and modeling dialog (Bunt, 2000). As such, engaging in a dialog is a collaborative activity between two or more conversational partners. This collaborative nature of dialog is a crucial premise for the analysis and processing of dialog as it gives rise to certain phenomena in dialog that we will explain in the first of the following sections.

3.1.1 Dialog phenomena in human interaction

The collaborative nature of dialog usually entails that the participants work together to achieve their respective goals and that they are dependent on each other's cooperation to achieve these goals. At a basic level, collaborative communication requires that the participants be able and willing to (a) communicate, (b) to perceive the message transmitted by the speaker, (c) to understand it and (d) to react to it, in particular to indicate whether they accept or reject it (Allwood et al., 1992).

Usually, the interpretation of contributions in a dialog relies considerably on the assumption that the dialog partner is cooperative. Grice (1975) posited this assumption as the *cooperative principle* which is realized in four maxims that are assumed to be obeyed by cooperative partners to make the conversation more efficient: The maxim of quality ("be truthful"), the maxim of quantity ("provide as much information as is necessary but not more"), the maxim of relation ("be relevant"), and the maxim of manner ("be clear").

The cooperative behavior of participants in a dialog is strongly determined by social norms and conventions, thus the cooperation often stems from obligations imposed by the culture to which the participants belong (Traum and Allen, 1994; Bunt, 2000). Even if the individual goals of the participants are in conflict, conventions and obligations usually make them compliant on the surface. Consider for instance an agent who wants to keep some fact to themselves – when asked about that fact, they will still provide a response, it just might not contain the desired fact (Traum and Allen, 1994).

Even though most dialog systems are based on a cooperative premise, there are applications which include a non-cooperative element. For instance, tutoring systems in which the goal of the system and the learner may be in conflict, or role-playing games which provide practice for dealing with inherently non-cooperative situations, e.g., as agents in a military conflict (Traum, 2008).

In the remainder of this section, we further describe aspects of dialog structure and interpretation that are tightly related to the collaborative nature of dialog. We will discuss how dialog participants take turns, how they ensure mutual understanding through grounding processes and how their utterances can be interpreted as acts on different levels.

Turns and turn-taking

Conversations consist of consecutive turns of the participants of the conversation. Due to the physical and cognitive constraints of speech-based conversation it is usually

impossible to speak and listen at the same time, hence conversations usually contain only a small proportion of speaker overlap[1]. Turn-taking rules govern when and how speaker shifts take place. One important turn-taking rule is that if in the current turn the speaker selects a next speaker the selected participant can and should have the next turn (Sacks et al., 1974). The selection of the next speaker can be achieved through an utterance that expects a response from another speaker. A prevalent example would be a question that should be followed by an answer. Two-part structures like question-answer are called adjacency pairs (Schegloff and Sacks, 1973) or dialogic pairs (Harris, 2005). These pairs are a small local structure of a conversation and knowledge about them is very useful for modeling dialog. Other examples for such pairs are greeting-greeting, offer–acceptance/refusal, request–grant/decline, thank–accept thank, apologize–accept/reject. Levinson (1983) argues that some second parts are preferred over others. For instance, the preferred response to a request is acceptance, whereas a refusal is dispreferred. Levinson understands preference in terms of linguistic markedness, in which preferred seconds are unmarked and therefore structurally simpler. In contrast, dispreferred seconds are marked through a more complex structure, which manifests as delays in delivery, some preface, and/or an explanation for why the preferred second cannot be given. The concept of preference relates to the *discourse obligations* discussed above that determine how participants in a conversation should react (Traum and Allen, 1994).

Grounding

An essential prerequisite for successful communication is that the participants share a certain number of mutual beliefs – common ground. Following the definition proposed by Stalnaker (2002), common ground is common belief, i.e., a set of propositions that all parties believe and that all parties believe that all parties believe. Common ground is central to dialog as a joint activity as the participants of a conversation presuppose some common ground and the contributions to a conversation modify the common ground. The modification of the common ground is known as *grounding*. It involves the hearer signaling to the speaker that they have understood the speaker's meaning and intention. By that, the hearer provides closure to the speaker, which is evidence that they have succeeded in performing their act of speaking (Clark, 1996). Grounding problems arise through lack of perception or understanding, through ambiguous utterances that lead to misinterpretations, and unknown differences of beliefs. These problems can be addressed by indicating the lack of understanding through clarification requests, and repeating, paraphrasing or otherwise repairing the original utterance. Through these processes, the common ground is constantly maintained, modified and re-assessed.

A prominent model for grounding was suggested by Clark and Schaefer (1989). They introduced the notion of *contributions* – joint linguistic acts that update the common ground. A contribution consists of two phases, the presentation and the acceptance. During the presentation, the speaker presents an utterance for the hearer to consider. In the acceptance, the hearer indicates whether they understood the mean-

[1] less than 5 percent in American English according to references cited in Ervin-Tripp (1979)

ing of the utterance. Clark and Schaefer list five means for the hearer to indicate that they understood the speaker and that the speaker's action was successful. The first one is *continued attention*, in which the hearer signals that they are continuing to attend. The second is to start the *next contribution* which is relevant to the previous. Thirdly, the hearer can express *acknowledgment* by nodding, uttering a continuer like uh-huh, yeah, okay or an assessment like "that's great". The forth method is *demonstrating* that they understood, by paraphrasing, reformulating or cooperatively completing the speaker's utterance. Finally, the fifth method is called *display* and consists of a verbatim repetition of all or parts of the speaker's utterance.

If the hearer did not hear or did not understand what the speaker said, they signal that, e.g., by looking puzzled or by asking for clarification. Such an expression of a problem in itself is considered the start of the acceptance phase and by clarifying, repeating or rephrasing all or parts of the original utterance, the speaker can proceed with the original contribution. The clarification process in itself is a contribution, too, which is subordinated to the original contribution (Clark and Schaefer, 1989).

A problem with Clark and Schaefer's model is that it is not well suited to computational treatment, as Traum (1999) points out. The main drawback according to Traum is that, given the nested structure, the function of an utterance can sometimes only be analyzed in retrospect, after the status of later utterances have been identified. This makes it hard for a conversational agent to choose an appropriate next utterance during the course of a conversation.

Opposed to that, Traum's approach, as put forward in Traum and Elizabeth (1992) and Traum (1994) is strictly incremental to the extent that each utterance can be assigned a status exclusively based on the course of the previous conversation. Instead of assuming a possibly recursive two-phase structure, Traum's model is based on grounding acts that do not extend over more than one utterance. Furthermore, the model defines a finite set of states and transitions between these states that are induced by the grounding acts. Such grounding acts are initiate, continue, acknowledge, repair, request repair, request acknowledgement, and cancel.

Speech acts and dialog acts

Related to the conceptualization of conversations as a joint action with a certain purpose is the insight that each utterance is not just a proposition about the state of affairs, but an action performed by the speaker. This idea goes back to Wittgenstein (1953/2009), who argued that "the meaning of a word is its use in the language" (Remark $ 43). Austin (1962) went on to analyze the meaning and effect of utterances on three different dimensions. According to his theory, each utterance encodes a locutionary act, an illocutionary act, and a perlocutionary act. The locutionary act refers to the utterance and its particular surface meaning, while the perlocutionary act refers to the effects that the act has on the feelings or actions of the addressee. The illocutionary act associated with an utterance is the act that is performed by uttering a meaningful sentence. The illocutionary dimension of an utterance is what Searle (1969) then conceptualized as the *speech act*. Searle (1976) gives a taxonomy of these acts, dividing them into five classes:

Assertives: Committing the speaker to the truth of the expressed proposition. Examples range from stating, complaining, to boasting and concluding.[2]

Directives: Attempts by the speaker to get the hearer to do something. May range from invitations or suggestions to fierce insistence.

Commissives: Committing the speaker to some future course of action. Examples are promises, plans, vows.

Expressives: Expressing the psychological state of the speaker about a state of affairs. Examples are thanking, deploring, apologizing.

Declaratives: Bringing about a different state of the world by the utterance. Examples are appointing, nominating, firing, resigning.

Searle's work on speech acts was primarily concerned with classifying the effect of a single utterance on the hearer or on the state of the world. Therefore it does not cover some of the phenomena that arise in a the collaborative effort that constitutes a conversation where one turn is highly dependent on another turn. In particular, as scholars such as Traum and Elizabeth (1992) have argued, it is based on a few assumptions that do not usually hold for conversations. One of these invalid assumptions is that each utterance is heard and understood correctly by the listener, who is, according to the second assumption, only a passive recipient and has no part in the plan or action executed by the speaker. A third assumption is that each utterance can only encode a single act. Starting from these limits, Traum and Elizabeth suggested an extension to the early speech act taxonomy, *conversation acts* which, in addition to the core speech acts, addresses conversational phenomena like turn-taking, grounding and argumentation. These are mapped onto four different levels on which to analyze a conversation. The constituents of each level are of different sizes, starting from the turn-taking level, whose components are usually smaller than an utterance to argumentation acts that can span over several utterances.

A similar approach is presented by Bunt (2000), who considers *dialog acts* in their function to update the context along multiple dimensions. He distinguishes between linguistic, semantic, cognitive, social, and physical-perceptual contexts and discusses how dialog acts change these different contexts.

Related to these conceptualizations is one of the most well-known and comprehensive classifications – the Dialog Act Markup in Several Layers (DAMSL) annotation scheme has been developed by the Multiparty Discourse Group in Discourse Research Initiative meetings and was presented in Core and Allen (1997) and Allen and Core (1997). DAMSL is intended to be domain-independent. According to the scheme, each utterance can be annotated with tags of four different layers: the communicative status, the information level, the forward-looking communicative function and the

[2] In his original work he mostly referred to them as "Representatives", but nowadays they are usually cited as "Assertives"

backward-looking communicative function. The communicative status of an utterance indicates whether it is intelligible, interpretable, completed or abandoned, or self-talk, i.e., not addressed at the partner(s). The information level classifies the content of the utterance as being relevant to either the domain task, the task-management, the communication-management, or something else, e.g., jokes, non-sequiturs, or small talk. The forward-looking communicative functions describe the effect of the utterance on the subsequent dialog and interaction and are thus similar to the original speech act classification. Among others, they comprise statements, info-requests, and acts to influence the addressees' future actions or commit the speaker to future actions. Backward-looking communicative functions, on the other hand, characterize how the utterance relates to a preceding utterance. Thus, they encode to what extent the speaker agrees with and understands a previous utterance, and whether it is an answer to a question.

3.1.2 Dialog systems

After looking into some phenomena present in human dialogs that are an important basis for a computational modeling of dialog, we are now going to present the foundations of building computer dialog systems that attempt to provide an interface for humans based on human conversation. We start by discussing the motivation for developing dialog systems and provide a general description of architectures and design features and issues. We then discuss in more detail the functionality of crucial components.

Motivations and applications

Motivations and applications for natural language dialog systems are manifold. A common, underlying goal for many systems is to make the interaction with a computer more natural and human like and thus easier or more fun to use. Apart from that, there are also more practical concerns that justify the use of a dialog system, in particular speech-based ones. There are application contexts in which more traditional interfaces based on visual displays and/or manual operation are impractical, dangerous, or impossible. This applies, for instance, to phone-based systems, or scenarios where users are driving vehicles or controlling other devices, or operate as surgeons. Related to that, speech based systems may also assist users who cannot use other devices due to inabilities. Finally, dialog systems may be used in systems in which the natural language is the only feasible medium to impart knowledge (tutorial dialog systems) or is even in the center of instruction, as in systems that support learning a language.

With the exception of purely conversational systems, most dialog systems serve practical purposes based on some task domain. In this view, natural language is considered as another possible interface alternatively or in addition to traditional user interfaces. A representation of the specific task and application domain of a dialog system must connect to the dialog-specific modules much like the logic of a regular software application must connect to the mouse gestures and dynamic screen content of a graphical user interface (GUI). Task-related knowledge may consist of a database

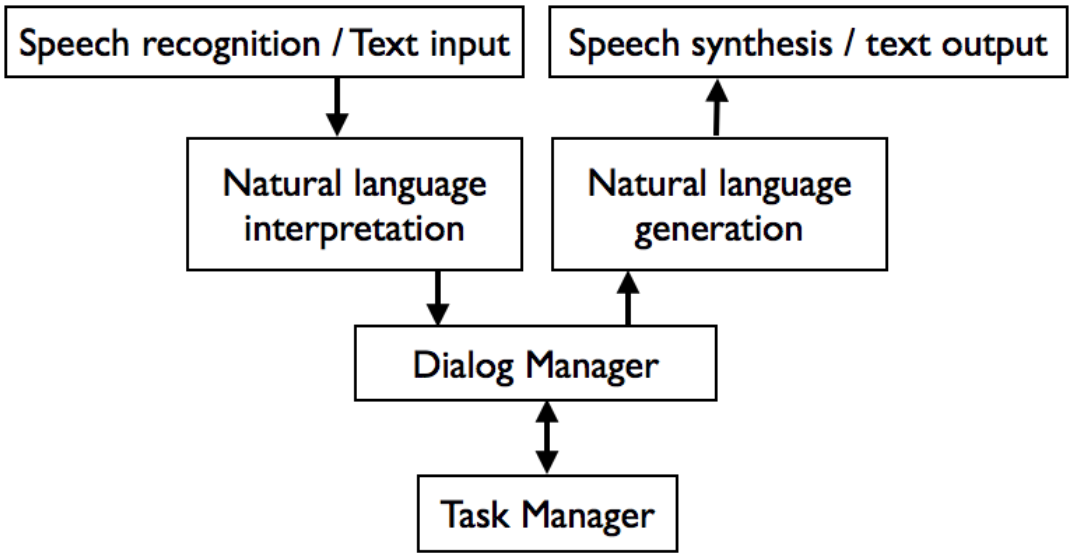


Figure 3.1 – Architecture for dialog systems

for an information retrieval system, map data for navigation systems, rich environmental information for robotic systems, or domain and didactic knowledge for tutorial systems. In most cases, the domain knowledge will be changeable, thus, the dialog system needs to have access to the latest state and also be able to trigger state changes. As a simple example, consider a booking application, in which a sucessful booking leads to the unavailability of the item in question. Depending on the application domain, management of the task can range from a trivial passing through of commands to the back-end application to highly complex models of collaborative multi-agent problem-solving (Allen et al., 2000). Collaborative approaches may also include the attempt to recognize user intentions, which requires more than just the literal interpretation of user utterances (Allen et al., 2001).

Architecture

Across all differences between the variety of dialog systems, there is a common set of components for the universal tasks. Figure 3.1 shows an overview of these components and the information flow between them. End-to-end dialog systems for human-computer conversation require an interface for input and output. The users can either type in their contributions or speak to the system, the latter relies on a module for automatic speech recognition (ASR). Likewise, the system needs an output interface, which can be based on text or speech, the latter requires a module for text-to-speech (TTS) synthesis. Based on the result of the ASR module or the type-written input, the module for natural language interpretation analyzes the input and provides a formal semantic representation of the user utterance. This representation is handed to the dialog manager, which decides how to react based on the current state of the dialog and task-related context. The dialog manager interfaces with the task manager which maintains knowledge related to the task of the dialog system and any relevant context

outside of the central conversation. Based on the dialog state and external context, the dialog manager issues a communicative goal to the natural language generation (NLG) module. The NLG module then is in charge of finding a linguistic realization of the communicative goal and sends it further to the synthesis module or simple text output.

Information flow

While most architectures for dialog systems share these components in one form or the other, they differ with regard to how the modules are connected and how the information flow is organized between them. The processing in simpler architectures follows a pipeline model, in which the information is passed in a linear fashion through ASR/text input, interpretation, dialog manager, generation and text/speech output.

More advanced architectures allow some additional exchange of information in a blackboard style, where each module can consult and contribute simultaneously to a central management component that stores the state of the dialog and external contexts. These approaches are also conceptualized as agent-based architectures, referring to the different modules that work independently but collaboratively (Kerminen and Jokinen, 2003; Ferguson and Allen, 2005). Advantages of these more sophisticated architectures are that they allow for continuous interpretation of user input and are therefore better suited to allow flexible initiative from user and system. Furthermore, they allow for the integration of different independent agents with different types of knowledge regarding the linguistic interpretation, domain knowledge, as well as collaborative concepts like a model of beliefs, desires and intentions (Ferguson and Allen, 2005).

Initiative

Depending on the specific application and task domain, the dialog system will implement a specific policy for initiative, which puts requirements on the architecture. Many systems implement a model which allows either the system or the user to initiate and proceed the dialog, whereas the respective partner only reacts and responds to the initiator's utterances. In system-initiative dialog systems, the system asks questions or makes announcements and waits for the user to respond, while in systems that implement user-initiative, the system awaits the user questions or commands and reacts. More sophisticated dialog systems provide mixed-initiative dialogs where both system and user can initiate in a more flexible manner. Mixed-initiative approaches are more natural but also more complex to implement.

Multiple threads

Natural conversation can comprise multiple topics, or threads, that can be embedded in one another or sometimes even interleaved. Humans usually have little problem managing thread switches. In terms of dialog management, a few approaches have been proposed (Rosé et al., 1995; Larsson, 2002; Lemon et al., 2002; Lemon and Gruenstein, 2004). Often, multiple threads arise out of multiple tasks that the dialog system

and user are pursuing concurrently. The ability to handle multiple threads and tasks increases the flexibility of a dialog system. At the same time, it poses additional demands for the interpretation module and management, since the range of possible user input widens and the system must keep track of the different threads.

Incrementality

Another method of making a dialog system more flexible and faster is the incremental processing of utterances. While the standard approach to treat language is to consider a complete utterance at once and pass it through the different processing steps, it has been proposed more recently to start processing with the smaller units at sub-utterance level. This can increase the reactivity of a system and make the conversation more natural as it is better suited to model phenomena like back-channels, fast turn-taking, self-corrections or collaborative utterance construction (Schlangen and Skantze, 2009). Further, an incremental approach to processing is also more similar to the way the human mind processes language.

Multiple modalities

While dialog systems use spoken or written language as their main modality, additional modalities for input and output are possible and can be useful for different applications. On the one hand, non-verbal channels that play a crucial role in human communication, as for instance, gestures, gaze, or facial expressions can be added. On the other hand, other conventional or novel user interfaces such as GUIs, touch, or body movements can be used to support the processing constraints or other physical constraints of the environment (Wahlster, 2006). Additional modalities increase the complexity of the system and add challenges to the overall processing and integration of all input and output channels.

3.1.3 Components

After presenting the general architecture of dialog systems and some of the relevant issues in more detail, we now describe each of the components of a dialog system in more detail.

Speech recognition

The key factor for *spoken* dialog systems is the quality of the speech recognition module. Speech recognition is the task to translate a raw speech signal into one or more hypotheses of what was said, usually expressed as a string of words, which is then used as input for the natural language interpretation module. This task is usually conceptualized in terms of the noisy-channel model which considers the original utterance to be distorted by some noise along the way with the goal to build a model on how the noise affects the signal in order to recover the original utterance given only the distorted signal.

Speech recognition requires as a first step to digitalize the speech signal that is recorded by one or more microphones. The digital signal is then segmented into

frames of about 10 to 20 ms, and from each frame acoustic features are extracted with the help of signal processing methods. Based on these acoustic features, a number of statistical models are applied in order to estimate the most likely utterance. The models comprise an acoustic model which contains the probability that the given acoustic features are realizing certain phones, further the probability of a sequence of phones realizing a certain word, and finally, the language model, which predicts the likelihood of word sequences in a particular language.

In general, the performance of the speech recognition depends on the size and variety of utterances that should be recognized. If the expected input is small and constrained, the recognition task is simpler than if the expected input is fairly unconstrained. Based on this insight, it is a common strategy to consider knowledge about the current state of the dialog to guide the speech recognition, as certain states make certain utterances more likely than others. Furthermore, the recognition of isolated words as in certain phone command systems is easier and more reliable than recognition of continuous speech. Speech recognition in dialog systems usually deals with speech that is directed at the machine which is different from speech recognition for automatic transcription of human-human conversation. Another parameter is the level of ambient noise in the signal.

Another determining factor for the quality of the recognition is the training data and how similar it is to the actual data. This is particularly relevant for the recognition of non-native speech, since standard recognizers are usually trained on native speech. Tomokiyo (2001) reports on word error rates (WER) between 33 and 75 percent for English spoken by native Japanese speakers, compared to 13 and 21 percent for native speakers. She also shows that the WER is related to the proficiency level of the speaker. Although there are ICALL systems that try to employ a standard recognizer trained on native speech (Morton and Jack, 2005; Anderson et al., 2008), it is usually more promising to adapt to non-native speech. One way is to train the recognizer on non-native speech data. However, given that there are fewer potential sources, it is hard and expensive to collect sufficient amounts of such data. It is even harder if the system is supposed to work with a variety of first languages and levels, since accents might differ considerably. Given these problems in collecting non-native data, there have been approaches to adapt native-trained recognizers based on known regularities about specific accents (Goronzy, 2004), or, in a more general approach, based on the observed differences for a set of different accents (Raux, 2004). For a more detailed account of these attempts, see Eskenazi (2009). Apart from being integrated in spoken dialog systems, speech recognition for ICALL has been also used for pronunciation training and correction in various applications (Eskenazi, 2009). Another, if somewhat dated overview of using speech-based ICALL applications is given in (Ehsani and Knodt, 1998). A recent example of such efforts is the IFCASL[3] project, which aims to provide automated individualized feedback for pronunciation errors. Part of this project is to build a bilingual corpus for French and German with the objective to predict the particular learner errors for the these two pairs of native and learner language (Fauth et al., 2014).

[3]Individualized Feedback for Computer-Assisted Spoken Language Learning

Speech synthesis

TTS synthesis produces an auditory signal based on text input. The process is usually divided into two phases: At first the textual input is translated into a phonemic representation, which is then synthesized as a waveform. There are two different types of approaches to synthesis, one is based on models of the vocal tract, the other is based on the concatenation of prerecorded units (Taylor, 2009). The former, first-generation approaches attempt to generate speech from scratch based on models about how acoustic features of speech arise from the physiological conditions of the human speech organs. A major disadvantage of these approaches is that the voices they produce do not sound very natural. Compared to the data-driven techniques of the second generation, however, they are more economical in terms of memory and processing demands. Nowadays, with the increases in available memory and processing power, concatenative synthesis became more feasible. For these approaches prerecorded speech is chopped up into units of different sizes and then recombined. Their main advantage is naturalness, which makes them particularly suited for ICALL applications. For very constrained domains a simpler approach is to use words as units and concatenate them, in this case, a phonemic representation may not be necessary.

For ICALL applications, speech synthesis is not only used as a part of dialog systems, but also as reading machines (including talking texts, talking dictionaries, and dictation systems) and a pronunciation model for practicing individual or combined sounds (phonemes), prosody, and intonation (Handley and Hamel, 2005). Apart from naturalness, other criteria for the suitability of speech synthesis for ICALL are comprehensibility, intelligibility, choice of pronunciation, accuracy, expressiveness, and appropriateness of register of the synthesized speech (Handley, 2009).

Natural language interpretation

The interpretation of utterances as part of dialog systems serves two purposes. For one, it is the precondition for generating an appropriate response. Furthermore, the content of the interpreted utterance is integrated into the existing knowledge base (Poesio, 2000). In order to achieve this, the linguistic input needs to be related to non-linguistic knowledge of the world. This requires (a) a formal representation of meaning and (b) computational methods that assign a meaning representation to the linguistic user input – semantic analysis.

Interpretation is challenging due to various factors. First of all, for speech-based systems, the result of automatic speech recognition is still not perfect and can lead to incorrect hypotheses to start from. Furthermore, spoken language is characterized by disfluencies like filled pauses, repetitions and corrections. In addition, utterances may be non-sentential, i.e., fragments that are not complete according to traditional grammars but can be resolved in the context of the preceding dialog. Consider for example, expressions such as “when?”, “at the post office”, or “exactly”, which can only be understood in relation to previous utterances. Similarly, referring expressions refer to entities in the context of the conversation and require a representation of the context for their interpretation. Consider deictic markers, like “here”, “today”, “this”, or “you” that refer to the particular spatial and temporal context of the conversation

and to objects and persons that are present. Anaphoric expressions refer to entities mentioned previously in the dialog, for instance the personal pronoun "she" that refers to some female person established previously. The resolution of deictic and anaphoric referring expressions, as well as non-sentential utterances increases the potential for ambiguity, which is a notorious challenge in NLP.

A very simple form of representation relies on extracting meaningful keywords or key phrases from the input and mapping them to system responses (Komatani et al., 2001; Zhang et al., 2007). This can be appropriate for very small and constrained domains, such as controlling devices. An application to control home appliances might spot the words "turn", "light", and "on" within the user input and translate this to a command to switch the light on. Simple keyword spotting may not be sufficient for systems that are supposed to handle more varied input. For such systems, the range of expected user inputs is described by a grammar augmented with information for semantic interpretation.

One common way of integrating semantic interpretation is to design a context-free grammar in which non-terminals directly correspond to the domain-specific semantic concepts. This approach is known as *semantic grammar* and goes back to Brown and Burton (1975). The result of a parse with such a grammar corresponds to a slot-and-frame (attribute-value matrix) semantic representation, in which the non-terminals correspond to slot-names (attributes) and the terminals correspond to the slot-fillers (values). A similar way of integrating semantic information is to add semantic tags to the rules of a context-free grammar. This approach has been realized in various grammar representations for speech recognition (see, for instance, the W3C specification *Semantic Interpretation for Speech Recognition*[4] or the Java Speech Grammar Format (JSGF) [5]). Because it is quite efficient and relatively easy to implement, the approach has been widely used. However, the disadvantage of this method is that its implementation is very domain-specific and therefore not easily adaptable to other domains.

A more general approach is to enhance the syntactic grammar with semantic attachments that specify how to compute the meaning representations of a construction based on the meaning of its constituents, using first order predicate logic and the λ-calculus. For grammar formalisms based on feature structures and unification, semantics can be represented within the feature structures and the composition of meaning as unification equations. An example for grammar-based interpretation is given in Van Noord et al. (1999).

While such a deep semantic analysis is arguably more general and thus less dependent on a particular domain, its development is relatively expensive. Approaches to cut down these costs, while still aiming for independence of a certain domain comprise a shallower analysis of semantics and machine learning techniques to automatically arrive at an interpretation. For semantic role labeling (SLR), which is also referred to as shallow semantic parsing (Gildea and Jurafsky, 2002), semantic roles are assigned to phrases of a sentence relative to a target predicate that invokes the semantic frame (Fillmore, 1976). While the role assignment is an automated process based on statistical learning techniques, it is dependent on annotated resources such as the FrameNet

[4] `http://www.w3.org/TR/semantic-interpretation`

[5] `http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/`

data base that require considerable effort for their construction. Coppola et al. (2009) present examples of successful SRL-based interpretation of spoken dialogs which rests on the English FrameNet database and a smaller domain-dependent database constructed by labeling a corpus of Italian help-desk dialogs. (He and Young, 2006, 2005) present another statistical parsing approach which reduces the dependence on annotated databases further by making do with annotations that contain no syntactic information and can be obtained easily from the associated SQL data base queries or parse results from a semantic parser.

A good overview and more details on semantic interpretation for dialog systems is given in De Mori et al. (2008) and Jurafsky and Martin (2009).

ICALL applications that attempt to interpret learner language need to take into account the nature of non-target like language and may include any of the error diagnosis approaches described in Section 2.3. We will discuss some of those attempts in the context of our detailed discussion of systems below Section 3.2. A very recent effort of parsing spoken learner language is described by Caines and Buttery (2014).

Natural language generation

Based on a communicative goal provided by the dialog manager, the generation module is responsible for finding the best realization of that goal. As in the interpretation step, a variety of methods is available that differ with regard to their flexibility, expressiveness and complexity. Simple approaches rely on canned utterances; slightly more advanced approaches make use of templates that contain slots which are filled with variable fillers. Such simple approaches lack in generality and are usually very application-specific, but have the advantage of easy maintenance. More powerful generation methods rely on syntactical and semantic representations. The generation process can be divided into different steps (Rambow et al., 2001; Walker and Rambow, 2002). In the first step, *content* or *text planning*, the communicative goal is decomposed into atomic subgoals that correspond to single utterances. In a second step *sentence planning*, sentences are planned based on atomic speech acts, by selecting lexemes and syntactic structures. These then feed into the third step — *surface realization*. In this step, function words (e.g., determiners, auxiliaries) are added, word order is determined, and lexemes are inflected according to morphological rules. For systems with speech output, the final step is *prosody assignment*, during which the surface string is enriched with intonation and stress patterns.

For the particular purposes of ICALL applications, the generation module may need to consider the limited vocabulary and knowledge of syntactic structures that learners a different stages might have. Furthermore, it may also consider the preference of particular structures or words that the learner should be exposed to. With a view on corrective feedback given in response to learner errors, the generation module may consider different parameters of feedback and the availability of information about the error, explained in more detail in Section 5.3 and 5.4.

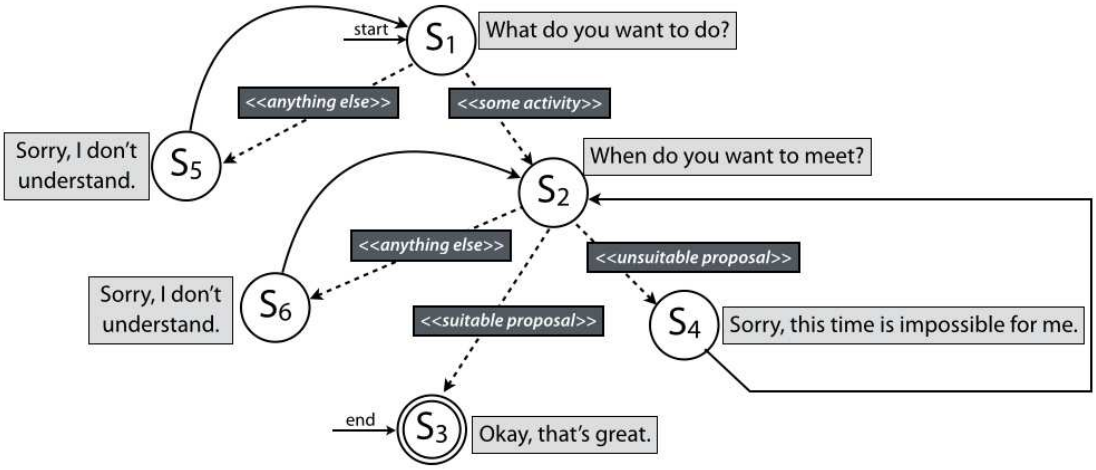


Figure 3.2 – A simplified example of a finite state automaton that models a dialog for making appointments. The labels of the nodes refer to system utterances, the labels of the edges are interpretations of the user response. The solid edges indicate transitions that are executed without conditions, the dashed edges indicate transitions that depend on the interpretation of a user response.

Dialog manager

At the heart of a dialog system lies the **dialog manager**, which is responsible for updating and maintaining the current dialog state and selecting communicative goals based on that state. Updates to the dialog state are usually triggered by results of the interpretation module, but, depending on the architecture, can also be induced by information from other processing modules and the external state and task manager. Similarly, the communicative goal selected by the dialog manager will be passed to the generation component, but there can be other, non-linguistic actions that the dialog manager passes to the task manager or modules for other modalities. A crucial part of the dialog manager is the dialog state representation. In the following section we will discuss in more detail the different approaches to model dialog state and dialog flow.

3.1.4 Approaches to dialog modeling and management

We briefly characterize the four most common models, which differ in their complexity and flexibility, following McTear (2002, 2004) and Jurafsky and Martin (2009). The simpler models are based on *finite-state* technology or *frames*. More powerful and complex are models based on *information state* and *AI planning* techniques.

Finite-state machines

Finite-state based models represent dialog as a network of states and transitions between states. At each state, the system produces utterances, executes domain-related actions, and recognizes user utterances. The interpretation of user utterances or other user actions usually trigger the transition to the next state.

Figure 3.2 gives an example for a simple finite state machine for a dialog model which serves for negotiating appointments.[6] At the beginning (S1) the system prompts the user by asking what activity the date should contain. If the user then responds with a valid date activity, the system transitions into S2 and ask for a time proposal. If the user suggests a time that is suitable for the system, it transitions into S3 and agrees. If the user suggests a time that is impossible for the system, the system transitions into S4 and refuses the suggestion. It then transitions back into S2 and asks for a time proposal again. If the user response cannot be interpreted as a valid response to the first system question about the activity the system transitions into S5, it utters "I don't understand" and transitions back to S1 to ask the question again. Similarly, if the system cannot interpret the user's response to its second question about the time, it goes into S6, signals its lack of comprehension and goes back to S2 and repeats its question.

In addition to this simple model, which contains only the contextual state transitions, there might be universal commands that can be understood at any time, for instance to end or reset the dialog or to get meta information.

The advantage of state-based dialog management is that in any given state, the system only expects a relatively small set of utterances, and sometimes even single words might suffice for arriving at an interpretation and triggering a state change. This very context-dependent interpretation makes these systems relatively robust to mis-interpretation.

However, at the same time, this approach is not well suited for modeling more flexible dialog phenomena, e.g., repairs, unforeseen information, or negotiation. Furthermore, a dialog based on a state machine is relatively restricted as the number of possible user utterances at each state is limited. For example, a system that needs several pieces of information from the user in order to fulfill a service, would prompt for these information bits in a certain order. The user would have to respond to the system's questions in the given order. This is very restrictive and may be inefficient. The user might prefer to provide information in a different order or to provide several pieces of information within one utterance. Even though, in theory, finite state automatons could be designed in order to cover that range of flexibility by adding states and state transitions, the design would be increasingly complex and hard to maintain. An extension to the basic state machine approach is to add variables that store additional values that can be used for generating the next system utterance. Another extension in that nature are statecharts, which make basic state automatons more expressive and powerful by adding hierarchy and concurrency (Harel, 1987). This approach is used for instance by State Chart XML (SCXML) initiative (Barnett et al., 2012).

State-based dialog systems are well suited for system-controlled dialog, where the user reacts to system prompts. They are less well suited to provide user initiative, where the user is in greater control of the interaction. Despite their limitations, these models are widely used in current commercial dialog systems.

[6] While this example may seem a bit odd as a task-based dialog, we chose it because the system developed for this thesis does negotiate appointments with the learner. In any case, it is not entirely inconceivable that a dating platform might require such information to help its users to find other users for particular activities for specific times.

Frame-based modeling

More flexible are frame-based models (also known as form-based or form-filling), which gather task-essential information from the user by filling slots in a template. Unlike in state-based systems, the order in which the slots are filled is flexible. This allows the user to provide input information in different orders and more than one at a time. It is possible to design more complex systems by combining several frames, but then additional methods to recognize and organize switches between frames may be required.

Table 3.1 provides a simplified example for a frame based on the previous appointment dialog. The system, aiming to fill the first slot ACTIVITY, starts the dialog with a question about the desired activity "What do you want to do?". The user then responds with "I want to go for walk tomorrow at 8, for about 2 hours". This response contains not only the activity *walk*, but also the time and duration. Thus, assuming that the interpretation captures the complete content of the utterance, all slot values can be filled at once.

Slot	Question	Response
ACTIVITY	What do you want to do?	walk
START-TIME	When do you want to meet?	tomorrow at 8
DURATION	How long do you want to do it?	2 hours

Table 3.1 – Example for slots, questions, and response instances in a frame-based dialog manager

Information state

An extension to frame-based models is an approach based on the information state of the interlocutors (Larsson, 2002; Traum and Larsson, 2003; Bos et al., 2003). The information state contains dynamic knowledge about what has been said, what can be assumed to be common ground, and what can be done at any state in the dialog. Update rules modify the information state based on the current state and the interpretation of the user input. Update rules consist of conditions which determine if a rule is applicable and the effects which describe the changes to the information state. User and system messages are interpreted as dialog acts[7], which generalize utterances according to the effect they have on the information state.

Figure 3.3 provides a simplified example. Again, the domain is appointment negotiation. The user proposes a time (*"What about Monday at 9?"*), which is interpreted by the system as the dialog act `Suggest`, with the time as parameter. The information state (IS) consists of three variables, storing the suggested time slot, a list of blocked times and the next move for the system to generate. There are three relevant update rules for this example. The first one sets the IS-variable `suggested-slot` to the time that was suggested. The second rule fires if the value of the `suggested-slot` variable

[7]Within the information state update framework dialog acts are traditionally termed *dialog moves*, but there is no conceptual difference between the two terms.

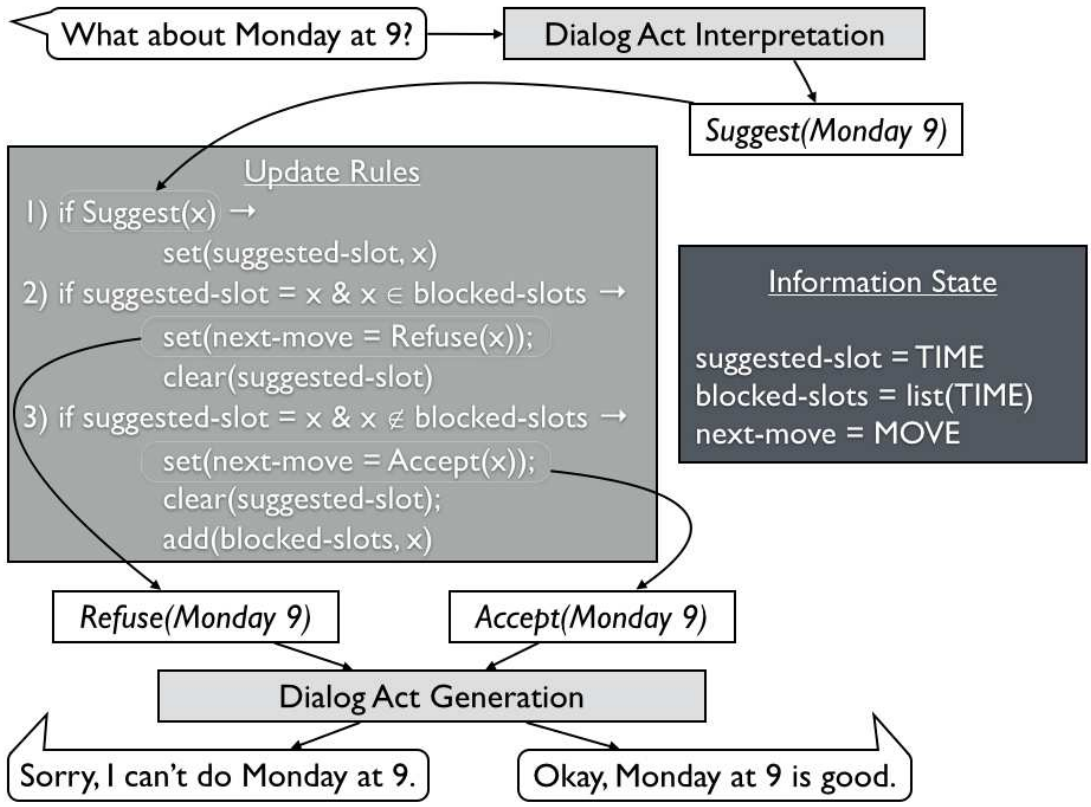


Figure 3.3 – A simplified example of an information state update modeling a dialog for making appointments.

is element of the `blocked-slots`-list. The effect of that rule is to set the next move to `Refuse` with the value of `suggested-slot` as the parameter and then to clear that variable. In contrast, if the suggested time is not one of the blocked slots, the third rule fires and sets the next move to `Accept`, clears the `suggested-slot` variable and adds the time to the list of `blocked-slots`. Depending on the result of the update rules, the system will generate the next move using the `next-move` variable.

The example only shows a tiny, simplified part of the information state and range of update rules. In practice, systems have a larger set of rules, which also necessitates a control strategy for deciding which rules to apply since more than one might be applicable in a given state. Furthermore, the information state can contain much more than domain related variables, but also beliefs, desires and intentions. In fact, the representation of the information state is likely to be more complex than just a set of atomic variables, as the example suggests.

The advantage of the information state update approach is that it can handle more general dialog phenomena that figure in different parts of a dialog. Larsson and Traum (2000) argue that this type of management can even approach the complexity and flexibility of more sophisticated AI-based approaches based on plans and intentions, but that in contrast to those, it is easier to modify because it is more declarative.

Agents, plans, and intentions

The most advanced dialog models employ AI planning techniques. They have been described, for instance, by Cohen and Perrault (1979); Perrault and Allen (1980); Allen and Perrault (1980). They are based on conceptualizing the system and the user as agents which both have beliefs, desires and intentions (BDI) that guide their behaviour in the dialog. Reasoning on those is necessary for the system to interpret and generate dialog moves and to be collaborative. At the same time, the system needs a representation of plans and actions to achieve the goal of the plan. This involves the specification of actions that includes preconditions, effects and a set of partially ordered states that must be reached in the course of the action.

As an informal example, consider again a dialog for agreeing on an appointment. If the user proposes a time slot that is impossible for the system, it must reject the proposal. If the system, however, has recognized the plan of the user to find a mutually agreed time, and is agreeing to that plan, it can be helpful by venturing an alternative time slot after the rejection.

Models based on plans and BDI are well suited for negotiation and problem-solving. However, their power and flexibility comes with an increased complexity and effort to specify the model.

Statistical approaches

All the dialog modeling approaches described above share a reliance on manually built models and representations. Recently, a new approach has evolved which attempts to learn dialog strategies automatically based on actual or simulated user data. Learning techniques range from supervised learning based on the results of Wizard-of-Oz data collections (Hurtado et al., 2005) to unsupervised reinforcement learning (Levin et al., 2000) or a combination of both (Rieser, 2008; Rieser and Lemon, 2011). In comparison with hand-crafted approaches, statistical ones are much better equipped to deal with uncertainties that arise in actual dialog. Besides the imperfection of speech recognition, certain dialog domains come with additional ambiguous sensor data, for instance, in human-robot interaction, which makes it necessary to account for ambiguity. Further, it is often challenging to anticipate actual user behavior when handcrafting a dialog strategy. Statistical approaches provide the flexibility and robustness to deal with unexpected and uncertain input. Furthermore, they offer ways to optimize dialog strategies based on actual data. One problem however, as in most machine learning approaches, is the scarcity of available training data. Another problem is the computational complexity of the learning models (Lemon and Pietquin, 2007). The first problem is often dealt with through the use of simulated or somehow extrapolated user data (Pietquin and Dutoit, 2006; Rieser and Lemon, 2008). The second problem can be addressed through considerable effort in devising the parameters for actions and states on one hand and various techniques to reduce the dimensions of the learning problem on the other hand (Young et al., 2010). However, despite these recent advances, most work so far can only deal with relatively simple slotfilling applications. Lison (2014) proposes probabilistic rules as a framework to combine symbolic and statistical approaches to dialog management and thus remedy the disadvantages

of both.

Within the area of ICALL, dialog systems are, to the best of our knowledge, so far exclusively built with hand-crafted dialog managers. Within their range, systems differ widely as to what type of model is used. In general, however, simple approaches seem to prevail, as we will see in Section 3.2.3. As we will explicate in Chapter 8.1, the dialog system that we developed for the present study is based on a finite state model but includes an additional very simple state presentation that takes account of the preceding discourse history.

3.2 State of the art in existing ICALL systems

3.2.1 Introduction

Following our characterization of the general state of the art in NLP-based ICALL, the description of approaches to error diagnosis in the previous chapter and the introduction of the background for dialog modeling and dialog systems in the preceding part of this chapter, we now illustrate the current state of the art for ICALL systems that include some form of interaction and feedback by describing a selection of specific systems and their approaches. We begin with a general characterization of the requirements and challenges that ICALL systems have to address, give a general overview, and briefly characterize commercially available systems before we detail the variety of research prototypes that have been developed.

Requirements and Challenges

Compared to the challenges for task-based dialog systems targeted at native speakers, interacting with learners comes with additional requirements. Learner language often contains particular errors that are more frequent and of different nature than errors of native speakers. These errors reflect the learning process. To recognize them, to incorporate them in the interpretation, and to provide corrective feedback is the main challenge of ICALL systems. Learner errors can make interpretation more difficult because they can increase the ambiguity. On the other hand, learner language is often simpler and more limited than native language. This can ease the task of language interpretation, since the language resources have to cover less. At the same time, it is also an additional design challenge to make sure that the vocabulary and syntactic structures used for system productions are appropriate for the targeted learner level. Giving feedback to learner language creates an additional thread in the dialog that has to be managed in relation to the content matter dialog thread.

Depending on the purpose of the dialog system, the content matter should be relevant and useful for the learner. In most cases, ICALL dialog systems will model a domain and task for the sake of practice, but in some cases, a real-purpose dialog system is adapted to non-native speakers (Raux and Eskenazi, 2004a,b)

Overview

The selection of the systems we present in this section is based on their relevance to the work undertaken in this thesis. We describe them under the aspects of input they expect from the learner and how this input is constrained, the error diagnosis and feedback they provide, the evaluation they have been subjected to, and the pedagogical theories they were informed by.

In general, a large part of publications on ICALL applications concentrate on descriptions of the system and the interaction it allows but do not include any evaluation. If the systems are evaluated, this is often done in terms of their performance, i.e., the amount of errors they make (Seneff et al., 2004), or in terms of usability by means of questionnaires given to the users (Wang and Seneff, 2007; Lech and Smedt, 2006; Johnson and Wu, 2008). Only a small number of ICALL publications include an evaluation of the language development or learning gains that their application can induce (Zhao, 2003). For an even smaller number of applications the learning gains are compared with alternative teaching means or across different parameters of the application. Related to that, publications on ICALL applications only rarely make explicit reference to theories of second language acquisition. If they do, it is usually with the purpose to justify design decisions, but not in order to investigate the validity of specific SLA theories. We will point to exceptions to this rule below. We will first summarize the state of the art for off-the-shelf systems that are available to private users and then look in more detail at systems that have been developed within research contexts.

Off-the-shelf Systems

Commercially available ICALL applications usually focus on exercises related to new vocabulary and grammar rules. The learner input is constrained and systems are not geared towards free communication. If they contain any dialogic material it is used as a means to impart new language content, i.e., lexical items and grammatical structures, similar to monologic lesson texts, rather than as a way to engage the learner in a conversation (e.g., "ActiveChinese" (Chiu, 2008) or "Side by Side Interactive" (Statan, 2006)). If called for at all, participation of learners is limited to advancing the presentation of the dialog by clicking a button. Sometimes, learners can choose one out of a set of semantically equal options. In another variant of this task, learners can order a set of utterances to render a meaningful dialog. None of these systems allows free input to engage in a dialog. Some systems allow the user to record pronunciations of textual prompts and then give feedback about the quality of the pronunciation (Lafford, 2004; Chiu, 2008). In "Tell Me More" (Lafford, 2004) learners engage in a dialog by choosing an appropriate response from a set of three given candidates and then pronounce their response. The system's speech recognition component then gives feedback about the quality of their attempt. Given that the learner input is very constrained, usually to multiple-choice questions, as described above, or fill-in-the-blank activities, the requirements for the the error diagnosis and feedback facilities are rather simple. In the simplest case, the system merely states whether or not the response was correct, which requires a simple comparison with the target response.

Another system that is open to the public and can be used for language learning is

the telephone-based bus schedule information system *Let's Go* in Pittsburgh (Raux and Eskenazi, 2004a,b). Although its original purpose and development was not geared to support language learning, it has been extended to cater for non-native speakers and it can give them implicit corrective feedback if their input deviates from the expected input. The application is meaning-based since its actual purpose is the real-life task of obtaining schedule information, but therefore its domain is very limited.

Since commercially developed off-the-shelf systems are rarely the subject of scientific publications, it is not surprising that there is a concomitant lack of evaluation of these systems, in particular regarding potential learning gains.

Unsurprisingly, there is a considerable gap between off-the-shelf systems and systems developed in research contexts as described in the literature. In general, research prototypes provide more freedom in input, richer communication and more informative feedback. However, since the systems are rarely accessible to the public, these claims are in general not verifiable. Research-driven systems are usually only available to a rather restricted number of learners, and in this context they are primarily used for the purposes of testing and further development. These systems can be roughly divided in those that support learning by providing distinct, often grammar-related exercises (Section 3.2.2) and those that support learning by engaging the learner in a dialog and meaning-based communicative interaction (Section 3.2.3). Although some systems include both aspects, one of them usually predominates.

3.2.2 Systems with a focus on grammar

The three systems described in this section offer a collection of exercises and provide detailed feedback on form-related errors. They have been used and tested within foreign language programs in universities.

E-tutor

The E-tutor system (previously known as German tutor) has been developed by Trude Heift and colleagues at the Simon Fraser University in Canada (Heift and Nicholson, 2001; Heift, 2003, 2004, 2010a). It is used by students of German as part of their regular language classes and covers the content of the first three beginner courses. In addition to texts that introduce the topic and grammar structures of each chapter, the core of the system consists of exercise activities. These exercises comprise listening and reading as well as writing tasks. For specific grammar-focused tasks the system is able to generate automatic feedback. These exclusively text-based exercises are gap-filling, sentence building, translation, and dictation. The feedback is implemented through a combination of generic non-linguistic matching algorithms and a linguistic analysis using constraint relaxation. The generic error module identifies spelling errors, missing or superfluous words, and incorrect word order by comparison to the set of correct answers. The NLP-based module diagnoses grammatical errors based on the syntactic analysis of the learner answer (using the *Head-driven Phrase Structure Grammar* (HPSG) formalism (Pollard and Sag, 1994)). It can identify agreement errors, e.g., mismatches between subject and verb or unsatisfied case requirements of verbs and prepositions (more details are given in Heift and Nicholson (2001); Heift (2003)). Feedback mes-

sages are explicit and provide different amounts of information that are specific to the level of the learner (see for more details Section 5.5.2). The activities in E-tutor that provide automated feedback do not allow free input. The developers argue that their goal is high accuracy of feedback which would be impossible to provide reliably for unconstrained input Heift (2003). The system is evaluated in terms of accuracy of the feedback it provides. In addition, learner errors and learner behavior in response to different types of feedback have been studied extensively, as described in more detail in Section 5.5.2 and Heift (2001b, 2004, 2010b). Heift (2004) refers to SLA feedback studies and the value of interaction and noticing (Section 4.2.3) as SLA principles and motivations for the system.

Robo-Sensei

Robo-Sensei is a system for learning Japanese, developed by Noriko Nagata at the University of San Francisco (Nagata, 2002, 2009). It covers grammatical structures that are contained in a standard 2- to 3-year Japanese curriculum. It is intended as a supplement to a text book, and its core consists of sentence production exercises. Learners are provided with a communicative context in English and an English paraphrase of what they should produce in Japanese. The system then provides immediate feedback to the learner response. Although the task is embedded in a real-life scenario, the sentence to be produced by the learner is not part of a larger dialog and the learner utterance is very much constrained by the English prompt that is to be translated.

The error diagnosis and feedback is based on a linguistically informed comparison between the correct answer and the learner answer. The linguistic analysis employs word segmentation, morphological and syntactic analysis and errors can be diagnosed at each of these levels. The error diagnosis can identify unknown, missing and unexpected words, modifier errors, word order errors, and predicate form errors, which include tense, negation, style, and auxiliary form errors. The feedback is explicit and very informative as it indicates not only the location of the error but also provides an explanation of the grammar rules that were violated. Some common spelling and conjugation errors are anticipated and handled in the morphological analyzer. Other errors are recognized by matching the syntactic structure of the correct target response with the syntactic structure of the actual learner response. In this way, errors are diagnosed through recognizing the difference to the model response, which can be considered as one instance of pattern-matching approaches (see Section 2.3). This means that errors are not anticipated explicitly, but, since the possible mismatches are identified related to very specific phrase structure rules, the feedback messages contain detailed information about the nature of the rule violation. The system, and in particular the learning effect of the feedback it provides, have been thoroughly evaluated (Nagata, 1993, 1997). We will summarize the results of this in more detail in Section 5.5.2.

TAGARELA

The Teaching Aid for Grammatical Awareness, Recognition and Enhancement of Linguistic Abilities - TAGARELA(Portuguese for "talkative") was developed by Luiz Amaral, Detmar Meurers, and colleagues at the Ohio State University (Amaral, 2007; Ama-

ral et al., 2011). It is conceptualized as an "electronic workbook that offers on the spot individualized feedback on spelling, morphological, syntactic and semantic errors" for learning Portuguese (Amaral and Meurers, 2011, page 14). The system provides listening and reading comprehension, picture description, rephrasing, fill-in-the-blanks, and vocabulary tasks as exercise activities. The linguistic analysis of the learner input comprises tokenization, spell checking, morphological analysis, lexical lookup and disambiguation for lexical information, bottom-up chart parsing based on a small custom-built grammar, and semantic interpretation based on shallow matching strategies. The feedback given by the system depends on the type of activity, which entails different kinds of learner input. Feedback for reading and listening comprehension and description tasks is meaning-based, while the rephrasing task provides feedback about syntactic errors. Vocabulary exercises, which expect a noun phrase as response, and gap filling exercises involve feedback about morphological or lexical errors. The work on TAGARELA is based on a number of SLA concepts, that we will discuss in the next chapter – task-based instruction and FOCUS-ON-FORM. The evaluation of TAGARELA is limited to small-scale usability studies and the observance of some specific problems for feedback efficiency (Amaral and Meurers, 2009). However, until now, there has been no principal evaluation in terms of learning gains that the system can support.

Summary

We have described E-tutor, Robo-Sensei, and TAGARELA as examples of systems that offer relatively focused and well-defined exercise activities and detailed feedback on form-related errors. This feedback is enabled by a combination of several steps of linguistic processing which at least comprise morphological and syntactical analysis. These systems are relevant for this thesis because they illustrate the state of the art in form-related feedback, and in the scope of this thesis, we will examine the effect of different types of such feedback. Since, for our study, we plan to provide feedback in the context of communicative interaction, we will now describe ICALL systems that focus on communicative activities in a meaning-based context.

3.2.3 Systems with a focus on communication

In this section we introduce systems that have a focus on communication and use some form of dialog as their primary means to impart new knowledge and provide practice. The systems differ (a) regarding how much freedom the learners have in contributing to the dialog and (b) regarding the amount and quality of feedback they obtain. In some systems, learners can merely choose one response from a given set, in others they are completely free to produce whatever they want. Some systems have rich expectations about form-related errors that learners might make and provide detailed, informative feedback, while other systems intentionally ignore any errors in the learner input. Published interactive systems further vary with regards to their domain, the input modality, the range of linguistic structures they practice, and the extent to which they put focus on those. They also differ in the number of involved conversational agents, the embodiment of those agents, the sophistication of the graphical

interface, and the specifics of the target group they were built for.

Chat bots

One popular, if low-tech class of applications for human-computer interaction are chat bots (or chatter bots) that communicate with humans in text-chat mode. They date back to the 1960s, with the most prominent example ELIZA, which simulates a psychiatrist (Weizenbaum, 1966). These chat bots were based on rather simple pattern matching algorithms to generate a response. Despite the simplicity of the underlying algorithms, these chat bots appeared to maintain a coherent conversation and humans could spend hours engaging in conversation with them. They had been built in an attempt to pass the Turing test, i.e., to display conversation behavior indistinguishable from human behavior (Saygin et al., 2000; Shieber, 2004). PARRY, a system of the same time, pretended to be paranoid and his behavior was actually indistinguishable from that of real human paranoia patients for a group of psychiatrist judges (Colby, 1975, 1981; Dennett, 1998). However, both ELIZA and PARRY exhibited a rather peculiar behavior, which is arguably entertaining and engaging, but probably easier to simulate than normal human behavior.

With the inception of the Loebner prize in 1990[8], which honors systems that attempt to pass the Turing test, the development of chat bots has picked up again. While to some there has been surprisingly little progress since PARRY and ELIZA (Wilks and Catizone, 2000), others do see considerable development (Coniam, 2008). However, many of the current chat bots are still based on the relatively simple pattern matching approaches and do not attempt a linguistic modeling. None has passed the Turing test yet. Coniam (2008) investigated the suitability of current state-of-the-art chat bots for language learning. Apart from one bot that proposed corrections for some ungrammatical utterances, most others were unable even to cope with spelling errors, let alone grammatical errors. Coniam's conclusion is that current chat bots are still not really suited as conversational practice tools for second language learners.

We will now describe communicative systems that were specifically developed for the purpose of supporting language learning. The first group of systems constrains the learner's input by providing a small set of options to choose from (E-daf, Let's Chat, CandleTalk, and Conversim). The feedback in these systems relates to content or pragmatic problems, and it is mostly implicitly provided through the reaction of the virtual conversation partner. Since all options are grammatically correct, there is no need for form-related feedback. The second group of systems allows the learner to freely produce their input and gives semantic feedback (MILT and TLTS) and form-related feedback (SPELL and Te Kaitito). The systems differ widely regarding the evaluation that they have been subjected to.

E-daf

[8] http://www.loebner.net/Prizef/loebner-prize.html

Chan and Kim (2004) describe a comprehensive ICALL system for learning German – "e-*daf*". Besides relatively decontextualized grammar activities (gap-filling, multiple choice, drag-and-drop exercises, etc.) it also includes more interactive dialog activities. However, in this activity the learner cannot freely produce their contribution to the dialog, instead they can choose one of three possible responses, which are all correct but different. Learners thus co-construct the dialog, and after completion they can review the complete dialog. Apart from this, learners are not required to attend to any formal aspects (they do not have to apply their grammatical knowledge), which is intended by the creators. The exercise is created to allow the learner to focus on meaning and discourse and to allow the learner to actively participate. The e-daf system also provides free response activities like open-ended writing tasks and web-chats, but the feedback to those is not provided by the system, but by peers, native speakers, or teachers. To our knowledge, there is no published evaluation of the e-daf system.

Let's Chat

Stewart and File (2007) describe a chat system that relies on communication through previously stored utterances, without using any NLP. It targets communicative skills in the area of introductory social conversations, a topic which is supposedly disregarded in classroom or other CALL applications. The learner can choose from pre-stored utterances and the system replies with appropriate pre-stored utterances, which are spoken and provided as text on the screen. While the learner has no obligation or opportunity to create free input, the authors argue that "the holistic assimilation of formulaic sequences and their frequent rehearsal" is important and beneficial for language acquisition (Stewart and File, 2007, page 101). Since the user merely selects one out of a set of pre-formulated grammatically accurate responses there is no need for form-related feedback. The only error learners can make is to select an inappropriate response, in this case, the system will respond by giving the advice to "choose again". Otherwise, it is tolerant regarding slightly odd responses for the sake of sustaining the communicative flow. To our knowledge, the system has not been evaluated.

CandleTalk

The CandleTalk system developed by Chiu et al. (2007) presents a collection of speech acts (greeting, parting, apologizing, requesting, complaining, and complementing) embedded within authentic dialogs. In order to participate in the dialog, the learner is supposed to select one of the available continuations and pronounce it. The continuations are semantically different, and the dialog will enfold differently according to the learner's choice. Since the focus of this activity is on the acquisition of pragmatically and socially acceptable speech acts, the options also contain inappropriate responses. Feedback on the appropriateness of the learner's choices is given as a summary only after all the dialogs of a unit have been worked on. The dialog only proceeds if the speech recognizer is successful in recognizing, but no explicit feedback on pronunciation errors is given. A native model pronunciation for all utterances is available for reference. The system has been evaluated in terms of user satisfaction and learning gains. Working with the system had a positive effect on the ability to use speech acts

appropriately but no recognizable effect on pronunciation accuracy.

Conversim

Conversim has been developed by researchers at Interactive Drama Inc. (Harless et al., 1999, 2003). It allows the user to engage in a dialog with a video avatar that is based on a real person. The system has been targeted at learners of Arabic with intermediate proficiency and provides them with the opportunity to practice and refresh their knowledge. The dialogs are motivated by a problem that the learner has to solve by obtaining information from the virtual interview partner. The dialog is scripted, which means that the learner can not freely produce their contribution but has to choose from a given set of options. These options are presented on the screen, and the learner is supposed to speak them literally or paraphrase them. Transcriptions and English translations of the character's responses are available to the learner at any time. The system has been employed and tested extensively in cooperation with the US armed forces. The developers measured gains in speaking, listening comprehension, and reading skills resulting from extensive, independent dialog with virtual native characters. Their nine subjects, members of the army, were required to use the system for one week for at least six hours a day. The pretest-posttest comparison showed a significant increase in reading and speaking skills. Listening skills increased too but not at a significant level. One objection to this evaluation is that there was no control condition to compare it with, which means the learning gains can only indicate that the system is successful, but there is no information on how its success relates to other learning material. With a perspective on the present study for this thesis, we should note that the intensity of the treatment – six hours a day for one week – may be hard to replicate in other contexts, since it is difficult to recruit subjects that are available for such long time spans.

Military Language Tutor

The Military Language Tutor (MILT) was a system developed by the Army Research Institute for training US soldiers in Modern Standard Arabic (Kaplan and Holland, 1995; Kaplan et al., 1998; Holland et al., 1999). The interaction is set in a simple 3D virtual microworld in which the learner can control a virtual agent through written and spoken commands. The commands correspond to a fixed set of possible actions targeted at objects within the microworld. The goal of the interaction is defined by a problem to be solved, for instance, "Where will the enemy attack?" There are two versions of the system, which differ in the mode of input they allow: text and speech. Reflecting the state of the art in speech recognition at that time, the speech-enabled system only allows the learner to read pre-defined sentences. The authors do not specify what kind of feedback the learners get in response to mis-pronounced utterances. In the text-based system, the learner can formulate their input freely, but the system is only able to understand commands related to the objects in the scene. Further constraints on the input are available to the learner in a help window. In Kaplan et al. (1998), the developers claim to use syntactic and semantic analysis for providing meaningful feedback to errors but they refrain from providing any details. Holland et al. (1999)

admit that these methods lacked the necessary robustness and were therefore replaced with a simple keyword matching approach. Feedback on the learner production is given implicitly through reaction of the virtual character – it behaves as intended if the command could be interpreted, otherwise, the character acts unexpectedly or says "I don't understand". Accompanying the microworld tasks was a familiarization lesson which provided all 72 commands that were available for the task as text and as a sound clip, spoken by a native speaker, plus their translation. The authors argue that the context of the tasks allows the learners to pursue an interesting goal, which supposedly motivates them intrinsically to work with the system long enough to approach automaticity of their language skills. Apart from assessing the user acceptance levels and attitudes towards their system, the developers of the MILT system also assessed the learning gains measured through pretest-posttest differences in the subjects' basic sentence-building skills. Participants were asked to translate 72 English sentences into Arabic – half of them as pretest and the other half as posttest. Each of the items was rated by a native Arabic speaker on a 5-point scale along the following four dimensions: vocabulary, grammar, pronunciation, and overall fluency. The test sentences were taken from the system's repertoire, implying that the participants were familiar with them through the interaction with the system. The participants were 16 soldiers who had different amounts of prior knowledge of Modern Standard Arabic. Each participant worked for one hour with the system. The difference between pretest and posttest was significant for 14 of the 16, although in absence of a control condition, it is not clear whether learning gains resulted from working with the system or possibly through exposure in the pretest alone.

Tactical Language Training System

The latest, and arguably most advanced system in the domain of military training is the Tactical Language Training System (TLTS) developed at the Information Sciences Institute at the University of Southern California (Johnson et al., 2004b,a; Johnson and Valente, 2009). The system was developed for Arabic, Persian, and other languages relevant for the US American armed forces. It is self-contained and teaches non-verbal behavior and cultural knowledge in addition to language skills. The system contains two complementary parts – the *skill builder* and the *practice environment*. In the beginning, the skill builder provides focused exercises that are used to impart new knowledge. For this constrained environment, it further provides individual feedback on pronunciation and grammar. The skill builder is thus comparable to the form-focused systems described Section 3.2.2. The practice environment is a virtual world with 3D landscapes and animated characters with whom the learner interacts. The scenarios are placed in local villages where learners have to interact with local people in order to pursue their mission. In this game-like environment learners can practice what they learned previously with the skill builder. Learners are supported by a pedagogical assistant character who offers hints that help to forward the game; the hints are specific to the stage and knowledge of the learner. Apart from these implicit directions, the learners' production is unconstrained. The feedback provided in the practice environment is meaning-based – if the learner's utterance is unintelligible or inappropriate, they will not be understood by the villager character.

The system was repeatedly evaluated in different ways as reported in Beal et al. (2005); Johnson and Beal (2005); Johnson and Wu (2008); Johnson and Valente (2009). The primary goal of evaluation was to improve the evolving system, in particular the speech recognition module. Evaluations regarding learning gains were kept very general, either characterized according to a general proficiency level (e.g., ILR[9] proficiency level of 0+ after 40 hours of training) or even more holistic as in "The marines who trained with Tactical Iraqi were able to perform many communicative tasks on their own, without reliance on interpreters. This enhanced the battalion's operational capability, enabled the battalion to operate more efficiently, and resulted in better relations with the local people" (Johnson and Valente, 2009, page 82). The system was not compared with alternative teaching methods or materials.

Only the small-scale study described in Beal et al. (2005) attempted to compare the effect of different system parameters. The goal was to assess the value of individualized feedback on pronunciation and the value of engaging in an interactive, meaning-based virtual game. Based on these two parameters – (a) the exposure to feedback and (b) the participation in an interactive game, four experimental groups were compared. The group that received feedback but did not participate in the game outperformed all other groups. The group that received feedback and participated in the game performed disappointingly, which was attributed to technical problems and the fact that they spent less time with the core of the tutorial on which the posttest was based. Unfortunately, the small number of participants (5 for each group) make the results somewhat inconclusive. However, to our knowledge, this was the only evaluation for TLTS that focused on different properties of the system and compared their effect on learning, as opposed to simply collecting usability gradings and a very rough estimation of learning progress.

SPELL

The SPELL (Spoken Electronic Language Learning) system developed at the university of Edinburgh provides a virtual world and animated characters with whom the learner interacts (Morton and Jack, 2005; Anderson et al., 2008; Morton et al., 2008). The scenarios for real-time conversations are based on real-life experience that are useful to the average learner, e.g., at a café or at the train station. The system was implemented and tested for Japanese and Italian as a second language. Informed by the INTERACTION HYPOTHESIS (see Section 4.5) it provides opportunities for the learner to modify their initially erroneous input, and it reformulates the output in case the learner indicates incomprehension. The speech recognizer is based on a grammar that explicitly models anticipated learner errors, supposedly by mal-rules although the exact formalism is not specified. The system gives implicit feedback using *recasts* – corrective reformulations embedded in the dialog flow (different feedback types are explained in more detail in Section 5.3).

Learning is organized in three levels. In the first level, the learner observes a sample dialog performed by two virtual characters without actively participating. In the second level, the learner is introduced to new vocabulary and grammar structures, by

[9] Interagency Language Roundtable http://www.govtilr.org/skills/ILRscale1.htm

interacting with a conversational agent who poses questions related to the scenario and gives feedback on errors. In the final level, the learner is immersed in the 3D virtual world and interacts with several characters according to the scenario. At this level, the learner's utterances are used to control the world, for instance, ordering a meal, will cause a waiter character to serve a meal within the 3D world. The system allows free input from the learner. Vocabulary, grammatical and cultural information is available at any level.

The SPELL system was evaluated in terms of usability and regarding the performance of the speech recognition module. The system's recognition was far from perfect – the accuracy for word-for-word recognition for utterances covered by the grammar ranged from 56% for Italian to 72 % for Japanese. However, the system's accuracy for meaning recognition was slightly higher: 66% for Italian and 79 % for Japanese. This difference is not surprising because one meaning can usually be realized by several different utterances that have a similar surface form. The learners' judgment regarding usability indicates that the system was engaging and fun to use despite its obvious failures. There was no evaluation regarding learning gains.

Te Kaitito

Another system that combines a communicative approach with form-related feedback is Te Kaitito for teaching Maori. It was developed in New Zealand at the University of Otago by Alistair Knot, Peter Vlugter and their colleagues (Knott et al., 2003; Vlugter et al., 2006; Knott and Vlugter, 2008; Vlugter et al., 2009). Unlike SPELL, it only handles written language and allows no speech input. It is bilingual in the sense that it engages the learner in a conversation in the Maori language, but provides metalinguistic explanations in English. Similar to SPELL, the authors refer to the INTERACTION HYPOTHESIS as the theoretical base of their work (Knott et al., 2003). The system is targeted for the beginner level of Maori learners and therefore covers only a small vocabulary of 381 words and a limited range of grammatical forms. The interaction is organized in lessons, which are associated with a set of grammatical forms which the learner is supposed to learn during that lesson. The system then makes use of these forms or tries to elicit them. The dialog is mixed-initiative insofar that system and learner can both start a new topic.

The interpretation of the learner input involves syntactical parsing based on the HPSG formalism and implemented by the *linguistic knowledge building* system (Copestake, 2002). The parser returns a semantic representation in form of *minimal recursion semantics* (MRS) (Copestake et al., 2005) which is then interpreted as a dialog act and represented as a discourse representation structure according to discourse representation theory (Kamp and Reyle, 1993) which updates the current discourse context.

The error recognition and correction suggestion are described in Knott et al. (2003) and Vlugter et al. (2006). In early versions, errors are modeled by special error grammars according to the mal-rules approach. Feedback is given in the form of metalinguistic explanation like: "Remember that objects must be introduced with *i*". In later versions, a new approach to error recognition is introduced. This approach is based on generating alternative variations of the actual utterance, so-called perturbations, that differ on character or word level. The perturbations are ordered according to their

likelihood and only the most probable are considered. In case the original utterance cannot be parsed, or its interpretation is hard to align with the given dialog context, the interpretations of the perturbations are considered and used as hypotheses about the intended production. They are then used for clarification questions or corrections of the form "I think you mean X".

One of the later versions of the system is extended to implement multiple dialog participants in order to teach Maori pronouns (Knott and Vlugter, 2008). In this version, the system can assume the role of two dialog participants. This system is evaluated in terms of how it affects the learners' knowledge about Maori pronouns (Vlugter et al., 2009). The learning gains are compared to those of learners who received regular teacher-based instruction and a control condition of learners who received no additional instruction. They show that learners who were tutored by the system performed comparably to the learners in the regular instruction in an immediate written test. However, in a delayed posttest one week later, the system group scored less well.

Summary

In this section, we presented different systems with a focus on communication. These systems let the learner participate in a natural dialog and the focus of the interaction is usually on meaning. Only two systems provide form-related feedback, but one of them, SPELL, does so in an implicit way that does not disturb the flow of conversation. Other systems provide feedback regarding content and pragmatics. The systems have been evaluated in different ways, including usability measures, technical performance, and, sometimes, learning gains. Of the systems that were evaluated in terms of learning gains, all but one were tested on their own, and not in direct comparison to alternative means of language instruction.

In this thesis, we want to assess the value of communicative interaction and the effect of different instructional settings, therefore, in contrast to the majority of previous ICALL work, we will employ a comparative test setting, in which different parameters are compared with each other. The system that we will employ for this study will integrate the communicative approach with form-related feedback.

3.3 Summary

This chapter provided the second portion of the technological background for this thesis by covering the linguistic and computational premises for modeling and processing dialog and portraying the existing approaches to providing foreign language instruction in an interactive way.

The first part of this chapter provided the background for computationally modeling dialog. It started off with a description of essential phenomena in natural conversation. We argued that a central premise in our understanding of dialog is that it is a collaborative activity. This explains how dialog is composed into a sequence of turns and how interlocutors constantly attempt to ensure mutual understanding in a grounding process. Furthermore, it is the basis of understanding how utterances can be classified as actions performed by the speaker. This part continued by introducing basic con-

cepts of dialog systems, comprising application contexts and features of architectures. It then described in more detail the common components of dialog systems and their purpose. This part finished with a characterization of the most prominent approaches for modeling dialog, including finite-state machines, frames, information state, agent and plan-based, and statistical approaches.

The second part of this chapter then provided a detailed summary of existing ICALL systems that provide feedback and dialog. We distinguished between systems that focus on grammar and systems that focus on communication. We characterized the systems with a perspective on the input they expect from the learner and how this input is constrained, the error diagnosis and feedback they provide, the evaluation they were subjected to, and the pedagogical theories they were informed by. We showed that only very few systems have been evaluated in terms of learning gains they enable, which is in contrast to the approach we pursue in this thesis.

After we spent the last two chapters expounding on the background of this study in relation to the linguistic and computational modeling of conversational interaction and error treatment in an ICALL context, we will use the next chapter to provide the essential background on theories and concepts of second language acquisition that inform the work on this thesis. The chapter after next will then combine both perspectives by focusing in more detail on the issue of feedback.

4

Second Language Acquisition

4.1 Introduction

The goal of this chapter is to discuss basic concepts, theories, and issues from the research area of second language acquisition (SLA) and thereby to provide the necessary background for our study from this perspective. The goal of research in SLA is to understand and explain the processes that govern non-native language acquisition. In general, it is desirable to apply findings of SLA research to the design of teaching materials and to teaching methods within classrooms. However, this transfer is not always smooth and straightforward, because there is a considerable gap between the context of theoretical SLA research on the one hand and the constraints of practical classroom pedagogy on the other hand. In essence, the underlying goal is to find the best, that is, the most efficient and most convenient, albeit realistic and feasible, methods for language instruction. *Instruction* is commonly understood as pedagogical guidance given to the language learner (Housen and Pierrard, 2005a), usually by an instructor in a classroom (Ellis, 1986). The instructional learning context is usually framed in opposition to *naturalistic* acquisition, in which learners acquire the second language through communicating spontaneously in authentic social situations, i.e., by living and acting in the second language context (Housen and Pierrard, 2005a). The contrast between naturalistic and instructed acquisition is related to two much debated issues that we will discuss in more detail in this chapter. The first issue concerns connections between form and meaning – "the essence of language" as DeKeyser (2007b) calls them. We will look into how instruction can establish these connections and how different kinds of instruction differ with respect to the weight they give to either meaning or form in Section 4.2. The second issue concerns the difference between implicit and explicit types of instruction, the respective learning processes they induce, and the nature of the resulting linguistic knowledge. In Section 4.3 we will discuss these differences and the role of implicit and explicit knowledge for language proficiency.

For practical reasons, research studies in the discipline of SLA often focus on very specific phenomena. Following a common approach in SLA research, for the scope of this study, we picked out certain linguistic forms – the so called *target structures* – and set up our experiment around these structures. Although the implications of the outcomes of a study are usually supposed to extend beyond the small scope of the target structures, it is important to note that linguistic structures can differ considerably from one another and are not equally well suited for different types of instruction and experimentation. In Section 4.4 we will discuss the properties of potential target structures and their effect on learnability and instruction.

We finish the chapter by discussing the role of *communicative interaction* for the acquisition process, how interaction can connect meaning and form and how it relates to the difference between implicit and explicit learning (Section 4.5). There, we will also present a teaching approach that uses *tasks* as a means to encourage interaction and to establish a focus on meaning.

4.2 Form and meaning in language instruction

The chief goal of second language instruction is to create proficiency in learners. The manner in which this goal is best achieved, however, is far from clear and has been subject of debate for decades. One reason for the dispute is that second language (L2) proficiency comprises different aspects that are potentially competing with each other. A common, widely accepted view is that proficiency can be described by the three dimensions of accuracy, fluency and complexity (Skehan, 1996b; Housen and Kuiken, 2009). *Accuracy* is understood as the formal correctness of the produced language and the ability to produce error-free utterances (Housen and Kuiken, 2009). *Fluency* is understood as the ability to communicate in real time in real-life situations with appropriate speed and with only few pauses and reformulations, approaching the speed of native speakers (ibid.) *Complexity* is understood as the extent to which the learner language is elaborate and varied (Ellis, 2003). Complexity can concern the syntactical structure or the vocabulary, where the former is often assessed by the average number of dependent clauses per independent clause, the latter by a type-token ratio for words. Before complexity was added to the proficiency spectrum, only accuracy and fluency were distinguished, for instance by Brumfit (1984). Brumfit considered fluency and accuracy under the perspective of classroom activities that were targeted at fostering the one or the other, namely, either fluent, spontaneous oral production or controlled production of formally correct L2 utterances. While Brumfit seemed to assume that both goals could be pursued in one and the same classroom, though maybe not at the same time, the dichotomy was sometimes more radically framed into two opposing approaches to language teaching.

The accuracy-oriented approach considers and treats language as an object, whereas the fluency-oriented approach sees language as a medium for communication (Long, 1991). Consequently, lessons according to the first approach consist mainly of explicit presentations of the language structures, while lessons according to the second approach emphasize meaning and thus are mainly concerned with how to use language to communicate successfully. In the remainder of this section we will characterize

the two different approaches in more detail, and discuss their respective merits and disadvantages. It will become clear that there is a need to bridge the gap between those two extreme positions, and we will present an approach that attempts just that. Following the terminology established by Michael Long (1991) we will distinguish between the accuracy-oriented FOCUS-ON-FORMS approach (4.2.1), the fluency-oriented FOCUS-ON-MEANING approach (4.2.2), and the integrated FOCUS-ON-FORM approach (4.2.3). The following characterization is based on the accounts given by Long (1991); Long and Robinson (1998); Lightbown (1998) and Doughty and Williams (1998c).

4.2.1 Focus on forms

For the accuracy-oriented FOCUS-ON-FORMS approach, language is taught in terms of linguistic structures (forms) in a step-by-step fashion. The order of the forms to be taught is determined by the perceived difficulty or frequency and relevance of the forms. This approach concentrates on formal aspects of language, usually by isolating and extracting individual linguistic constructs from a meaningful communicative context (Doughty and Williams, 1998b). The instruction treats language as an object as opposed to a means of communication, and the content of lessons are the forms themselves (Long, 1991). This approach was the dominant approach until the 1980s and is often called the "traditional" approach. It is still widely used around the world, although it has now incorporated modifications influenced by approaches that place greater emphasis on meaning.

The FOCUS-ON-FORMS approach is based on several assumptions, one of which is that learners will learn what they are taught immediately after they are taught. This assumption involves the notion that learners learn a linguistic form in a categorical fashion, going from zero knowledge to perfect mastery in one step, rather than in a gradual approximation. Further, the way to present language as distinct forms seems to suggest that language can be learned piece by piece. Finally, there is the expectation that learners will be able to transfer knowledge about language structures taught in relative isolation from a meaningful context smoothly onto communicative meaning-driven contexts.

These assumptions have, however, been challenged. It is obvious that many learners experience difficulties in applying the theoretical knowledge they have about the L2 in practical situations (Kadia, 1988; Long, 1991). Another objection revolves around the order of taught items. Often, the order that is taught in the classroom does not reflect the so-called "natural order of development" (Dulay and Burt, 1973; Ellis, 1984). This order was derived from the observation that the development of second languages follows certain patterns, in which some structures are consistently acquired prior to others. The most prominent work on this phenomenon is that of Pienemann and colleagues (Meisel et al., 1981; Pienemann, 1984, 1988), who identified stages for word order rules for learners of German. As we will discuss in more detail below (Section 4.4.4) such developmental sequences have been identified for a diverse range of phenomena besides word order, for instance, question formation and relative clause formation, and they seem to override any instructional sequences. The effect of such developmental sequences is that instructed learners follow the natural order just like naturalistic learners despite the fact that the instruction they received differed from

the natural order (Pica, 1983; Ellis, 1989). Related to this, Pienemann (1984) argues that learnability determines teachability, supposedly making it impossible to teach forms for which learners are not developmentally ready (for more details see also Section 4.4 below).

A further objection to the FOCUS-ON-FORMS approach regards the piece-by-piece fashion of language instruction: Long (1991) argues that learners rarely master a linguistic form in one step when starting from zero knowledge, but rather that they approach the target forms gradually. After all, learning a language is not just the accumulation of items, but a much more inter-related process. Minor arguments against the FOCUS-ON-FORMS approach critique the lack of need analysis for a particular learner group and an undue simplification of language input, resulting in unrealistic and inauthentic language (Long, 2000). All these objections and alleged problems have fueled the development of a very different approach to language instruction - the FOCUS-ON-MEANING approach.

4.2.2 Focus on meaning

The premise of the FOCUS-ON-MEANING approach is that language is a tool for communication, and therefore that learners should learn how to use language for communicative purposes. As such, this approach is driven by the learners' needs. Linguistic structures and grammar are never made the topic of a lesson. The approach is supposedly based on the assumption that the acquisition of a second language follows the same processes as the acquisition of a first language. In particular, it assumes that innate acquisition processes override any potential effects of explicit instruction. This means that grammatical structures can presumably be learned incidentally and without awareness. We will elaborate on the role of awareness when we discuss implicit learning processes in Section 4.3.1 below. One of the most prominent proponents of incidental and unaware nature of second language learning is Stephen Krashen, who posited the INPUT HYPOTHESIS, which claims that comprehensible input is sufficient for language acquisition (1982). In general, advocates of this strictly meaning-based approach (also known as the *non-interventionist* position) are convinced that grammar-based instruction has no or only a negligible effect on learners' L2 proficiency.

Although there had been no clear evidence for the alleged insufficiency of form-oriented instruction – for instance in form of controlled experimental studies – the alternative meaning-based approach presumably arose out of a general dissatisfaction with the traditional grammar-based approach and its apparent failure to produce highly proficient learners. However, the meaning-only approach and has never been clearly shown to be superior to its predecessor either. Part of the reason for this may be the difficulty of conducting controlled comparative studies. However, at the beginning of this controversy, methodological problems were seldom addressed explicitly. Instead, the debate was primarily based on theoretical arguments and anecdotal evidence. Efforts to substantiate the respective claims have only been made later. For more details, see for instance Long (2007, chap. 1) who illustrates in his characterization of history in SLA research how the idea of accountability and evidence for theories only arrived relatively late.

First objections to the strictly meaning-based approach were grounded in experi-

ence from Canadian immersion classrooms, in which native speakers of English were taught a major portion of their lessons in French. These students acquired native like comprehension skills but, in the absence of any attention to form, the accuracy of their production was far from native-like, even after years of immersion (Swain, 1985).

Further, the main premise of the FOCUS-ON-MEANING approach, namely that second language acquisition works exactly as first language acquisition seems to be invalid. There is evidence that a second language cannot be acquired in the same way as a first language (L1) after a certain age, probably due to maturational processes in the brain (DeKeyser, 2008; Johnson and Newport, 1989; Newport, 1990). The most convincing evidence for this is the fact that the overwhelming majority of L2 learners never achieve native-like proficiency: Their pronunciation is very often non-native, their grammars are incomplete, and their vocabulary seldom reaches native-like breadth ever after years of exposure to the language (see Schachter (1996) for a review).

Another objection to a solely meaning-based approach is that some L2 structures are unlearnable through positive evidence alone, if there is a certain contrast between L1 and L2. If the L2 is more restrictive than the L1 and the L1 allows constructions that are not possible in L2, negative evidence or some form-related instruction is necessary for the learner to become aware of the difference (White, 1987, 1991, and see also more details below, 5.2.1, page 85). If the incorrect form does not hinder comprehension, an exclusively meaning-focused manner of instruction is unable to alert the learner to the mismatch.

Finally, an important argument against the FOCUS-ON-MEANING approach is that it seems *inefficient* for fostering formal accuracy of the L2. Building on the experience with immersion classes, more controlled studies have sought to compare a solely meaning-based instruction with instruction that also addresses formal aspects of language. These studies give convincing evidence that the latter is more advantageous for increasing the grammatical correctness of learners (Doughty, 1991). In the next section we will present details about how it is possible to combine a focus on meaning with a focus on form.

4.2.3 Combining both approaches - focus on form

The apparent disadvantages of the FOCUS-ON-FORMS and FOCUS-ON-MEANING approaches and their negligence of one aspect at the cost of the other led to the attempt to combine them both. The most prominent approach in that tradition was developed by Michael Long. In Long (1991) and Long and Robinson (1998) he proposed what he called, a little ambiguously, "focus on form" as opposed to focus on formS. Arguably, the terms are a little unclear; nevertheless they are now established and commonly used, and we will therefore adhere to this terminology. The FOCUS-ON-FORM approach shares with the FOCUS-ON-MEANING approach an assumption, that the underlying content of the lesson is meaning-driven and communicative. However, unlike pure focus on meaning, the attention is occasionally shifted to form if the need arises: *"focus on form [...] overtly draws students' attention to linguistic elements as they arise incidentally in lessons whose overriding focus is on meaning or communication"* (Long, 1991, p.45-6) and *"focus on form involves an occasional shift in attention to linguistic code features"*(Long and Robinson, 1998, p.23). The two definitions contain two views on

attention that illustrate two different perspectives on the FOCUS-ON-FORM approach. One is about what teachers intentionally seek to establish (*draw* attention), the other is about what learners actually do (*shift* attention). These two clearly do not always correspond. If the teacher tries to draw attention to some formal aspect, it is not guaranteed that the learner will attend to that aspect. At the same time, learners can shift their attention to some formal aspect, which the teacher had no intention to put in focus (Long, 1991). Despite this reservation, teachers need to assume that learners' attention can be influenced and directed to a certain degree by the instruction they receive, otherwise all teaching would be futile.

The rationale for FOCUS-ON-FORM is based on two hypotheses: the NOTICING HYPOTHESIS and the INTERACTION HYPOTHESIS. The noticing hypothesis states that learners have to notice, i.e., register forms in the input in order to learn them (Schmidt, 1990, 1993). However, noticing does not necessarily entail that learners understand the meaning of a form. The interaction hypothesis states that interaction between learners and other speakers is beneficial for language development, because it enables *negotiation of meaning* (see for more detail Long (1981) and Section 4.5.1).

Long argues that in FOCUS-ON-FORM (as opposed to FOCUS-ON-FORMS) the forms are determined by the developing language of learners and the learners' needs that come about in a communicative situation. In addition, learners are likely to (at least partially) comprehend the meaning and function of the forms, because they arise out of authentic language use (Long, 1991; Long and Robinson, 1998).

The advantages of form-focused instruction (FFI), that is, instruction that addresses formal aspects - whether exclusively (FOCUS-ON-FORMS) or integrated within a meaning-based context (FOCUS-ON-FORM)[1] - compared to solely meaning-focused instruction (FOCUS-ON-MEANING) or mere exposure in the context of naturalistic acquisition are the following: FFI increases the rate of acquisition (Doughty, 2003), it leads to a higher ultimate level of attainment, (Long, 1991; Doughty, 2003), and it increases the accuracy with which forms are used (Leeman et al., 1995; Doughty and Varela, 1998). Further, focusing on form(s) provides negative evidence for forms that are incorrectly used by learners due to L1 influence but do not lead to communicative problems in a exclusively meaning-based context, as described above (White, 1987, 1991).

Although the FOCUS-ON-FORM approach is often presented in contrast to the FOCUS-ON-FORMS approach, it should be clear from the account given above that they should not be considered as "polar opposites" (Doughty and Williams, 1998c). Rather, FOCUS-ON-FORM lies in the middle ground between the two extremes of exclusively considering either forms or meaning.

An objection to the integrated FOCUS-ON-FORM approach is the potential limit of attentional capacities, that might render it impossible for the learner to simultaneously attend to meaning and form. This argument is based mainly on work by VanPatten, who tested the ability of learners of Spanish to comprehend the content of a text while paying attention to form features (VanPatten, 1990). Learners' performance indicated that it was difficult even for those of an advanced level to attend to form and meaning simultaneously. This brings us back to the introduction of this section, in which we discussed fluency and accuracy as aspects of language proficiency. According to

[1] see Ellis (2001) for more background on FFI

Skehan (1998) and Skehan and Foster (1999), the two objectives to be accurate and fluent compete with each other, because attention and processing capacities are limited. VanPatten's observations seem to support this position. However, this view is controversial. For instance, Long and Robinson (1998) argue for the plausibility of a model of cognition which uses multiple resources that can be accessed in parallel. We will now go into more detail regarding the FOCUS-ON-FORM approach and consider its different manifestations.

4.2.4 Parameters of focus on form

Although the differences between the three approaches seem clear at a general level according to the definitions given above, the concepts have been interpreted and appropriated differently by different researchers and have in consequence slightly shifted their meaning over the years (Doughty and Williams, 1998b). This has resulted in some disagreement over whether or not certain types of instruction can be considered as FOCUS-ON-FORM. These conflicts illustrate that FOCUS-ON-FORM has been realized in various ways, which differ in important respects. In the remainder of this section we will discuss two important dimensions along which FOCUS-ON-FORM realizations can vary. The first dimension regards the extent to which the focus on form is planned beforehand. The second dimension is about how to integrate the form focus into the meaning-based lesson. This question also includes the extent to which a form focus relies on some distinct, explicit explanation of forms as preparation. We discuss this in that much detail because it is important for the choice of the FOCUS-ON-FORM realization that we adopt for the present study.

Reactive/unplanned versus proactive/planned FOCUS-ON-FORM

FOCUS-ON-FORM can be reactive and unplanned or proactive and planned. For the first approach, the instruction is driven by problems that arise within a meaning-based context. The instructor notices these problems and consequently focuses on them. This reactive approach demands the ability of the instructor to notice problems immediately and react promptly and appropriately. The open, unplanned and incidental approach also seems to be what Long had in mind when he first defined FOCUS-ON-FORM. However, given the open nature of this approach, it is hard to test its effectiveness, and there are indeed only a few classroom studies that have aimed to investigate this (Doughty and Williams, 1998c).

A study by Spada and Lightbown (1993) found indirect evidence for the effectiveness of such an unplanned, reactive FOCUS-ON-FORM. Spada and Lightbown had originally sought to investigate the effect of explicit form-focused instruction and corrective feedback compared to the default FOCUS-ON-MEANING instruction style practiced in the given school context, when a teacher of the meant-to-be FOCUS-ON-MEANING control group behaved unexpectedly and implemented a reactive FOCUS-ON-FORM approach. It turned out that the control group, which had apparently been subject to this FOCUS-ON-FORM instruction for months, outperformed the experimental groups who had been taught according to the FOCUS-ON-MEANING approach in the months preceding the 2-week treatment. Although the performance was only measured for

one phenomenon, namely English question formation, and there were no long-term records of the actual instruction apart from the 2-week experiment period, this observation suggests that a comprehensive, reactive FOCUS-ON-FORM can be effective.

However, in many instructional contexts, another, more pro-active and planned approach is easier to implement. For this approach, the instructor plans in advance which forms to focus on, either by setting up the tasks and meaning-focus of the lesson in such a way that the target forms are likely to occur (see Section 4.5.2 for more details on how to achieve this), or by filtering incoming problems, such that the instructor will only focus on a subset of problematic structures and ignore the others. Such an approach is easier to control and is often used when evaluating the effect of specific forms of feedback.

Another method that implements a planned approach is the a-priori provision of form-focus by techniques that increase the perceptual salience of forms, known as *input enhancement*. The term "input enhancement" denotes various kinds of techniques that manipulate or enhance the input in order to draw learners' attention to formal aspects of the language (Leeman et al., 1995; Sharwood Smith, 1993). For example, linguistic forms can be highlighted by using a different font. Input enhancement is another technique to integrate meaning and form in a simultaneous and thereby unobtrusive way.

For all of the mentioned pro-active and planned approaches, however, it can be argued that they are not consistent anymore with Long's original definition of FOCUS-ON-FORM which emphasizes that the focus should be *incidental*. As a matter of fact, Ellis (2001) addresses this meaning shift and argues that planned FOCUS-ON-FORM differs in an important respect from incidental FOCUS-ON-FORM: The former is intensive - focusing on a single form many times, while the latter is extensive - focusing on a subset of a wide range of linguistic forms. This contrast also raises the question of whether intensive or extensive types of instructions are more effective, an issue which is beyond the scope of this thesis. Ellis hypothesizes that the re-conceptualization of FOCUS-ON-FORM is motivated by the need of researchers to conduct controlled experimental studies, which is hard, if not impossible, for purely incidental, reactive and unplanned FOCUS-ON-FORM.

As we will illustrate in more detail in the next chapter, the feasibility of reactive and proactive FOCUS-ON-FORM approaches is particularly relevant for attempts to realize such an instruction through a computer system. Given that the reactive approach already requires considerable skills on the part of a human instructor, one can assume that it would be an ambitious and challenging task to realize it through an artificial agent. As we will see in Chapter 2, the current state of the art in computer-assisted language learning is in general not fit to realize a fully reactive FOCUS-ON-FORM approach and therefore most engineering attempts settle for the more proactive approach if they attempt to give meaningful corrective feedback.

Integration of meaning and form

The challenge of the FOCUS-ON-FORM approach is to focus on linguistic structures without interrupting a primarily meaning-driven activity (Doughty and Varela, 1998). Since engagement in meaning is required before focus on formal features can be estab-

lished, communicative goals can be used to motivate the need for attention to form, to the extent that certain forms are required to realize these goals (Skehan, 1996a). However, it is not clear how exactly to integrate both aspects.

Doughty and Williams (1998c) distinguish three degrees of integration - simultaneous, sequential, and with preparation. In the first approach, form and meaning are in focus at the same time. However, this strictly simultaneous integration is not always feasible, either due to the cognitive limits of the learner or due to specific characteristics of the form. In the first case, the learner might not be capable of paying attention to both meaning and form at the same time (VanPatten, 1990). In the second case, the form might not be essential for transporting that particular meaning and therefore inaccuracy in production or non-noticing in comprehension will not cause a breakdown in communication. For such forms it is impossible to create a communicative task for which the form is essential. This poses a problem, which we will discuss in more detail in Section 4.5.2. If a simultaneous integration is impossible, it is more feasible to integrate form and meaning sequentially. However, Doughty and Williams (1998c) argue that sequential attention to both should occur within a limited time frame.

As a third way, there is the method of preparing the FOCUS-ON-FORM session with a distinct FOCUS-ON-FORMS session, in which the forms are explicitly explained and potentially also practiced in a more controlled way before they are used within a meaning-based context (DeKeyser, 1998; Lightbown, 1998). This approach raises the question of whether it is necessary or desirable to explain forms separately from the meaning-based communication in which they will be used. DeKeyser (1998) argues that this is so on the basis of skill acquisition theory, which assumes that skills are developed based on explicit knowledge that is gradually proceduralized and automatized through practice (see also Section 4.3.5). Others, however, rule out such preparatory forms-only sessions for the FOCUS-ON-FORM approach and argue that such separation is inconsistent with proper FOCUS-ON-FORM (Doughty and Williams, 1998c). For the current study, we exclude a preparatory session for practical reasons, but we assume that the learners had been exposed to some amount of instruction before. We will discuss this in more detail in Section 6.6.3.

This section served to present different approaches to instruction and characterize how they give different weight to meaning and form. We argued for the FOCUS-ON-FORM approach as a method which combines attention to form and meaning. In general, one can argue that form-focused instruction makes form explicit at some point, while in meaning-based instruction forms are usually treated more implicitly. The difference between explicit and implicit learning and knowledge is the topic of the next section.

4.3 Implicit and explicit learning, instruction, and knowledge

The difference between implicit and explicit learning processes and the nature of the resulting L2 knowledge is an important issue in second language acquisition. Beyond the general agreement that learners' attention contributes to language acquisition, the extent to which this attention has to be conscious remains controversial. Does the

learner need to be aware of the language structures, or is it possible to learn implicitly and incidentally like children learn their native language? Should grammar be taught explicitly or implicitly (DeKeyser, 2008)? How is the L2 knowledge represented by the learner and how is it accessed during L2 use (Doughty and Williams, 1998c)?

Before discussing the meaning of the implicit/explicit dichotomy in more detail for the three areas – learning, knowledge, and instruction – we will begin with a general characterization of the two terms, following the definitions provided by Doughty and Williams (1998c) and Gove et al. (1993) (Webster's Dictionary). *"Implicit"* indicates that something is implied and potentially inferable from something else, but not clearly expressed or revealed and thus not readily apparent. *"Explicit"*, on the other hand, indicates that something is fully and clearly expressed and therefore clearly observable, leaving little room for vagueness or ambiguity. When applying these general meanings to the specific areas of knowledge and learning, the defining criteria are *consciousness* and *awareness*: Explicit knowledge is conscious and learners are aware of what they learn when they learn explicitly, while implicit knowledge is usually unconscious, and implicit learning proceeds without awareness. In parallel, explicit instruction seeks to make learners aware of linguistic forms by making them overt, noticeable and salient. Implicit instruction on the other hand creates conditions in which learners are exposed to linguistic forms without paying conscious attention to the forms.

In the following sections we will discuss the dichotomy at each level in more detail, starting with learning Section 4.3.1 and instruction in Section 4.3.2. We will review the existing evidence for both types of learning in Section 4.3.3. In Section 4.3.4 we will then focus on implicit and explicit knowledge and discuss how these two types of knowledge are related Section 4.3.5. We finish this part by introducing possible measures to assess implicit and explicit knowledge in Section 4.3.6.

4.3.1 Implicit and explicit learning

There is a general agreement that the key criterion that distinguishes explicit from implicit learning is the learner's **awareness**: During implicit learning learners are not aware of what they are learning, while in explicit learning they are (Ellis, 2009a; DeKeyser, 2008). However, beyond this agreement, there is some dissent about the meaning of awareness. Schmidt (1994, 2001) distinguishes two levels of awareness. At the lower level of awareness, learners do notice certain elements of the surface structure of the language input they receive, but they do not analyze these elements or reason about them. At the higher level, the metalinguistic level, or "level of understanding" (Schmidt, 1990, page 145) learners analyze the input elements and create generalizations or rules. Regarding the relation between implicit learning and these two levels of awareness, there is agreement that implicit learning does not happen at the higher, metalinguistic level of awareness. This means that integration and restructuring of new input take place autonomously and without conscious control during implicit learning (Ellis, 2009a). However, researchers do not agree on the connection between the lower level of awareness and implicit learning. Schmidt claims that learning is impossible when the learner does not notice certain elements in the input (1994; 2001). Williams (2005) on the other hand, has found evidence that seems to contradict Schmidt's contention and indicates that learning may indeed happen without the

learner noticing.

Another controversy concerns the criterion of intentionality. Explicit learning is often characterized as being intentional (Ellis, 2009a), whereas implicit learning supposedly excludes any intention to learn something. However, this view has been questioned by DeKeyser (2008), based on the argument that subjects in an experiment (as well as students in the classroom) may indeed have the intention to learn something. However, the intention may not be directed at the particular linguistic structure or rule that is the target of the instruction.

Since learning processes are influenced by the instruction that the learner receives, it is important to make the distinction between explicit and implicit also on the level of instruction.

4.3.2 Implicit and explicit instruction

There is a range of criteria for distinguishing between explicit and implicit types of instruction. According to Ellis (2009a), implicit instruction tries to create conditions in which learners are exposed to specific instances of language rules or patterns with the goal to enable learners to infer rules and internalize them without being aware of them and without paying conscious attention to them. In contrast, explicit instruction has the goal to make learners aware of metalinguistic rules, either by providing the rules or by discreetly directing learners to discover the rules themselves. Another perspective on the dichotomy is given by Doughty and Williams (1998c), who characterize implicit instruction as unobtrusive and as minimizing any interruption to the communication of meaning and avoiding any metalinguistic discussion. In contrast, explicit instruction makes reference to metalinguistic concepts (or pedagogical grammar) and is overt and obtrusive. Further, Doughty and Williams distinguish the two types of instruction with reference to learner attention: Implicit instruction tries to *attract* learner attention, while explicit forms of instruction try to *direct* learner attention. Another, more practical definition based on DeKeyser (1995) is used in Norris and Ortega (2001) for the purpose of classifying a wide range of instruction techniques for a meta-study: Instruction is considered explicit if it contains an explanation of the language phenomenon in question or asks learners to attend to particular forms in the target language. In any other case, it is considered to be implicit.

The parallel between the implicit/explicit dichotomy and the classification of instruction as either focusing on meaning, form, or forms is evident. Indeed, Housen and Pierrard (2005a) seem to confound the two dimensions by framing the implicit-explicit distinction in terms of the distinction between FOCUS-ON-FORM and FOCUS-ON-FORMS. In addition to the above they list the following criteria: Implicit instruction presents target forms in context and encourages their free use, while explicit instruction presents them in isolation and provides controlled practice of them. However, this kind of blending is not supported unanimously. For instance, Doughty and Williams (1998c) argue that FOCUS-ON-FORM comprises both implicit and explicit types of instructions. According to their view, it makes sense to consider FOCUS-ON-MEANING as a very implicit type of instruction, FOCUS-ON-FORMS as a very explicit type of instruction and FOCUS-ON-FORM as somehow in the middle, ranging from rather implicit to rather explicit types of instruction. Table 4.1 summarizes the different criteria for

implicit and explicit instruction.

As with the distinction between FOCUS-ON-FORM and FOCUS-ON-FORMS, it is important to note that the instruction can only be defined from the perspective of the instructor, but not from the learner's perspective. It cannot be taken for granted that implicit instruction results in implicit learning, just as explicit instruction does not necessarily entail explicit learning (Ellis, 2009a).

Implicit Instruction	Explicit Instruction
provision of instances, rules should be inferred and internalized	rules are either provided or learners are guided to discover them
unobtrusive (minimal interruption of meaningful communication)	obtrusive (interrupts meaningful communication)
no metalinguistic explanations	metalinguistic explanations
attracts attention to forms	directs attention to forms
forms are used in context, free use of them is encouraged	forms are used in isolation and practice is controlled

Table 4.1 – Implicit and explicit instruction

4.3.3 Empirical evidence

The effectiveness of explicit and implicit instruction and learning processes has been investigated in a wide range of studies, the results of which we will summarize below. In general, these studies are based in two different fields – one is cognitive psychology, which has a more general perspective on learning, the other is second language acquisition with a more specific perspective to the language learning context. The underlying question of many of these studies is to what extent implicit learning is possible and how it compares to explicit learning.

Artificial languages

Studies from the field of cognitive psychology that are cited as evidence for the existence of implicit learning in general usually involve the learning of artificial grammars. A more detailed review is provided in Ellis (2009a) and DeKeyser (2008). For instance, in an experiment described by Reber et al. (1991), subjects were exposed to a set of strings or symbols that were constructed according to a set of rules. After this learning phase, they were presented with another set of strings and asked to judge if these are consistent with the rules. Subjects were not told about the rules, and they did not know that they would be tested later. There were two conditions, one for explicit and one for implicit learning. During the learning phase of the explicit condition, participants had to figure out the underlying rules by means of a test which asked them for the next letter according to the rules, but they did not receive any feedback on whether or not their hypothesis was correct. The implicit condition contained no such task and

no additional information, participants were just presented with the symbols. The results of this study and similar ones show that implicit learning of artificial grammars is possible and that it can be more effective than explicit learning for complex rules. Yet, when the goal is to learn simple rules, there seems to be no difference in effectiveness between implicit and explicit learning. In this experimental paradigm, explicit learning is operationalized by asking the learner to derive rules from the input, which is supposed to make the learner aware of underlying rules. However, since learner awareness was not measured, it is not clear that learning processes were either implicit or explicit respectively (Ellis, 2009a). Further, there is some doubt about this experimental paradigm regarding the extent to which it can be generalized to the learning of natural languages. The rules of natural languages are supposedly different from the rules of artificial symbol sequences. Further, the context of natural acquisition is usually less controlled and there may be additional factors that play an important role.

Natural languages

Although the foundational research that we have just discussed is still considered important and valuable, recent studies have focused more on investigating the learning of real languages in authentic classrooms. Doughty and Williams (1998c) and DeKeyser (2008) provide an extensive review of studies that sought to compare the effect of implicit and explicit instruction directly. In summary, these studies indicate that instructions which contain explicit rule presentations are more effective than more implicit instructions that provide no rules, but only presented language instances. Ellis (2009a) comes to similar conclusions in his review of studies on implicit and explicit language learning. According to Ellis, there is no study to date that has shown that implicit learning is superior to explicit learning. In a more systematic attempt to compare the effectiveness of explicit types of instruction with implicit types of instruction, Norris and Ortega (2000) conducted a meta-study that included a large range of relevant studies. They found a slight superiority for explicit types of instruction over implicit types of instruction. However, they caution that the assessment measures in most studies were inadvertently biased in the favor of explicit knowledge. This objection is addressed in a successor meta-analysis conducted by Spada and Tomita (2010) which includes newer studies but reaches similar conclusions as Norris and Ortega (2000). Explicit instruction yields larger effect sizes and can therefore be considered advantageous. The new studies employed free response measures, in which learners are relatively unconstrained in their use of language. These are arguably a better measure for implicit knowledge, since the elicitation is embedded in a communicational context. Spada and Tomita conclude that explicit instruction was beneficial not only for explicit knowledge but also for implicit knowledge as exhibited in the ability to use target forms spontaneously. Further, explicit instruction was advantageous in short- and long-term treatments and for simple and complex features.

Further notes

In summary, up to now, the empirical evidence suggests that more explicit forms of instruction seem to be more effective. However, it is important to ensure that the measures that are employed for assessing learning gains tap into implicit as well as explicit knowledge; we will discuss such measures in Section 4.3.6. At the same time, implicit learning may in general take more time than explicit learning, therefore it is disadvantaged by the relatively short time spans that many comparative studies cover (Ellis, 2009a; DeKeyser, 2008). The relative effectiveness of implicit and explicit instruction also depends on the nature of the structure, which we we will discuss in more detail in Section 4.4, and the proficiency level of the learner. A study by Gass and colleagues suggests that learners with lower proficiency seem to benefit more from explicit instruction than highly proficient learners (Gass et al., 2003). As we will see later in Section 5.5.1, the proficiency level of the learner also has an effect on the effectiveness of implicit feedback. Further, the advantages of explicit instruction seem to diminish with more complex structures (Reber et al., 1991; Robinson, 1996). Similarly, the study described by Green and Hecht (1992) indicates that a grammatical phenomenon which is determined by relatively vague rules is more likely to be mastered implicitly – as evidenced by the ability to correct an erroneous utterance – than explicitly – as evidenced by the ability to provide a metalinguistic rule. Green and Hecht investigated implicit and explicit knowledge about different English phenomena and found that German learners had difficulties in providing metalinguistic rules about vague phenomena like the *some/any* distinction, or verb aspect, i.e., when to use the continuous form or the perfect tense. At the same time, these learners were able to correct errors regarding these vague phenomena, which allows for the interpretation that the learners had no explicit, but implicit knowledge.

As we have stated above, instruction does not correspond exactly to the learning processes it entails. In general, the experimenters have more control over the instruction than over the consequent learning processes. Further, learning can only be tested indirectly through testing the knowledge that results from the learning, based on the assumption that implicit learning results in implicit knowledge and explicit learning results in explicit knowledge. Although this assumption seems reasonable, there may be cases in which it does not hold. In the following section we will characterize implicit and explicit knowledge and look into possible connections between the two types of knowledge. If explicit knowledge can turn into implicit knowledge or vice versa, the above assumption may be invalid.

4.3.4 Implicit and explicit knowledge

Knowledge is the result of learning processes. Obviously, there is a difference between knowing how to ride a bike and knowing the capital of your state or its number of inhabitants. Similarly, there is a difference between knowing how to speak your mother tongue and knowing how to conjugate a verb or decline a noun in a foreign language that you just started to learn. Riding a bike or speaking your native tongue requires the ability to *do* something without necessarily being aware of the rules that determine your actions. On the other hand, reciting facts or applying grammatical

rules relies on conscious awareness and the retrieval of factual knowledge. The difference between implicit and explicit knowledge involves different levels that have been discussed most comprehensively by Rod Ellis, e.g., 2005; 2009a. We will briefly summarize his account below.

Criteria for the distinction between implicit and explicit knowledge include awareness, representation, accessibility, verbalizability, and learnability. The first criterion regards the **level of awareness**: While explicit knowledge is conscious, implicit knowledge is intuitive and tacit. Learners may intuitively know that a sentence is ungrammatical but only explicit knowledge enables them to know why it is ungrammatical and what rule it breaks. Recall that the awareness criterion was also the most significant criterion for distinguishing explicit and implicit learning processes. The second criterion concerns the **representation** of the knowledge. Explicit knowledge is declarative and encyclopedic, it consists of facts. Declarative knowledge about a language can comprise abstract rules or concrete exemplars. Implicit knowledge, on the other hand, is procedural. If a procedure is a series of actions that accomplish a goal, and these actions are dependent on one or more conditions, procedural knowledge can be understood as a set of condition-action rules. Procedural knowledge is thus revealed in behavior according to these rules, "that is, knowledge, how to do things" (Anderson, 1983, page 215). Procedural knowledge of a language allows learners to encode meaning into a surface form and decode the surface form of an utterance to arrive at the meaning.

The third criterion is **accessibility** of knowledge. Implicit knowledge is accessible through automatic processing, that is, it can easily and rapidly be accessed in unplanned language use. In contrast, explicit knowledge is only accessible through controlled processing, and thus usually not as fast as implicit knowledge. This issue is somewhat controversial however. DeKeyser (2008) argues that explicit knowledge can be automatized through practice up to a point where it is 'functionally equivalent' to implicit knowledge. Opposed to this point of view, Hulstijn (2002), considers such automatization as the development of implicit knowledge.

The forth criterion relates to the ability to **verbalize** the knowledge: While implicit knowledge is only evident in verbal behavior, explicit knowledge is verbalizable (Ellis, 2009a). Verbalization does not necessarily require metalinguistic terminology, it is possible to describe explicit linguistic knowledge with plain language as well. Implicit knowledge, on the other hand, can only be verbalized after reflecting on it and generating explicit knowledge through this reflection (Bialystok, 1994).

The fifth dimension regards **learnability**. There is convincing evidence that there are age constraints for the acquisition of implicit knowledge, while explicit knowledge can be acquired without such constraints (Ellis, 2009a; Schachter, 1996). Recall that we shortly discussed this issue in the context of meaning-focused instruction in Section 4.2.2. In summary, implicit knowledge is reflected in the ability to produce and comprehend a second language fluently and accurately, while explicit knowledge is factual and conscious knowledge about the second language, which involves metalinguistic awareness (Andringa, 2005). It is important to note that both implicit as well as explicit L2 knowledge is not perfectly target-like, but imprecise, inaccurate and incomplete.

4.3.5 Interface debate

The relationship between implicit and explicit knowledge, learning, and instruction has been of interest to many SLA researchers, and different positions on it have formed the "interface debate". This debate comprises the following questions: Are there any connections between implicit and explicit knowledge? Can implicit knowledge be made explicit? Does explicit knowledge convert into implicit knowledge or facilitate the acquisition of implicit knowledge and if so, how? An important motive in this debate is to evaluate the utility of explicit knowledge and instruction for the development of implicit knowledge, based on the assumption that the final goal for L2 proficiency is implicit knowledge (Andringa, 2005). In summary, the interface debate revolves around the question, of whether there is an interface between the two kinds of knowledge. There are three positions regarding this issue, which we will summarize below – the non-interface position, the strong interface position, and the weak interface position.

The non-interface position

Proponents of the non-interface position argue that implicit and explicit knowledge representations are completely separate and that there is no transfer from one to the other. This also means that they are acquired through different processes, are located in different areas of the brain, and are retrieved in different ways (Ellis, 2005).

Krashen (1981, 1985) is usually cited as the most prominent proponent of this view because he was the first to propose the distinction between *acquisition* and *learning* and consequently the distinction between acquired (implicit) knowledge and learned (explicit) knowledge. This view was probably based on the common observation that the explicit teaching of grammatical rules does not directly lead to learners who can use this knowledge fluently, that is, to a degree of automaticity that would suggest that they have implicit knowledge. Note how this is closely related to how Krashen argued for a focus on meaning as opposed to a focus on form(s) in language instruction, as we have discussed above in Section 4.2.2. However, since Krashen's theory was not based on empirical evidence and did not include any criterion of falsifiability, it was subject to strong criticism (Ellis, 2009a).

Independently from this criticism, Krashen's position was strengthened by evidence for the neuroanatomical separateness of the two knowledge systems. Paradis (1994) argues that the two kinds of knowledge are located in different areas in the brain. His primary evidence are bilinguals who lost their implicit knowledge of their L1 due to brain damage, but retained the ability to use their L2. Ellis (2004) also argues for the separateness of the knowledge representations, but based on a connectionist view of learning and knowledge (Christiansen and Chater, 1999). In this view implicit knowledge is considered as "weighted content", i.e., an elaborate network of node connections with different strengths, which determines the probability of following these routes. Ellis considers such weighted content as incompatible with the representation of linguistic facts. However, since he does not elaborate further on this argument, it does not become clear why linguistic facts could not equally be represented as strengths of connections. Interestingly, unlike Paradis, Ellis does not conclude from

the separate representation that it is impossible for the two types of knowledge to interface and be converted into one another.

The strong interface position

Proponents of the strong interface position assume a strong relation between the two knowledge systems. They argue that explicit knowledge can be rendered implicit and vice versa. Supporters of this view agree that explicit knowledge can convert into implicit knowledge through practice. However, there is no agreement about the nature of this practice. As an advocate of this position, DeKeyser (1998, 2007a) proposes skill acquisition theory (SAT) (based upon Anderson's Adaptive Control of Thought (ACT) model (Anderson, 1983)) as a model of how explicit knowledge gradually becomes implicit. According to this model, the development of knowledge goes through three stages. In the first stage the knowledge is declarative, in the next stage it is proceduralized by applying it, and finally, in the last stage, it is automatized. DeKeyser argues that proceduralization works through engaging in target behavior with the temporary help of declarative knowledge. Automatization leads to robustness of knowledge and fluency in the usage of this knowledge. Furthermore, automatization results in a decrease of the reaction time and error rate as well as the amount of required attention (DeKeyser, 2007a). According to DeKeyser, while the transition from the declarative stage to the proceduralization stage can be quite quick, automatization is a more tedious and costly process and takes a lot of practice.

The weak interface position

Finally, the weak interface position holds that explicit knowledge can convert into implicit knowledge, but only under certain conditions and in certain ways. One such condition regards the developmental readiness of the learner according to developmental sequences of grammatical features. As we have briefly mentioned above in Section 4.2.1, and will discuss in more detail below in Section 4.4.4, it has been observed that the development of second languages follows certain temporal orders, where some structures are consistently acquired before others (Meisel et al., 1981; Pienemann, 1989). As a consequence for the weak interface position, it is argued that explicit knowledge can only turn to implicit knowledge if the learner is in the appropriate stage for learning a specific grammatical phenomenon. The most prominent proponent of the weak interface position is Ellis (1994c). He argues that explicit knowledge facilitates the acquisition of implicit knowledge indirectly. Explicit knowledge is considered to have a positive effect on the perception of formal features in the input by making them more salient (Ellis, 1994b). As a result, learners are more likely to notice them. With regard to learner output, explicit knowledge is considered to work as a monitor to control the accuracy of the learner's production (Paradis, 1994).

4.3.6 Measures of explicit and implicit knowledge

In order to take a position in the interface debate and examine how implicit and explicit knowledge develops from different types of instruction and through different

kinds of learning process, it is essential to find and use appropriate measures to assess knowledge. The choice of measures is also in particular relevant for the study we have conducted in the scope of this thesis and we will come back to this question in Section 7.3, when we argue for the measures we employ in the current study. Traditional measures for assessing learning gains are usually directed at explicit knowledge, and therefore indirectly favor explicit instruction when they are used to compare the effects of implicit and explicit learning and instruction (Long and Robinson, 1998; Norris and Ortega, 2001; Doughty, 2003). In order to overcome this imbalance, it is crucial to consider that different measures tap into different types of knowledge. In light of the existing preference for explicit knowledge, it is particularly important to find and employ measures for implicit knowledge.

Measures of explicit and implicit knowledge are obviously related to the characteristics of the respective knowledge types. Ellis (2009b) identifies four criteria to discern implicit from explicit measures. The **level of awareness** characterizes the extent to which learners are aware of their linguistic knowledge. Learners respond according to their *feel* when they use implicit knowledge, but according to *rules* when they use explicit knowledge. Related to that is the criterion **utility of metalinguistic knowledge**. While tests for explicit knowledge invite or even require learners to use metalinguistic knowledge, tests for implicit knowledge do not encourage the learner to use such knowledge. Another criterion is the **focus of attention** in a test measure: The focus can be either on meaning (implicit) or on form (explicit). The first can be realized as communicative free production (i.e., activities that involve unplanned language use and are directed at fulfilling some communicative purpose (Ellis, 2009b)), while the latter usually tests forms in isolated contexts (Andringa, 2005). Similarly, Norris and Ortega (2000) distinguish between 'free constructed response' measures, which target implicit knowledge and, on the other hand, 'meta-linguistic judgments', 'selected responses', and 'constrained constructed responses', which all measure explicit knowledge. For constructing a free response, test takers produce language with a communicative goal. The target structures are not strictly required by the test task, but their usage and correctness is analyzed for the subsequent evaluation.

Finally, an important criterion is the existence of a **time limit**. Among others, Han and Ellis (1998) and Ellis (2009b) argue that a limit for response time can prevent learners from accessing their explicit knowledge, since explicit knowledge is not as fast and easily processed as implicit knowledge (recall our discussion about accessibility on page 69). If there is no time pressure, learners have enough time to access their explicit knowledge and to monitor their response production. If there is an appropriate time pressure, learners can be forced to use their implicit knowledge. It is not entirely clear, though, how to determine the appropriate length of a time limit that forces the learner to draw on their implicit knowledge. An adequate time limit should (a) allow enough time for the learner to process the item semantically and at the same time (b) be short enough to prohibit the use of explicit knowledge (Loewen, 2009). With regard to time limits in grammaticality judgment tests, which ask learners to indicate if an item is grammatical or not, Ellis (2004) identifies three consecutive operations. First, the learner has to understand the meaning of the item (semantic processing), second, the learner has to search the item in order to determine if something is incorrect (noticing),

and third, the learner has to consider what is incorrect and possibly why (reflecting). Ellis argues that a timed test should prevent the last operation, but allow enough time for the first two. However, the exact time that meets these requirements is hard to determine. In addition, the appropriate limit might be different for individual learners. As an objection against a time limit, Purpura (2004) cautions that time pressure might increase the level of anxiety, which could add undesirable variability to the test. We will return to the time limit in Section 7.3.1 when we present the measures that we employed in the current study.

In the current section we have characterized implicit and explicit language learning, instruction, and knowledge. We have discussed how they differ and summarized the ongoing debate about the possible relation between implicit and explicit knowledge. In the face of limited time and resources, experimental studies in the field of second language acquisition usually concentrate on a small subset of language phenomena, so-called *target structures*. Since the properties of these structures have an impact on the experiment results, they need to be examined cautiously and taken into account for planning and evaluating experiments. In the next section we will discuss the properties of target structures with a view on how they interact with implicit and explicit learning and form- and meaning-focused instruction.

4.4 Properties of target structures

When teaching language and grammar in particular, it is important to decide which grammatical structures to teach and in which order. It is obvious that some structures are learned more easily and thus earlier than others, but the reasons for this are not yet entirely clear. This section gives a summary of the factors that have been hypothesized to influence the learnability of a particular linguistic form. As a consequence, one can argue that structures may not only have to be taught in certain orders, but also that different structures should be taught in different manners. Important factors for the learnability of a structure are salience, regularity, and functional value, (Ellis, 2006; DeKeyser, 1998; Doughty and Williams, 1998c; Hulstijn and de Graff, 1994; Goldschneider and DeKeyser, 2001). In addition, learnability is argued to be affected by the developmental readiness of the learner and other individual characteristics of the learner, e.g., motivation, language aptitude, memory capacity, learning style, and age, as well as first language(s).

The characteristics of specific structures also need to be considered for the design of experimental studies. Studies that compare the effect of different types of instruction usually focus only on one or a few structures because a more comprehensive set would be unfeasible. Therefore, the choice of these target structures needs to be well-founded. As we will show below in Chapter 5, where we discuss the role of feedback for language learning, the effectiveness of feedback is affected by properties of the grammatical structures as well. In the remainder of this section we will discuss the determining properties of target structures salience, frequency and regularity, and functional value.

4.4.1 Salience

Salience is understood as the inherent, and therefore permanent, property of a structure to attract the attention of the learner and, as a consequence, to be noticed by the learner. It is widely accepted that noticing structures is a prerequisite for acquiring them (Schmidt, 1990). Following this view, the salience or noticeability of linguistic structures is an important variable. Although it is hard to determine exactly the salience of a particular structure, different factors have been proposed. These features regard phonological, morphological, and syntactic properties (Goldschneider and DeKeyser, 2001; Witzel and Ono, 2003). On the phonological level, phonetic substance (the number of phones) and sonority (loudness) play a role, as well as whether the feature is stressed or unstressed and syllabic or not. On the morphological level, salience is influenced by how regular the morpheme is and if it is free or bound. For bound morphemes, the position within a word has an influence too. On the syntactical level, the position of a structure in a sentence and its complexity have an effect on the salience.

For the purpose of facilitating instruction, the inherent salience of a target structure can be modified and increased externally. For instance, for written input, it is possible to enhance forms typographically, for example, by using a different font or color. This relates to the range of techniques known as *input enhancement* that we discussed above in Section 4.2.4 (Sharwood Smith, 1993). Similarly, in spoken language, the salience of structures can be increased by applying atypical stress patterns. The frequency of a structure and its semantic properties are considered as factors for its salience by some researchers (Witzel and Ono, 2003), but we will treat them as separate features below.

4.4.2 Frequency and regularity

The frequency of a grammatical structure has been shown to have an impact on its learnability. Ellis (2002) argues that frequent forms are easier to learn than infrequent forms and that humans are very sensitive to frequency effects. For the purpose of instruction, frequencies can be manipulated fairly easily in order to facilitate the acquisition of structures that are rare in authentic texts. Related to frequency is the *regularity* of a grammatical structure. Regularity comprises the *scope* and *reliability* of the rule which governs the structure (Ellis, 2006). Hulstijn and de Graff (1994) define scope as the absolute number of instances that a rule covers and reliability as the percentage of cases in which the rule holds. The more exceptions there are to a rule, the less reliable it is. An example for a rule with wide scope is the plural marking of English nouns using the affix -s (Doughty and Williams, 1998c). Another example of a rule with wide scope regards the relation between the gender of German nouns and their surface form. There are around 15.000 singular nouns ending in *-e* and about 90% of them are feminine (Hulstijn and de Graff, 1994). Opposed to that, the rule that predicts that monosyllabic nouns that start with *Kn-* are masculine has a very narrow scope – it covers only 15 instances, because there are no more than 15 monosyllabic nouns starting with *Kn-*, 14 of which are masculine (Hulstijn and de Graff, 1994). Hulstijn and de Graff consider scope and reliability as factors for assessing the utility of explicit instruction. They argue that teaching reliable rules with large scope has the greatest

effect. In contrast, rules with low reliability and/or small scope are less effective, considering the cost to teach them and the probable outcome.

4.4.3 Functional value

The functional value, also termed *"semantic complexity"* (Goldschneider and DeKeyser, 2001) or *"functional complexity"* (DeKeyser, 1998), refers to the relationship between form and function. Forms that express exactly one meaning (i.e., there is a one-to-one correspondence between form and function) are easier to learn than forms which express several different meanings (1 form - n meanings) and multiple different forms that express one and the same meaning (n forms - 1 meaning) (Housen et al., 2005). As an example for a form that expresses several meanings, consider the *-s* suffix in the third person singular verb forms in English. It encodes information about person, number, and tense (present). Even more complex are German articles, which simultaneously encode gender, number, and case information. In addition to that, they are also ambiguous in that one article can encode several combinations of theses three features (Doughty and Williams, 1998c).

Related to the functional value is the concept of *semantic redundancy* (Hulstijn and de Graff, 1994) or *communicative value*: A structure can be essential for conveying a certain meaning, (e.g., *-s* plural noun suffix in English) or it can be purely formal and semantically redundant (*-s* suffix in the third person singular verb forms in English). It is supposedly harder to acquire (and notice) forms that are not semantically essential, i.e., carry no meaning (Ellis, 2006). However, note that so far, since the concept functional value is not well enough defined to assign discrete complexity values to a given structure, it cannot be operationalized in a straightforward manner (Goldschneider and DeKeyser, 2001).

4.4.4 Developmental readiness and processability

The concept of developmental readiness is based on the observation that acquisition follows relatively fixed routes, (a "natural order") (Dulay and Burt, 1973; Meisel et al., 1981), which are not influenced by pedagogical interventions (Ellis, 1989). According to Pienemann's TEACHABILITY HYPOTHESIS (1984; 1989), the success with which certain forms are taught depends on the developmental readiness of the learner to *process* the forms. Instruction cannot change the order of acquisition, but probably increases the rate of acquisition. Closely related to that, Pienemann conceived the PROCESSABILITY THEORY, which attempts to identify the relationship between properties of grammatical structures and the difficulties involved in processing these structures. Essentially this means that forms "that involve little manipulation or little demand on short-term memory tend to be acquired early" (Doughty and Williams, 1998c, page 215). Likewise, salient and continuous elements are also easier to process and therefore learned earlier than less salient and discontinuous elements.

Developmental stages have been identified most prominently for German word order (Clahsen, 1984; Clahsen and Muysken, 1986), English morpheme acquisition (Larsen-Freeman, 1975; Goldschneider and DeKeyser, 2001), and English question formation (Pienemann et al., 1988).

However, such an order has not been or could not be established for all grammatical phenomena. For instance, in a study of francophone adolescents in Switzerland, Diehl et al. (2002) could not find a systematic order for acquisition of German nouns, genders and numbers. Further, for some of the developmental stages, there are discrepancies between the results of different studies. For instance, Clahsen (1984) found that subordinate word order (finite verb in final position) is acquired as the last word order rule in German, and therefore only after subject-verb inversion, while Diehl et al. (2002) found that subordinate clause word order is acquired before subject-verb inversion. As Eckerth et al. (2009) argue, such inconsistencies make it questionable to deduce a grammatical curriculum based on any such found order.

When considering developmental readiness for instruction in the classroom, it is important to note that it is not always trivial to diagnose the current stage of a learner reliably and stages might differ between learners of one group, which poses an additional problem. Furthermore, a curriculum based on developmental sequences may be difficult to integrate with communicative goals because these two objectives are not necessarily consistent (Eckerth et al., 2009).

We will revert to the topic of structure properties below in Section 7.1, where we discuss the choice of target structures for the current study. When we discussed the meaning-based approaches for language teaching in Section 4.2 we did not go into detail about the means to realize these approaches. We will do this now, in the last section of the current chapter, in which we will discuss methods, concepts, and underlying principles of meaning- and communication-based language instruction.

4.5 Conversational interaction and task-based instruction

This section presents the rationale behind *conversational interaction* (Section 4.5.1) and the concept of *tasks* (Section 4.5.2) as tools to create focus on meaning and meaning-based instruction. Both concepts are related to the FOCUS-ON-FORM approach to language learning that we described in Section 4.2.3. Conversational interaction can support the FOCUS-ON-FORM approach by drawing attention to formal aspects of the language within a primarily meaning-focused context. Similarly, tasks, in the sense of task-based language learning, are designed in a way to engage learners in a meaningful goal while, at the same time, providing the opportunity to use certain linguistic forms.

4.5.1 Conversational interaction

The role of conversational interaction for second language acquisition has been the subject of a large body of research dating back to the 1970s. Among the first researchers who pointed out the importance of communication for language learning was Evelyn Hatch (1978). In contrast to the then dominant view, which considered the communicative use of the L2 as the *outcome* of the learning process, she proposed a consideration of the communicative use of the L2 as *leading to* the learning of the L2 – "Language learning evolves out of learning how to carry on conversations, out of learning how to

communicate" (Hatch, 1978, page 63). Besides the work of Hatch, there are other researchers who have emphasized and defined the role of communication for language learning; for instance, Breen and Candlin (1980) and Brumfit (1984). This approach is also referred to as "communicative language teaching" or simply the "communicative approach". In contrast to previous pedagogies, it emphasizes the role of language as a tool for achieving goals in the real world and as a means of social interaction. The goal is to enable learners "to use language meaningfully and appropriately in the construction of discourse" (Ellis, 2003, page 28).

Related to the rise of the communicative curriculum, Long (1981) developed the INTERACTION HYPOTHESIS, which we have discussed above in Section 4.2.3 in the context of the FOCUS-ON-FORM approach. Long's hypothesis was based on an examination of conversations between L2 learners and native speakers or more proficient non-native speakers. He had observed that problems in comprehension often led to a modification of the interaction – a process which was later termed *"negotiation for meaning"*. In a refined version of this hypothesis, Long (1996) argued that negotiation for meaning provides (1) comprehensible input for the learner, (2) feedback, and (3) the opportunity for the learner to modify their output. By considering these three factors, the interaction hypothesis amalgamated two contemporary theories that considered *input* and *output* respectively as important factors for acquisition.

The importance of comprehensible input is the key of Krashen's INPUT HYPOTHESIS (Krashen, 1982, 1985). Krashen claimed that input is the primary factor for acquisition, while output merely presents what has been acquired and thus has no beneficial effect for L2 development. However, this view was questioned by Swain (1985). As we have discussed above in Section 4.2.2, Swain had observed the language skills of Canadian immersion students who had achieved native-like comprehension skills, but lagged behind considerably in producing target-like utterances, especially regarding grammatical accuracy. These observations led Swain (1985) to consider the importance of learner output for language acquisition (OUTPUT HYPOTHESIS).

In order to clarify the role of conversational interaction for the learning process, we will explain in a little more detail how negotiation for meaning, input, and output contribute to language learning and acquisition.

Negotiation of Meaning

Negotiation of meaning is usually triggered by problems in comprehension. It is thus an attempt "to repair breakdowns in communication or ensure mutual comprehension of meaning" (Pica, 1994, page 510). The receiver of the problematic message will request clarification or confirmation, and the original sender follows up by repeating, rephrasing, segmenting, simplifying or elaborating the original message. In conversations between learners and competent speakers, negotiations initiated by the competent speaker push the learner to modify her output to make it more comprehensible. In case the learner initiates the negotiation, she will receive modified, more comprehensible input by the competent speaker, who ideally adapts his initial utterance to resolve the initial problems.

Work by Pica (1994) and Gass (1997) has further specified the role of negotiation, suggesting that it not only facilitates comprehension, but can also direct attention to

formal aspects of the language. Negotiation thus provides access to forms and helps to make the connections between form and meaning clearer, which is essential for the FOCUS-ON-FORM approach (as described above in Section 4.2.3).

Input

The most important role of input is to provide learners with positive evidence for possible L2 utterances. However, in order to be effective, the input needs to be comprehensible to learners. Interaction allows learners to signal their problems in comprehension, and as a consequence, to be provided with more comprehensible input. In order for the input to provide positive evidence about language forms, however, learners have to *notice* the forms (Schmidt, 1990, 1993). During interaction, usually only those forms will be noticed that contribute to comprehensibility or that are salient enough. Certain forms might not be relevant for any communicative task. Pica (1994), for instance, suggests that tense and aspect are grammatical phenomena, which are hard to make relevant in communication tasks. This means that interaction alone might not be sufficient for facilitating the learning of the entire form inventory of a language.

Output

Two functions are usually ascribed to learner output: One revolves around the notion of noticing and awareness, the other is related to practice. When learners try to produce output they might *notice* that they have problems in conveying their intended message and consequently be more attentive to aspects of the L2 that would help them to express the meaning (Swain, 1985, 1995). Thus, noticing also plays a crucial role with regards to the output.

It has also been argued that producing output can support the transformation of declarative knowledge into procedural knowledge, thereby promoting automaticity and fluency (de Bot, 1996; DeKeyser, 1997). Note however, that the possibility of such a transformation is controversial and subject to the interface debate that we have discussed above (Section 4.3.5). The results of the study described by DeKeyser (1997) suggest that comprehension practice alone is not sufficient for improving productive skills, but that production practice is necessary, too.

Swain (1985, 1995) argues further that output allows learners to test hypotheses about the L2. By producing output, learners invite feedback, which they can use to conclude whether or not their hypotheses were correct. Swain also emphasizes that learners need to be *pushed* to modify their output in order to produce correct and appropriate output and develop target-like production skills.

In summary, it is important to note that the positive effects of output production hinge on (a) communication partners who provide feedback and (b) the ability of the learners to notice the feedback, interpret it appropriately and, finally, to integrate it into their developing interlanguage. Since feedback plays a crucial role in these processes, we will discuss its properties, efficacy, and limits in more detail in Chapter 5.

After recounting the theoretical foundations of conversational interaction, we will now describe a practical method to evoke interaction in the classroom – teaching and

learning based on practical *tasks*.

4.5.2 Tasks

The goal of language learning is to gain knowledge about the language, but more importantly, to develop the skills to use that knowledge appropriately. In other words, the goal is to be able to use the foreign language for communicating in the real world. For language teaching, it seems only natural then to provide opportunities to use the language in a situation that is similar to situations in real-life. *Tasks* have been introduced as a means of providing such opportunities. Tasks are communicative activities that help to achieve non-linguistic goals via the use of language. The general meaning of "task" is a piece of work that is undertaken or attempted and is often to be finished within a certain timeframe Gove et al. (1993) (Webster's Dictionary). Within the scope of the communicative approach and Task-based Language Teaching and Learning (TBLT), tasks have been more narrowly defined as a "piece of classroom work which involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is principally focused on meaning rather than form" (Nunan, 1989, page 10). Further developments and refinements of the task concept have resulted in more detailed properties, which we describe below referring to Ellis (2003) and Skehan (1998):

1. Tasks have a primary focus on meaning.
2. Learners are free to use the linguistic forms they need for completing a task.
3. Tasks have a clearly defined communicative outcome.
4. Tasks are related to real-world activities.

The first criterion follows directly from a tenet of the communicative approach: Language should be a tool for communication, not an object of study. The second criterion is the consequence of the first: If meaning is primary, then there should be no restrictions on the forms that can be used to express the meaning. The outcome of a task provides the goal for the learners and determines the completion of a task. Tasks are grounded in real-world activities, because, eventually learners will want to use their language in the real world – tasks are supposed to prepare them for that.

However, there is problem with the primacy of meaning. As we have seen via the example of Canadian immersion students, an exclusive focus on meaning does not necessarily lead to grammatical accuracy. In particular, this is the case for linguistic forms that are not essential for conveying a particular meaning and are thus unlikely to cause problems for comprehension. However, although the communicative approach is based on the primacy of meaning and using language as a tool, there is a concern for accuracy as well. The development of formal accuracy is not neglected as a goal. Therefore it is generally accepted that some focus on form should be provided to allow interlanguage development (Skehan, 1998). However, it is not clear how to introduce attention to form within a task without compromising the primacy of meaning. How to find the appropriate balance between meaning and form? There are two approaches

of solving this problem. One approach uses pre- and post-task periods and general parameters for the execution of the task to draw attention to forms. The other approach is limited to the task itself – it tries to devise tasks in which certain forms come naturally into focus.

The first approach was advanced most notably by Skehan (1998), who proposed a comprehensive range of methods to create a focus on form. Some of these methods are used to prepare the task, for example, to introduce the form explicitly or to raise consciousness of the form in an implicit manner. Other methods implement post-task strategies that let learners reflect on the forms and consolidate them. Further, there are activities that raise the likelihood that accuracy and forms are attended to during the tasks. This can be achieved by announcing that students will be tested after the completion of the task, or that they will have to present the results of the task in a performance. During the task, the capacity for learners to attend to form can be manipulated through the setting of parameters like time pressure or the modality – spoken versus written. Attention to form can be promoted by decreasing time pressure because a lower time pressure usually allows learners more time to attend to form. Since oral production and perception usually imposes a higher time pressure than written production and perception, attention to form is better supported by the latter.

The second approach to creating a focus on form within a meaning-based task is known as **focused tasks** (Ellis, 2003). This approach is employed for this study, as we will describe in more detail in Section 7.2. Focused tasks are based on the idea that certain tasks lend themselves more readily to the use of certain linguistic structures. However, the degree to which it is feasible to design a task that creates a focus on certain forms is dependent on different factors, including (a) the kind of skill that is trained by the task, and (b) features of the structure itself. In general, pushing learners to perceive and comprehend a structure is easier than forcing them to produce a structure. Given their free choice of linguistic forms, learners can avoid producing a structure, but they cannot avoid being subjected to input. Characterizing the range of relationships between a task and a linguistic form in focus, Loschky and Bley-Vroman (1993) proposed three graded properties: The use of the structure can be *natural*, *useful*, or *essential* for the task. A structure is natural for a task if the structure is likely to arise naturally and frequently during the task. A structure is useful for a task if the task can be completed more efficiently with the structure. A structure is essential for a task if it is impossible to complete the task without using the structure. Of all the three properties, essentialness is the hardest to achieve in task design. Loschky and Bley-Vroman indeed admit that it can be difficult, if not impossible, to devise tasks for which a specific structure is essential.

4.5.3 Evaluating tasks and communicative interaction

A variety of studies have been conducted with the goal of gaining empirical support for the theoretical arguments for communicative interaction and task-based instruction. Evaluating task-based instruction can be done on two levels, the macro- and the micro-level (Ellis, 2003): The macro-level considers a complete task-based program, usually one or several courses spanning weeks, months, or even years. Such evaluations are usually requested by stakeholders of the language program, and only in

a second step used for scientific purposes. Given the inherently practical purposes, there are a number of problems for the scientific interpretation of such evaluations. For example, it is usually not possible to randomize the samples of the experimental and control groups. Further, it is normally also not feasible to assess prior knowledge by administering pretests. Finally, there is a lack of control about what is actually happening in the classroom – some teachers might not be able to implement the task-based approach fully as intended by the program designers (Ellis, 2003).

One example for a macro-level study is the evaluation of the Bangalore/Madras Communicational Teaching Project by Beretta and Davies (1985). It is usually cited as positive evidence of the effects of task-based language learning. With a battery of carefully selected posttests, which included tests that were more task-oriented and tests that were more similar to the traditional grammar-focused instruction, as well as neutral tests, Beretta and Davies compared the performance of a task-instructed group with the performance of a group that received traditional instruction. The results showed that the task-instructed learners were better in the acquisition of forms that they had not been explicitly taught and more ready to actually use the forms they had learned than the traditional group. However, the authors cautioned that these results should be considered as a probe rather than as proof of the effectiveness of task-based instruction, given that their study suffered all of the limitations mentioned above. According to Ellis (2003), macro-level studies in general fail to produce convincing evidence for the superiority of task-based instruction because it is hard to overcome these limitations.

An alternative to examining entire language programs on a macro-level are micro-level studies, which focus on the evaluation of one particular task tested with a particular group of learners. Such an evaluation is easier to control and therefore the results are usually more reliable. Ellis (2003) distinguishes between three aspects of task evaluation: students, responses, and learning. A student-based evaluation investigates the attitude of the students towards the task by probing whether they found it enjoyable and/or useful. A response-based evaluation examines the outcome of the tasks; it checks whether the learners solved the task as it was intended by the designer. For tasks that focus on specific linguistic forms, a response-based evaluation checks whether the learners used the targeted forms. Finally, a learning-based evaluation tests if the task has created any learning gains. The problem with learning-based evaluations is that the learning effect of a single task might be too subtle to be measured. This is problematic in particular for unfocused tasks, but as we will discuss later, it can also be an issue with focused task.

Communicative interaction is usually implemented in task-like contexts, even though the researcher may not always explicitly refer to the task-based approach. Studies that examine the effect of communicative interaction show a tendency that is similar to general task-based evaluations. The effect of communicative settings that have no particular focus is weaker than the effect of communication that targets specific linguistic forms (Muranoi, 2007). An example of a successful communicative task is described by Mackey (1999). She showed that learners of English improved their ability to form questions through engaging in an interactive task with native speakers. However, this positive result may be restricted to certain target structures and may not be generaliz-

able to a wider set of linguistic forms.

In general, it has been shown that communicative tasks are a good means of prompting learner output (Muranoi, 2007), but as Swain (1995) has shown, the mere production of output does not necessarily entail an increase in accuracy for linguistic forms.

4.6 Summary

In this chapter, we have established the relevant background of SLA research. The chapter started by discussing different approaches to language instruction that put different emphasis on meaning and forms. We introduced a classification that distinguishes between FOCUS-ON-FORMS, FOCUS-ON-MEANING, and FOCUS-ON-FORM. The third was introduced as an attempt to combine the advantages and compensate the disadvantages of the previous two. It is characterized by the effort to draw attention to forms only when they become relevant in a primarily meaning-oriented context. The vagueness of that definition led to a variety of practical implementations that differ among others with regard to the degree the instruction is planned and proactive or spontaneous and reactive. They further vary with regard to how exactly meaning and form are integrated.

The chapter then looked at the dichotomy of implicitness and explicitness regarding learning, instruction, and knowledge. After defining both implicit and explicit learning and instruction, it presented a summary of existing evidence for implicit and explicit learning and the effects of instructions. It then defined implicit and explicit knowledge and gave an account of the interface debate that is concerned about how implicit and explicit knowledge relate to each other and to what extent one can be transformed into the other. In the end of that section we presented possible measures for both types of knowledge.

The chapter further discussed how certain properties of linguistic structures have an impact on how these structures are learned and how they can be taught. Relevant properties comprise salience, frequency and regularity, functional value, and processability. Furthermore, we also discussed the concept of developmental stages that learner go through and that impose constraints on the order in which certain structures can be acquired or learned.

The chapter finished with a presentation of conversational interaction and task-based instruction which are both closely related to the FOCUS-ON-FORM approach. Conversational interaction comprises a set of different concepts and hypotheses that attempt to explain how the participation in a conversation can benefit foreign language learners. Tasks are tools to create a FOCUS-ON-FORM approach by creating a meaningful context and providing opportunities to use language in situations that resemble real life situations. Focused tasks try to elicit the use of certain linguistic structures in a natural unforced way.

Central to understanding the benefit of conversational interaction is the effect of feedback. The following chapter will discuss in more detail the value of feedback. It will combine the two perspectives of SLA and ICALL by classifying feedback that is provided in SLA contexts and relating it to the technological conditions to provide such feedback through an ICALL application.

5 Feedback

5.1 Introduction

One of the essential elements of communicative interaction is the feedback that learners receive in response to their production. We have already briefly discussed above the role of feedback both within conversational interaction (Section 4.5.1) as well as in the context of language learning software systems that provide feedback (Section 2.2 and 3.2). In this chapter, we will take a closer look and elaborate further on the aspects that are relevant to set the background for our study.

Feedback, in a very general sense, is understood as the reaction or response to some antecedent activity which contains information about the effect or consequence of the activity. In the context of language learning, feedback is the response to learners' language production; it provides learners with information that indicates how successful their production attempt was. The feedback can refer to the accuracy, the communicative success, or the content of the learner production (Leeman, 2007). Feedback that responds to a problem or an error in a learner's production is of particular interest in the language learning context, given the frequency of erroneous (or non-target-like) utterances in learner language. This kind of feedback is termed *corrective feedback*, since its aim is to correct the learner error. Beyond the general purpose of correcting, it is possible to make a more fine-grained distinction between different objectives and the corresponding effects of corrective feedback. Chaudron (1988), for instance, distinguishes three different levels. The most basic goal of corrective feedback is to simply inform the learners that they made an error. A more ambitious goal is to elicit a revised learner response. Finally, the most ambitious goal of corrective feedback is to induce a permanent adaption of the learner's L2 knowledge, which would prohibit any future errors of the same kind (Chaudron, 1988).

An important concept related to feedback is *evidence*: information about whether certain structures are permissible in the target language (Leeman, 2007). Positive ev-

idence consists of information about what is grammatical or acceptable, while negative evidence is information about what is ungrammatical or unacceptable in the second language (Ellis and Sheen, 2006; Leeman, 2007). Positive evidence is primarily provided through naturally occurring utterances, but also through explicit grammar teaching and corrective feedback. Negative evidence is provided particularly through corrective feedback, but also through explicit examples of incorrect structures.

The role of negative evidence and corrective feedback for acquiring a second language is rather controversial (Truscott, 1996; Ferris, 2004; Long, 2007). For one thing, it is disputable whether corrective feedback is necessary or beneficial at all. For another thing, it is an unresolved issue what kind of corrective feedback may be the most effective. In the next section (5.2), we will shortly discuss the first issue by presenting the different positions and arguments for and against corrective feedback. In Section 5.3, we will then give an introduction to the different types of feedback that language teachers can use and discuss their properties and the implications on their effectiveness. In Section 5.4, we will consider feedback from the perspective of a computational system and elaborate on what kind of information and knowledge is necessary to provide automated feedback. We will conclude with a more extensive discussion of two specific types of feedback - recast and metalinguistic feedback, which we implemented and tested for the present study (Section 5.5).

5.2 Necessity and benefit of feedback

Arguments in the discussion about feedback are based on theoretical assumptions about the process of language acquisition, on intuition and anecdotal evidence, and, increasingly, also on the results of empirical studies. However, empirical evidence is still fragmentary and at times inconclusive, which has resulted in contradictory conclusions and passionate debates (Truscott, 1996; Ferris, 1999). In general, the discussion about the role of corrective feedback comprises two strands. The first strand regards the theoretical necessity of feedback, while the other strand revolves around the effectiveness and potential disadvantages of feedback.

5.2.1 Necessity of corrective feedback

There are two opposite positions on the necessity of corrective feedback. Whereas one views it unnecessary for acquisition, the other considers it as indispensable for acquisition. The first position is based on the nativist view on language acquisition, which claims that humans acquire a language by the virtue of their innate ”language acquisition device” (Chomsky, 1965). Arguments for the existence of such a device are based on the rationale that the input that language learners receive is insufficient to account for their eventual competence. In particular, it is argued that children receive no or only negligible negative evidence during their first language acquisition. By assuming that second language acquisition (SLA) is similar to first language acquisition (FLA), it follows that second language learners do not need negative evidence in order to acquire the L2. Empirical support for this position is provided by studies that show that learners do acquire certain L2 principles that they have not been explicitly taught

(Bley-Vroman et al., 1988; Cook, 2003; Kanno, 1997; Pérez-Leroux and Glass, 1999).

However, there is reason to believe that SLA and FLA do in fact differ. The most convincing indication for this is that there are only rare cases of adult L2 learners who achieve native-like proficiency in their L2, as we have discussed before in Section 4.2.2. Furthermore, it has been argued that negative evidence is indeed necessary for acquiring certain structures (White, 1987, 1991). White argues that positive evidence alone cannot inform a learner about the fact that, for instance, null subjects are ungrammatical in English. For learners with an L1 that allows null subjects, this information is hard to conclude from positive evidence alone. A nativist objection to this argument is given by Schwartz (1993). Schwartz claims that – even if negative evidence might be necessary – it cannot be used for the development of implicit knowledge. She argues that, opposed to positive evidence, which consists of natural language utterances, negative evidence consists of information *about* the utterance. According to Schwartz, only the former is appropriate and processable information for the direct development of linguistic ability, i.e., implicit knowledge. Due to its form, negative evidence can only be used for the development of explicit knowledge. This view goes back to Krashen (1982, 1985), who distinguished between language acquisition (implicit) and language learning (explicit), which, in his view, are mostly unrelated. However, as we have discussed in our account of the interface debate (Section 4.3.5), there is also a view that explicit knowledge can indeed be turned into implicit knowledge and that feedback can directly or indirectly contribute to implicit knowledge. Ellis et al. (2006) provide empirical evidence that supports this position. In their study, learners who received metalinguistic information about their errors regarding the English past-tense *-ed*, did improve their implicit knowledge, as measured by an oral imitation test.

5.2.2 Effectiveness, benefit and potential harm of corrective feedback

Independently of the question whether feedback is necessary or not, it is reasonable to ask whether feedback can support language acquisition. Even if one argues that feedback is not necessary, it might be the case that feedback accelerates the acquisition process and makes teaching more efficient. After all, in consideration of limited time and resources, teachers and learners are interested in achieving maximal effect with minimal effort.

Truscott is a prominent opponent of corrective feedback in the classroom and has repeatedly argued that it is not effective (Truscott, 1996, 1999b). The heated debate that followed his objections indicates that there was a general assumption that feedback is useful for learners, which he challenged insistently (the debate is covered in Ferris (1999); Truscott (1999a); Ferris (2004); Chandler (2003); Truscott (2004)). Truscott argues that it is difficult for teachers to always correctly interpret the cause of an error. This makes it hard to give appropriate and clear feedback to learners. Further, teachers might not notice an error, or, if they notice it, they might consciously abstain from correcting in certain situations because a correction may conflict with the primary goal of the lesson. According to Truscott, these problems are likely to result in inconsistent feedback, which is hard to interpret for the learner. In addition, it cannot be taken for granted that learners do notice the feedback.

Even if learners do not notice the teacher's feedback or only part of it, the teacher's

efforts might not be spent wisely. Apart from the potential lack of efficiency and impact, an even stronger objection against feedback is that it can have a detrimental effect on learner production. Truscott (1996) cites references that indicate that correction of written text decreases the amount, content quality, and complexity of subsequent writing of students (Semke, 1984; Kepner, 1991; Sheppard, 1992). He attributes this to the unpleasantness of corrections and the aim of learners to avoid future corrections. Truscott (1999b) further argues that public oral correction has the potential to embarrass, inhibit, and produce feelings of inferiority for some students. However, he does not support this argument with empirical studies about the actual pervasiveness of such a negative effect. On the other hand, proponents of feedback cite the pervasive wish of learners to receive feedback and argue that ignoring these learner needs can lead to frustration (Ferris, 1999).

Another argument put forward in favor of corrective feedback is that its absence would promote *fossilization*, i.e., the permanent halt of any further interlanguage development towards the target language (Selinker, 1972). A prime example for this is the experience with immersion students in Canada who received no grammar-related feedback and exhibited poor grammatical accuracy despite years of instruction (Swain, 1985). However, the mixed evidence collected so far does not justify a general claim that corrective feedback has the potential to inhibit fossilization, nor that the lack of feedback causes fossilization.

Apart from that, more theoretical arguments for the benefit of feedback draw on the INTERACTION HYPOTHESIS and the NOTICING HYPOTHESIS (Schmidt, 1990, 1993). According to the interaction hypothesis (Long, 1996), feedback is a crucial factor for learners to modify and improve their production, as we have discussed in more detail above in Section 4.5.1. With regard to the noticing hypothesis, feedback supposedly helps learners to notice particular forms because it draws attention to errors (see also Section 4.2.3 (p. 60) and Section 4.3.1 (p. 64)).

When we consult empirical evidence for supporting any of the more theoretical position, the existing studies present no consistent picture. The reasons for inconsistent results lie in the variety of contexts, parameters, and evaluation methods used in the different studies. However, recent summaries and meta-analyses on corrective feedback draw a predominantly positive conclusion – corrective feedback does have a positive effect on learners' language development (Russell and Spada, 2006; Sheen, 2010b; Mackey and Gass, 2006; Mackey, 2007; Lyster and Saito, 2010). While Truscott's objections (1996; 1999b) were successful in challenging an overly intuitive assumption about the worth of feedback, the selection of evidence he cited for making his point was arguably biased and has since been qualified by the increase of new evidence that documents the beneficial effect of feedback. Truscott is justified, however, in pointing out the practical problems of implementing corrective feedback in the classroom. The fact that feedback seems to be more effective in laboratory settings than in classroom settings (Nicholas et al., 2001; Li, 2010) supports this objection. Nonetheless, there is positive evidence for the benefit of corrective feedback in classroom settings as well (Loewen and Philp, 2006; Doughty and Varela, 1998).

5.2.3 Further issues in feedback research

Despite the existing positive evidence for the benefit of corrective feedback, the body of research is still fragmentary as the role of different additional parameters has not been examined thoroughly yet. While the study that we are going to present in this work explores the effect of feedback in a ICALL context, it does not address other interesting factors that are important for an assessment of feedback. Individual differences between learners, like aptitude, motivation, age, learning styles, memory, personality, anxiety, and learner beliefs have been relatively neglected so far (Ellis, 2010). Related to the issue of classroom versus laboratory settings discussed above, feedback might also operate differently depending on how much input learners get outside the classroom. Learners who live in a context where the target language is spoken by the majority of people outside of class might benefit differently from feedback than learners whose only exposure to the language takes place during class time (second language versus foreign language context), and yet differently from students in immersion programs, who are exposed to the target language throughout their entire school day in all lessons and activities outside of class (Lyster and Saito, 2010; Mackey and Goo, 2007; Li, 2010).

Another issue relates to the evaluation methods for assessing the learning outcomes. We already discussed implicit and explicit knowledge, the means to measure them, and the importance of considering them both (Section 4.3.6). A further distinction of learning gains is made by Ellis (2010) who discusses the different meanings of the term *acquisition* and how they figure in assessment of learning. Ellis distinguishes (a) acquisition of an entirely new linguistic feature, (b) the increase in accuracy of partially acquired features, and (c) acquisition as characterized by a progress along a sequence of stages. Ellis claims that most studies that evaluate the effect of feedback on acquisition are based on (b) and a few are based on (c). He was, however, not aware of any study that measured learning outcome as the acquisition of an entirely new feature (a). He attributes this to the difficulty of finding linguistic features that are entirely unknown. Although Ellis seems to have overlooked at least the study of Long et al. (1998), in which they do ensure that the subjects have no prior knowledge of the Spanish target structures, this instance does not contradict the observation that this kind of measure is only very rarely considered.

5.3 Classification of feedback

In this section we will discuss different types of corrective feedback that have been identified through examining interaction in language classrooms. After a general description we will further introduce different parameters that are relevant for language learning and use them to characterize the feedback types further. This is important for this thesis because it provides the necessary background to decide which type of feedback we will examine in relation to the dichotomy between explicit and implicit instruction and focus on meaning versus form.

5.3.1 Types of feedback

Feedback is most commonly classified according to the taxonomy established in the seminal work of Lyster and Ranta (1997), who analyzed the feedback strategies used by teachers of French in Canadian immersion classes. Note that these strategies are used in the context of primarily oral classroom interaction. The classification of error correction for improving the quality of students' writing is different and not considered here. Although the computer interface that we employ to study feedback in a ICALL context is based on type-written interaction, the communication is immediate in nature, therefore the oral interaction feedback types are more relevant here. Lyster and Ranta listed the following types of feedback:

- Explicit correction
- Recast
- Clarification Request
- Metalinguistic Feedback
- Elicitation
- Repetition
- Translation

An **explicit correction** provides the correct form and clearly indicates that the learner utterance was problematic. As an example consider (1-a), in which the determiner "der" is incorrect. The correct determiner is the dative masculine form "dem":

(1) a. Learner:
Das Kino ist neben *der* Buchladen.
The cinema is next to the$_{incorrect}$ book shop.

b. Teacher:
Es muss heissen: "neben **dem** Buchladen".
It should be "next to the$_{correct}$ book shop"

Recasts are understood as reformulations of all or part of the learner's preceding utterance, in which one or more errors are replaced with the correct forms. However, it is not necessarily obvious to the learner that the reformulation is meant as a correction. (2-a) and (2-b) provide two examples for a recast of (1-a):

(2) a. Das Kino ist neben *dem* Buchladen.
The cinema is next to the$_{correct}$ book shop.

b. neben *dem* Buchladen.
next to the$_{correct}$ book shop.

Recasts can further be classified with regards to whether they are isolated (as in (2-b)) or merged with additional information. The latter contain or seek additional content-matter information in addition to the reformulation (Lyster, 1998). As an example for an incorporated recast, consider (3)

(3) Also das Kino ist neben *dem* Buchladen. Und was ist hinter dem Kino?
Ok, the cinema is next to the$_{correct}$ book shop. And what is behind the cinema?

Clarification requests indicate that the learner utterance was either incomprehensible or inaccurate so that a repetition or reformulation is required. This feedback move is usually realized as a non-specific "Pardon?" or a more elaborate "What do you mean by X?"

Metalinguistic feedback provides "either comments, information, or questions related to the well-formedness of the student's utterance without explicitly providing the correct form" (Lyster and Ranta, 1997, page 47). This type of feedback indicates that there is an error, and can also give hints about the nature of the error, usually by using linguistic terminology. (4-b) provides an example:

(4) a. Learner:
Das Kino ist neben *der* Buchladen.
The cinema is next to the$_{incorrect}$ book shop.

b. Teacher:
Der Artikel "der" in "neben **der** Buchladen" ist nicht richtig .
The article "der" in "neben **der** Buchladen" is not correct.

Elicitations prompt the learner to reformulate their erroneous utterance by asking questions like "How do you say that?" or "How is this called in German?" or by explicitly asking the learner to reformulate. The teacher can also elicit a reformulation by providing the first part of an utterance and then pausing to allow the learner to complete the utterance. The expected completion would contain the problematic form.

Repetitions indicate that there is an error by repeating the erroneous utterance in isolation and by using a distinct intonation to emphasize the error.

Translation is feedback given in response to unsolicited L1 utterances of the learner. Translations, like recasts, reformulate a non-target-like learner utterance. Unlike recasts, they do not follow an erroneous L2 utterance but an obvious failure of the learner to produce an L2 utterance. Translations were not part of the original set of feedback types presented in Lyster and Ranta (1997) because they were so infrequent in the data they had analyzed. However, Panova and Lyster (2002) have treated translations as a separate category, because they were considerably more frequent in their data set.

Distribution of feedback types

Feedback type	Explicit/ Implicit	Prompts learner modification	Provides correct form	Indicates location of error	Indicates nature of error
Explicit correction	explicit	no	yes	yes	no
Recast	implicit	no	yes	yes	no
Clarification request	implicit	yes	no	depends	no
Metaling. feedback	explicit	depends	no	depends	depends
Elicitation	depends	yes	no	depends	depends
Repetition	explicit	depends	no	depends	no
Translation	implicit	no	yes	n.a.	n.a.

Table 5.1 – Feedback strategies and their properties

Lyster and Ranta (1997) found that recasts were the most frequent feedback strategy in the data they analyzed – they constituted 55% of all feedback moves. The prevalence of recasts has been confirmed by subsequent studies that analyzed the distribution of feedback types in different classrooms – the proportion of recasts was between 54% and 65% (Panova and Lyster, 2002; Suzuki, 2004; Lyster and Mori, 2006; Mc Carthy, 2008). In the study of Lyster and Ranta, elicitation and clarification requests are the second most frequent feedback strategy with 14% and 11% each. The other studies reveal a slightly different distribution for the further types of feedback. For instance, while Suzuki (2004) finds 38% clarification requests and only 6% elicitation, Mc Carthy's data show 27% elicitation and only 3% clarification requests (Mc Carthy, 2008). While these differences might be due to the different countries and teaching cultures, Mc Carthy's comparison of three different teachers suggests that there is a considerable difference between individual teachers. Metalinguistic feedback, repetition, and explicit correction each make up less than 10% of feedback moves in the considered studies. Translation as a feedback strategy was only coded by Panova and Lyster (2002), where it amounted to 22%. In the other studies, it was either non-existent, or merged with recasts. The absence of translations is not surprising in a second language learning context (as opposed to a foreign language learning context), since students in this context cannot assume that the teacher has knowledge about their native language, which makes it unlikely that they attempt to use their L1 in the classroom.

5.3.2 Parameters of feedback

The different types of feedback can be further classified according to criteria that are relevant for the language acquisition process. Feedback can be explicit or implicit, it can provide a correction or prompt the learner to provide one, and it can include the location and/or nature of the error or not. Table 5.1 summarizes the criteria for the different feedback types, which we will discuss in more detail below.

Explicitness

One important criterion is whether the feedback is explicit or implicit in nature. This criterion is closely related to the distinction between explicit and implicit learning, instruction, and knowledge that we discussed in Section 4.3. Explicit feedback expresses clearly that an error occurred and thereby it usually interrupts the meaning-based flow of conversation. Implicit feedback, on the other hand, is integrated into the subject matter conversation. As a consequence, it is harder for the learner to recognize and correctly interpret the corrective intention of implicit feedback, a fact that can undermine the beneficial effect of feedback. This problem has been widely discussed for recasts, the prototype for implicit feedback, and we will recount this discussion in more detail below in Section 5.5.1. Similarly, clarification requests are considered an implicit type of feedback since they indicate a problem with understanding but give no explicit hint to the cause of misunderstanding (Loewen and Nabei, 2007). Explicit correction and metalinguistic feedback are explicit types of feedback, since they make explicit that an error occurred. Elicitations can be implicit or explicit, depending on the particular form they take. Rezaei et al. (2011) argue that an overt request for reformulation and an open question are rather explicit forms of elicitation, while the use of pauses to allow learners to complete an utterance is more implicit since it is less intrusive to the flow of communication.

Repetitions are characterized as an implicit feedback type by Rezaei et al. (2011) and Loewen and Nabei (2007) without further explanation. However, we would argue that repetitions tend to be more explicit since they are normally not a plausible part of communication, and even if so, they serve other purposes, for instance, expressing disbelief. A repetition disturbs the flow of communication, although it might not explicitly point to the error. Translations, on the other hand, are implicit in that they do not indicate overtly that an error was produced, because the error was rather a failure to produce by using the L2 altogether (Rezaei et al., 2011). Although translations cannot be considered part of the subject matter dialog since the dialog contribution was already given by the learner in their L1, they do not severely disturb the flow of communication, since the learner is not required to react on them. They can be understood as a comment to the subject matter dialog which does not require a response.

Pushing for modification

Another defining criterion of feedback is whether or not it pushes learners to modify their production. This criterion is related to the previous one since the obligation to modify output is, similarly to explicit feedback, likely to disturb the flow of communication and divert the focus from meaning to form. Feedback that does not prompt a learner modification is less likely to disturb the flow, similar to implicit types of feedback. This criterion is important because the obligation to modify output and correct errors has repeatedly been argued as being facilitative for language learning (Swain, 2005; Lyster, 1998; Panova and Lyster, 2002, and references therein). De Bot (1996) reasons that production (of output) is more effective than perception (of input) in strengthening the connections in memory and proceduralizing knowledge because it requires more attention. In general, this argument can be traced back to research in

psychology that shows that the depth of processing is correlated with retention rates (Craik and Lockhart, 1972). Attention and engagement of the learner deepen the level of processing. More specifically, this phenomenon has been conceptualized in the *generation effect* (Buyer and Dominowski, 1989; Jacoby, 1978), which purports that subjects retain items better when they have to actively retrieve (*generate*) them instead of passively perceiving them.

In the context of second language acquisition, Ferreira et al. (2007) show that feedback that pushes learners to modify their output results in a higher rate of self-repair than feedback that provides the expected target forms. The authors argue that self-repair indicates that the learner is aware of the mismatch between their initial utterance and the target utterance, which "can be a first step towards improvement", but within the scope of the study, improvement is not assessed (Ferreira et al., 2007, page 18). Actual improvement as measured by four different tests before and after the instruction has been found by Lyster (2004) - he found that feedback that prompts for self-repair from the learner yields greater learning gains for assigning the correct grammatical gender to French nouns than feedback that does not. Similarly, Ammar and Spada (2006) show that feedback that pushes learners to correct is more effective than feedback that provides the correct form for low-proficiency learners, but not for high-proficiency learners. French learners of English had received recasts or prompts regarding third-person possessive determiners *his* and *her*. Izumi (2002) investigated the acquisition of English relativization, he indicates that instruction that requires learners to reconstruct a text is more effective for learning than instruction that does not. Opposed to that, however, Lyster and Izquierdo (2009) cannot find a difference between the effect of recasts (providing corrections) and prompts (asking for corrections) in dyadic interactions for learning the grammatical gender of French nouns. A limitation for prompting for correction is that it requires that learners have the necessary knowledge to correct their error, therefore it is problematic for structures that are unknown to the learner (Loewen and Nabei, 2007).

Table 5.1 shows for each feedback type whether it provides the correct form or not. Explicit corrections, recasts, and translations provide the correct form and thereby do not oblige the learner to modify their output. Clarification requests and elicitations do not provide the correct form and directly request the learner for a modification of their output. Metalinguistic feedback and repetitions also withhold the correct form, however, the obligation for the learner to repair their error is arguably not as obvious and direct as in clarification requests and elicitations.

Information content of feedback

The provision of the correct form, as we just discussed it in relation to pushing for modification is also part of the information content of feedback. Furthermore, there are two other related criteria that regard the error-related information given to the learner. The feedback can indicate the location of an error, and in a more informative version, the linguistic nature of the error. For the learner, the location of an error is a useful support for correcting it. But only a more detailed explanation of the error supports the learner in generalizing the problem beyond the specific context.

Explicit correction and recast both contain information about the location of an

error, but do not explain the nature of the error. If the recast is not isolated but embedded in a larger utterance, the location of the error may not be obvious to the learner. Whether clarification requests, metalinguistic feedback, elicitations, and repetition indicate the location of the error depends on their specific realization. Clarification requests and repetitions do not contain information about the nature of the error, whereas for metalinguistic feedback and elicitations, it depends on their specific realization.

From the perspective of an intelligent ICALL application, the criteria that we discussed above are important because they determine how much and what kind of additional linguistic knowledge is required for realizing the respective feedback. For instance, the ability to provide a correction as part of the feedback requires a hypothesis about what the learner wanted to say. Deriving such a hypothesis is not a trivial task. Given that even human instructors sometimes have difficulties with this task, it can be a considerable challenge for a computer system. So far, we have presented the different types of feedback that have been observed in classroom interaction and we have characterized their properties mostly from a learner's perspective. The feedback, as we have discussed it, is given by teachers or competent speakers. We will now discuss feedback with a view on how to generate it automatically within a ICALL application.

5.4 Feedback in the ICALL context

In this section we present problems that are relevant if feedback is given by computer applications rather than human teachers. We will first consider general issues of automatic feedback provision (Section 5.4.1) and in the second part characterize the requirements for implementing specific types of feedback (Section 5.4.2).

5.4.1 General issues in ICALL feedback

General issues for feedback given through intelligent systems that we will discuss in this section regard (1) the specific requirements that learners have opposed to native speakers, (2) the hypothesis about the intended utterance of the learner, and (3) the balance between the cost and benefit of informative feedback.

Learner requirements

A particular requirement for feedback provided by a computer to a learner is that it be reliable. Since the learners' knowledge about the language they learn is incomplete and just developing, they cannot be expected to have the competence to recognize inappropriate feedback, unlike native speakers (Amaral and Meurers, 2011). It is therefore very important to avoid such inappropriate feedback, because it can confuse or even mislead learners (Heift and Schulze, 2007). One strategy for this is to only give corrective feedback for unequivocal, very certain cases, but to avoid feedback in less certain cases. This approach entails that some errors may be left unnoticed, but this is

considered as less harmful than to falsely correct a non-erroneous utterance. A complementary strategy to compensate for this shortcoming is to make learners aware of the limits of the system such that they will know that not all of their errors will be recognized (Levin and Evans (1995) provide references for such an approach).

Error hypothesis and extent of expected learner input

Another issue that we have discussed already above in the context of ambiguity (Section 2.2.2), is the problem of determining what the learner had intended to produce. Such a hypothesis is often essential for reasoning about the cause of an error and for knowing which grammatical rule was violated (Nerbonne, 2003). Knowledge about the causes is necessary to give more detailed information about the error and provide one of the more informative variants of feedback. However, as we have illustrated before, this is not a trivial problem, because some errors have multiple possible sources (Heift and Schulze, 2007; Meurers, 2012; Schwind, 1995).

A factor that affects this problem is the space of possible and expected learner input. The error diagnosis approaches that we have discussed above (Section 2.3) were implicitly based on the assumption that the learner produces free, unconstrained written text. However, in the context of the current study, as we will describe in more detail in Section 7.2, the learner production is guided by a task-driven real-time dialog with a computer system. In this context, the space of possible and expected utterances is usually much more constrained than it is with free unrestricted monologic text production. If the expected input is more constrained, the representation of possible input can be less sophisticated and the coverage can be less comprehensive. Along with a restriction for the input, the space for potential errors is more restricted, which makes it more feasible to use anticipation-based error diagnosis.

Balancing cost and benefit of informative feedback

Regardless of the extent of possible learner input, the informativeness of feedback is usually related to the complexity of the error diagnosis approach. There is evidence that more informative feedback can be beneficial for the language learner, as we will discuss in more detail below in Section 5.5.2. Nagata (1993), for instance, showed that intelligent feedback that explains the nature of an error, based on a linguistic analysis of the learner input, can be more effective for the acquisition of Japanese particles than traditional, less informative feedback. However, Heift (2010b) argues that "from a computational point of view, the more detailed the error explanation, the more laborious and elaborate the error checking mechanism. For this reason, a reasonable prospect of benefit must be weighed against the development cost both in terms of time and expertise" (page 204). In order to appropriately balance the expected benefits against the efforts to spend, one needs an appropriate estimation for the effects of different types of ICALL-delivered feedback. One further needs an estimation of the development costs for specific feedback types. We will present the existing evidence about the benefit of specific types of feedback in the next section (5.5). In the following section 5.4.2 we will approach the assessment of development effort by characterizing different feedback types in terms of the information and models that are needed to

realize them. This extends the foregoing discussion about the information content of feedback (5.3.2) by adding a computational perspective.

5.4.2 Information requirements for different feedback types

Based on our characterization of information content in Section 5.3.2, we will now discuss for each of the feedback types what kind of information is necessary to provide the feedback. Recall that the information content of feedback can be defined by whether or not it provides a correction, indicates the location of the error, or explains the nature of the error. Note that the information contained in a certain feedback is not necessarily equal to what the learner perceives and to what the feedback provider must know. Here, we will focus on the latter. We will see that some of the feedback types of the classification based on Lyster and Ranta (1997) have different variants that differ in their specificity and information content, therefore there is not straight-forward one-to-one mapping between feedback type and the kind of information it requires.

For a **clarification request**, it is, in general, not necessary to know about the nature of an error, its location or correction. The system can produce an utterance like "Pardon?" or "Sorry, can you say that again?" to indicate that the learner utterance is erroneous or unexpected. There is no need to anticipate or model errors, it suffices to reject everything that is not within the range of expectations. However, for the learners, it will be unclear whether their production was erroneous or just not expected and interpretable by the system. For more specified, and therefore, more informative clarification requests, like, for instance, "What do you mean with X?" or "Do you mean X?", the system needs more information. Consider the clarification request "What do you mean with X?", where X is a placeholder for an unknown or incorrect word or phrase. In order to identify the unknown word or phrase, which is part of a larger utterance, it is necessary to analyze the learner input as being composed of smaller units instead of complete utterances. Such smaller units can be, for instance, words, and, given a lexicon of known words, the system can identify unknown words.

For a clarification request like "Do you mean X?" it is necessary to have a hypothesis about the intended utterance. Such a hypothesis is also necessary for an **explicit correction**. However, for the explicit correction, the system's confidence that this is indeed the correct version should be sufficiently high. Opposed to that, for a clarification request, the confidence can be lower, because a wrong hypothesis would not be so severe, if it is given as a suggestion (as in a clarification request) instead of as a command (as in an explicit correction). In summary, the amount of knowledge required for providing a clarification request depends very much on the informativity of the request which can range from simply identifying an input as being unexpected or incomprehensible to reasoning about what the learner might have intended to produce.

To deliver a **recast** or an **explicit correction**, it is necessary to have a hypothesis about what the intention of the learner was. This does not necessarily require knowledge about the location of the error nor does it require a notion of the grammar rules that have been violated, but both might be helpful to arrive at a possible correction of the error. An approach that does not need error-specific information is to compare the

learner input with the set of possible expected inputs and adopt the closest match as a correction hypothesis. Finding the closest match can be based on a string similarity metric, e.g., the edit distance (Gusfield, 1997), but it can also take into account the similarity between larger units of the string, e.g., words, chunks, or phrases, and their combinations. An approach that uses words as units is also be better suited to discover word order rule violations.

Recasts can vary with regard to how much of the context of the utterance is re-used – it can be limited to just the corrected form or it can include additional, correct parts of the original utterance. Some recasts are also integrated into new content that forwards the conversation. For integrating the recast with new material, the system requires techniques for language generation.

Similarly, explicit corrections can vary in informativeness. The most economic but probably for the learner least valuable option is to deliver a closest match introduced by the explicit announcement that the original utterance was wrong, e.g., "This is incorrect. You should say <CLOSEST-MATCH>!" This requires a certain level of confidence, because suggesting an inappropriate alternative can be harmful. A disadvantage of giving the closest match as a whole is that the difference between the original utterance and the target utterance might not be evident to the learner, in particular, if the error is not very salient. It would be easier for the learner to recognize the erroneous form if it was provided in isolation.

In order to provide isolated recasts as well as focused corrections, the location of the error has to be known. This requires to identify the mismatch between the actual learner utterance and the expectation in order to provide the corrected portion. Depending on the type of error, however, it is desirable to provide a meaningful portion that includes the dependent constituents. Finding meaningful parts of the original utterance then demands a deeper understanding of the error. For instance, if the error regards a wrong agreement between subject and verb phrase, the system would need to provide only those in order to focus on the error. This requires a notion of the involved constituents and dependencies, i.e., knowledge about the syntactical structure. On the other hand, if the error is limited to a word form, e.g., an orthographic error or a wrong plural form, the system can just focus on this word. However, in order to correctly recognize the wrong plural, it might need a model that predicts wrong plural forms. In conclusion, the requirements for providing a recast and an explicit correction depend on the type of error and on the specific didactic purpose of a lesson, but cannot be characterized per se.

Metalinguistic feedback as defined by Lyster and Ranta (1997) does not necessarily require reference to linguistic concepts, as it can also be realized by giving a simple indication of the missing well-formedness of the learner utterance, as in "That is not quite right." However, in the scope of this thesis, we will narrow down the definition as to include a reference to linguistic concepts or grammatical rules that are relevant for the error. Understood in this way, metalinguistic feedback requires, in addition to aforementioned techniques to detect errors, a representation of grammar rules, the ability to detect how they are violated in an utterance, and the means to reason and communicate about it.

Repetition of an error requires only to recognize that an error occurred. If only

the erroneous bit should be repeated, the location needs also to be known. This type of feedback is more suitable in speech-based interaction, and a model of intonation is required in order to give the repetition the appropriate emphasis.

For **elicitation** feedback, the amount of knowledge depends on the specificity of the elicitation. A very general form, as, for instance, "How do you say that?" does not require any information about the location or nature of the error. A more specific elicitation that provides the first part of an utterance and requests the learner to complete it, however, is contingent on knowledge about the location of the error.

In summary, we have described on a general level what kind of information and processing techniques are needed for different variants of feedback, in order to be able to estimate the effort that is required for realizing different feedback. Since the effort needs to be balanced with the expected benefit, we will now proceed with a detailed account of the benefits that have been observed for the two feedback types that we examined in the present study.

Note that we have not discussed the additional issues and requirements that are related to the uncertainty of speech recognition in speech-based systems because within the scope of this work, we limit our focus on type-written dialog interaction.

5.5 Recast and metalinguistic feedback

We will now review the existing research regarding the effects of recasts and metalinguistic feedback. In order to understand why we selected these two feedback types for further examination in our study, note that one objective of the current study is to examine the difference between implicit and explicit instruction. A convenient way to control implicitness and explicitness is through the properties of feedback. Of all feedback types we consider recasts and metalinguistic feedback as the most prototypical feedback types for implicit and explicit instruction, respectively. This view is supported in related work, consider for instance the operationalization of implicit and explicit feedback discussed and implemented by Ellis et al. (2006).

Furthermore, within the context of the ICALL application we use in the current study – a type-written dialog system, recast and metalinguistic feedback have some advantages over other types of feedback. *Elicitation* and *repetition* are inconvenient as they do not lend themselves easily for type-written feedback, since they are typically realized with means reserved for speech. The teacher pauses and/or marks the problematic form by a different intonation, e.g., by raising their voice. While it might be possible to find type-written counterparts to prosodic features, ambiguity is likely to arise. *Clarification requests* are also potentially ambiguous, not in general but within the context of the tasks and target structures, that we chose (explained in more detail in Section 7 below). Learners could interpret clarification requests as being targeted at the content level of the task and not at the formal aspects. *Explicit corrections* of learner errors are another option for a rather explicit form of feedback. However, in contrast to metalinguistic feedback, which does not provide a correction, but merely indicates that there is a problem and thus prompts the learner for finding the solution on their own, explicit corrections provide the forms and do not require the learner to find the

solution. As we have argued above in Section 5.3.2, pushing the learner for a modification of their erroneous production has been shown to be beneficial, therefore we chose a prompting feedback.

We will take up the choice of the two feedback types again in a more comprehensive manner in the next chapter, at which point we will derive and explain the general approach we pursued for this thesis. For now, we hope the reader can accept this limited explanation ahead of a more principled justification.

In the remainder of this chapter, we start with a discussion of recasts, which includes factors that influence their effectiveness, and problems related to their implicit nature as well as ways to address these problems (Section 5.5.1). We then discuss metalinguistic feedback and its implementations in ICALL systems (Section 5.5.2). In the end of this section we review studies that explicitly compare the effectiveness of recasts and metalinguistic feedback (Section 5.5.3).

5.5.1 Recasts

The study of recasts as an *implicit* and *incidental* type of feedback is strongly related to the interest in FOCUS-ON-FORM approaches (see Section 4.2.3), since recasts allow to deal with students' language problems *incidentally* while working on content matter (Long, 2007) and they do not interrupt the flow of conversation. Long (2007) defines a recast as "a reformulation of all or part of a learner's immediately preceding utterance in which one or more non-targetlike (lexical, grammatical, etc.) items is/are replaced by the corresponding target language form(s), and where, throughout the exchange, the focus of the interlocutors is on *meaning*, not language as object" (page 77, emphasis in the original). The potential advantages of recasts, according to Long, are that they provide linguistic information in context when interlocutors share a joint attentional focus. Long further argues that the mapping of form to function is facilitated by the fact that the learner has a prior comprehension of (parts of) the utterance. Because learners are involved in the exchange, they are supposedly motivated and attending and thus more likely to notice forms. Since a recast immediately follows an erroneous utterance, the incorrect form is brought face to face with the correct form. This juxtaposition supposedly highlights the contrast between the correct and the incorrect form and makes it easier for the learner to notice their error (Saxton, 1997). It has to be noted that the involvement of the learners, their partial understanding of the utterance, and the direct juxtaposition of incorrect and correct form are not exclusive features for recasts only but also pertain to other types of feedback given in a synchronous communicative exchange.

Several empirical studies have provided evidence that the provision of recasts can have a beneficial effect for the acquisition of specific target structures (see, for instance, Loewen and Philp (2006); Mackey and Philp (1998); Long (2007) and references therein). However, is has also been shown that the effectiveness of recasts depends on different factors, most notably, the developmental stage and proficiency of the learners. Philp (2003), for instance, showed that learners of higher proficiency are more likely to notice and correctly interpret recasts than learners of lower proficiency. Apart from the proficiency level, individual differences in phonological sensitivity and working memory have an impact on the effectiveness of recasts – learners with a higher de-

gree of phonological sensitivity or working memory capacity benefit more from recasts (Robinson, 2001; Mackey et al., 2002). The efficiency of recasts is further modulated by characteristics of the linguistic structures they target. In general, it seems that recasts are more effective for more salient and meaning-bearing structures (Long et al., 1998). This relates to our discussion in Section 4.4.1, in which we explained that more salient structures are more likely to be noticed by the learner and therefore more likely to be learned.

Perception of recasts

While the implicit nature of recasts is a desired feature within the context of meaning-focused instruction, it is at the same time the cause of an important problem: Recasts are potentially ambiguous and hard to notice: "[...] learners might have no conscious awareness that the recast is intended to be corrective" (Ellis et al., 2006, page 341). Along these lines, Lyster (1998) argues that the corrective intention is hard to recognize for learners, because recasts can be mistaken for non-corrective repetitions, which have similar pragmatic functions. Recasts, as well as non-corrective repetitions, provide or seek confirmation of the learner's message and they also can both provide or seek additional information related to the preceding message.

Although there is evidence that learners do notice recasts under certain conditions, the amount of noticing might be considered insufficient. The learners in a study conducted by Roberts (1995) noticed one third of full recasts and 43% of partial recasts. However, noticing was tapped only afterwards: three student volunteers watched a recording of a 50 minutes class session they had attended several days before and were asked to indicate any instance of correction that they detected. It is questionable if this method can accurately measure the actual amount of noticing that takes place during the interaction, because it is likely that during the actual interaction the noticing rates are lower.

Another sign that shows that learners noticed a recast is *uptake* – an immediate response to the feedback that indicates that the learner has noticed the corrective intent of the feedback (Lyster and Ranta, 1997). In comparison to other types of corrective feedback, recasts induce a relatively small amount of uptake – roughly between 20 and 30%. For instance, Oliver (1995) examined dyadic interactions between non-native speakers and native speakers of English aged between 8 and 13. On average, learners incorporated one third of the recasts in their following utterance. Similarly, the data reported by Lyster and Ranta (1997) show that learners reacted on 31% of the recasts they were given. Although the other types of feedback elicited considerably more reactions, these numbers still indicate that at least some of the recasts are perceived as corrective. This finding is consistent with the results reported by Doughty (1994) – beginner level university learners of French repeated on average 22% of the corrective recasts they received during classroom interaction but only 2.3% of non-corrective, exact repetitions.

No matter whether one considers these rates of uptake as sufficient or not, it remains debatable if uptake is an appropriate indicator for learning gains. It has been shown that learners are able to employ the information that was provided through recasts regardless of the uptake they show (Loewen and Philp, 2006; Mackey and Philp,

1998). Further, Mackey et al. (2000) were able to demonstrate in a stimulated recall study that learners are able to incorporate feedback that they have not consciously perceived. Similarly, Smith (2005) found no relationship between uptake and the acquisition of target lexical items in his examination of computer-mediated text chat. All this suggests that uptake cannot be used as a direct measure to determine the learning effects of feedback.

Another problem with uptake as a measure of learning is that an immediate response to recasts is often impossible or inappropriate in the communicative setting of the classroom (Oliver, 1995). Given their status as implicit feedback moves, recasts are arguably not even intended to induce a repetition by the learner, since this could interrupt the flow of the meaning-based conversation.

Considering the perceptional problems, instructors and researchers have proposed methods to increase the perceptibility of recasts. One way is to focus on a small subset of target structures instead of reacting on the entire range of appearing errors (Nicholas et al., 2001). Another way to reduce the ambiguity and to increase the salience of recasts is to use prosodic and extralinguistic cues, e.g., facial expressions. In the study described by Doughty and Varela (1998), recasts were preceded by a repetition of the learner's error and the recast itself was realized with emphatic stress on the correction part.

In addition to this, Ellis and Sheen (2006) illustrate more means to increase the explicitness of a recast. The recast can be repeated, or a single word can be recast, instead of embedding the corrected part into a larger utterance. As Ellis and Sheen rightly notice, these modifications turn recasts into a rather explicit form of feedback and it is problematic for these cases to maintain the notion of recasts being an implicit type of feedback. Related to that, another reason to consider some types of recasts as explicit is given by studies that show that learners gain explicit metalinguistic knowledge after exposure to more explicit recasts (Long et al., 1998; Han, 2002).

Another factor that influences the perceptibility of recasts is the communicative context. In their review of recast studies, Nicholas et al. (2001) show that recasts in laboratory settings in dyadic interactions tend to be more effective than recasts in classroom settings. The authors attribute this difference to the fact that laboratory settings and the limitation to one-on-one interaction help learners to recognize the intention of the recasts, while in otherwise meaning-focused classroom contexts, recasts are more likely to be interpreted as confirming the communicative content of an utterance. In other words, recasts seem to be most effective when learners are aware that they refer to the form and not the content of their utterances.

As will become clear further below, the recasts employed in our study fall into the implicit end of the implicit-explicit range, because they are integrated into a longer utterance, and their salience is not increased by any enhancement. Also, they are not cast in a way to invite uptake. Although the interaction is dyadic between one learner and the system, the setting does not include any explicit hints that the system feedback refers to the form of the learner production.

Recasts in written chat interaction

Recasts and their perceptibility have been examined primarily in oral interaction, but there are also a few studies that analyze recasts in written chat interaction. Sachs and Suh (2007), for instance, explored the effect of textually enhancing recasts in a text-based chat between dyads of native speakers and learners of English. Important forms were underlined and set in boldface respectively. Although this kind of enhancement led to greater amount of reported awareness and awareness was related to posttest performance, no significant relation between the enhancement and developmental gains could be found. In general, one could hypothesize that recasts in written chat interaction are more noticeable than recasts in oral interaction, because the transcript of the interaction is more permanent and learners have more time to process the input that they perceive. However, at the moment, there seem to be no empirical results that would support this assumption. The only study that directly compares the rate of noticing of recasts in written chat versus oral face-to-face conversations cannot find a significant difference between the two modes (Lai and Zhao, 2006). However, the data that was available for this comparison was probably insufficient – it was based on the interaction protocols of only four participants. In absence of more comprehensive data, it would be premature to draw any general conclusion.

A potential problem of chat interaction is that the sequence of turns that constitute the conversation is often interleaved and related turns are not necessarily adjacent. For instance, a question does not have to be followed immediately by its response, but other turns belonging to a different topic can be issued in between. Since the direct adjacency of recasts with the erroneous utterance has been argued to be supporting the learners to notice their error, the question arises whether learners do notice recasts that are not directly following the erroneous utterance. Lai et al. (2008) examined the rate of recast noticing in text chat through think-aloud protocols and stimulated recalls and found that learners are more likely to notice contingent recasts than non-contingent recasts. Further research on recasts in chat interaction is discussed below in Section 5.5.3 when we summarize studies that compare recasts with metalinguistic feedback.

Recasts in ICALL

The amount of research dedicated to recasts indicates that this particular type of feedback has probably drawn the most interest among all types of feedback. As we have seen above in Section 5.3.1 (page 89), it is also the most prevalent form of feedback in classroom interaction. In contrast to that, recasts have been implemented and tested in ICALL systems only very rarely. One example is the SPELL system described in Morton and Jack (2005) which offers assistance in spoken natural language interaction for learners of Japanese and Italian by providing recasts on grammatical errors (we have described this system in more detail above in Section 3.2.3). However, to our knowledge, this system has not been evaluated in terms of learning gains. Another example is Petersen (2010), who compared the effects of recasts in a text-written ICALL system with the effects of recasts in oral face-to-face learner-teacher interaction. He found a positive effect for English learner question development as well as morphosyntactic accuracy in both modes, but no difference between the two modes. Apart from these

two examples of recast implementations in ICALL systems, ICALL developers seem to be reluctant to develop applications that provide recasts. They are more likely to implement and examine more informative and explicit types of feedback, as will become evident in the following section, when we discuss the existing work on metalinguistic feedback.

5.5.2 Metalinguistic feedback

Metalinguistic feedback, as defined by Lyster and Ranta (1997), indicates that an error occurred and it can include hints about the nature of an error. Lyster and Ranta further distinguish three different types of metalinguistic feedback: comments, information, and questions. Metalinguistic *comments* indicate that the learner utterance is not well-formed, (e.g., "That is not correct.", "There is an error."), but do not provide any further details or explanations. Metalinguistic *information* gives more details about the source of the error, typically using linguistic terminology, e.g., "You should use the dative case here!". Metalinguistic *questions* indirectly provide hints to the source of the error, by asking the learner about linguistic properties of their attempt, e.g., "Is it feminine?" or "Which case should you use here?". The different types of metalinguistic feedback illustrate that it can differ widely in the amount of information it contains - from merely indicating that there is an error to a detailed explanation of the sources of the error. According to Lyster and Ranta's definition, metalinguistic feedback does not provide the correct form and we will adhere to this property in our further discussion. However, it should be noted that in some studies, metalinguistic feedback is operationalized as including the correct form. We will point this out in the discussion of the concerned cases.

Metalinguistic feedback is in several regards complementary to recasts. It is explicit while recasts are usually implicit. It interrupts the flow of meaning-based communication while recasts can blend into the communication. These properties makes metalinguistic feedback in general easier to notice for learners than recasts. Finally, in contrast to recasts, metalinguistic feedback does not provide the correct form, but prompts learners to generate it on their own, which has been argued to have a positive effect on learning (as we have discussed above in Section 5.3.2.

We will now summarize the results of research on metalinguistic feedback. In the area of human-human interaction, this type of feedback has not inspired nearly as great an amount of research as recasts have; in contrast to the area of ICALL applications, in which metalinguistic feedback has generated much more research. Since the present study is also concerned with feedback within a ICALL system, our summary focuses on existing ICALL research. After that, we will discuss studies that compare metalinguistic feedback and recasts, and, on a more general level, explicit and implicit forms of feedback.

Research about metalinguistic feedback in ICALL systems can be divided into two strands. In one strand, the goal is to examine how learners perceive and use the feedback. In the other strand, the goal is to examine the effect of the feedback on the development of linguistic knowledge. Van der Linden (1993) and Heift (2001a) follow

the first strand in investigating if learners do attend to metalinguistic feedback if they have the choice. Heift (2004) and Yang and Akahori (1999) extend this line by analyzing the reaction of learners towards metalinguistic feedback compared to other, less informative feedback. Nagata (1993, 1997) and Nagata and Swisher (1995) are examples of the second strand – they evaluate metalinguistic feedback by the learning gains it induces. We are now going to summarize these studies in more detail.

Do learners attend to metalinguistic feedback?

Van der Linden (1993) examined how learners of French used the feedback facility of a computer exercise program. The goal of the exercises was to manipulate sentences, more specifically, to replace the nouns in a given sentence with pronouns. Van der Linden examined learners' strategies by logging their interactions with the program and additionally by conducting think-aloud protocols and interviews with a subset of the participants. For the exercises, learners had the option to try as often as they want and consult detailed metalinguistic feedback. It turned out that only about half of the 23 participants consulted the feedback after an incorrect response and attempted to correct themselves. The other half made only one attempt at solving the exercise and then proceeded to the next question. Across all types of learners, it appeared that longer feedback – defined as more than three lines – was rarely read until the end. Some of the students who did consult the feedback seemed not to be able to use it – as evidenced by the fact that they repeated some of their incorrect responses. The conclusion of this study is that if learners are given the option to receive metalinguistic feedback, roughly half of them prefer to go without. Further, lengthy and complex feedback is less likely to be considered by learners than short and simple feedback.

Heift (2001a) built on Van der Linden's study and examined if learners do attend to metalinguistic feedback given in a ICALL system and how they use the feedback. Her system offers form-focused exercises for learners of German, among others, for instance, a sentence building task based on a set of noninflected word forms (see also the description of the system above in Section 3.2.2). When the learner's answer is incorrect, the system provides metalinguistic feedback about the error without providing the correct answer and learners have the choice to either try to correct their error or to take a look at the sample solution. In the study, 33 students from introductory German classes spent three one-hour sessions with the system and its sentence building exercises. The interaction with the system was logged. The exercises covered a broad range of grammatical structure which had all been introduced and practiced before in classroom activities. After the first failed attempt, on average, in 73% of the cases, learners considered the feedback and tried to correct the mistake, in 27% of the cases, learners requested the correct answer without trying again. This indicates that metalinguistic feedback is indeed used and appreciated by the majority of learners, who did not request the correct answer although is was accessible.

In another study, Heift continued to research the use of metalinguistic feedback in her ICALL system (Heift, 2004). This time, she compared the effect of three different variants of feedback – (a) simple metalinguistic feedback, (b) metalinguistic feedback plus highlighting and (c) repetition plus highlighting. Highlighting means that the erroneous part of the learner production is set in bold font. In condition (c), the learner

utterance was reproduced in the feedback area, the erroneous part was highlighted and a general comment about the category of the error was given, for example, grammar vs. spelling mistake. In addition to the sentence building exercise, the study also covered dictation, fill-in-the-blank, and translation tasks. Participants were 177 students enrolled in German courses in three Canadian universities, with levels ranging from beginner to intermediate. The students worked with the system for a period of 15 weeks for approximately 8 to 12 hours in total. As in the previous study (Heift, 2001a), Heift logged if learners tried to correct their error in response to the feedback or if they quit the task by either querying the system for the sample answer or skipping the exercise altogether. Each participant received a balanced amount of each feedback type. After receiving (a) - metalinguistic feedback, learners tried to correct their error in 86.9% of the cases. When they received (b) - metalinguistic feedback plus highlighting, they corrected in 87.4% of the cases. The difference is not significant. However, after (c) - repetition with highlighting only 81.7% of the answers were resubmitted. This is significantly less than the other two conditions, which suggests that informative, metalinguistic feedback is slightly more conducive to evoking learner self-repair than relatively uninformative highlighting.

In a similar study, Heift (2010b) investigated how the specificity of metalinguistic feedback affected if learners tried to correct their error. She compared two types of feedback, metalinguistic clues and metalinguistic explanations. Clues indicate the location of the error by highlighting the involved word and show whether the error is based in grammar or spelling. Explanations provide a metalinguistic explanation of the error and are thereby considerably more informative. The results of the study indicate that more informative explanations lead to significantly more learner repair.

In addition to examining the behavior of learners with regards to feedback, there have also been studies which queried the learners explicitly which feedback they preferred. When Heift (2004) asked learners for their subjective opinions about the different feedback types, 85.5% affirmed that they would prefer the most explicit feedback at all times. This is similar to the findings of Yang and Akahori (1999), who compared two different ICALL systems for Japanese that differed with regard to the flexibility in input they allow and the informativity of feedback. Learners had to work with both systems and were then asked about their experience. With a high majority, learners preferred detailed, metalinguistic feedback over simple feedback that merely displayed the correct answer, independently of the user input.

Effect of metalinguistic feedback on language skills

The research discussed so far examined how learners react to feedback if they have options and how they perceive it. However, this perspective cannot provide information about the effectiveness of feedback for improving language skills. Nagata and colleagues conducted a series of studies that target this question.

Nagata (1993) and Nagata and Swisher (1995) showed that metalinguistic feedback that points to the error and explains its nature using linguistic terminology is more efficient for improving the learner's accurate use of Japanese particles and passivization than metalinguistic feedback that only describes the error by listing which words in the answer were missing, incomplete or not expected to be used. Thirty-two

university students, enrolled in 2nd-year Japanese courses attended four treatment sessions, in which they first read a grammatical explanation and then completed exercises with a ICALL system. There were two variants of the system that differed regarding the feedback they provided (in Section 3.2.2 we already discussed this in more detail). For the exercises, learners were given a communicative context and a Japanese prompt - produced by an imaginary conversational partner. Their task was to respond to the prompt using a Japanese sentence. Input and output were in typewritten mode. Learners who had received the more informative type of metalinguistic feedback, achieved better posttest results for complex sentence-level structures - particles and passive constructions - compared to the learners who had received less informative feedback. The two groups did not differ, however, regarding their performance on word-level structures, i.e., vocabulary and conjugation. Note that both types of feedback that were compared are covered by Lyster and Ranta (1997)'s definition of metalinguistic feedback. The difference was the presence of additional grammatical information – which proved to be beneficial for the acquisition of more complex phenomena.

In a follow-up study, Nagata (1997) compared the effect of informative metalinguistic feedback with the effect of translation feedback, i.e., English translations of Japanese phrases with particles. The results showed that metalinguistic feedback is more efficient than translations for the acquisition of Japanese particles.

In summary, if learners have the choice to consider metalinguistic feedback for correcting their initially erroneous responses, the proportion of learners who consider it as opposed to the learners who neglect it varies between roughly 50% to 80%. Factors that influence the choice may relate to individual learning styles but also to the nature of the feedback – longish feedback is less likely to be considered (van der Linden, 1993). To our knowledge, other potential factors have not been investigated specifically, for instance, the nature or the linguistic structure may also influence the choice. If learners can choose between more informative metalinguistic feedback and less informative types of feedback they seem to prefer the former (Heift, 2004; Yang and Akahori, 1999). Learners who are not given the option to skip feedback seem to profit from more informative feedback than from less informative feedback (Nagata, 1993, 1997; Nagata and Swisher, 1995).

After having discussed recasts and metalinguistic feedback separately, we will close this chapter by presenting studies that have directly compared the two feedback types, since this is one goal of the present study as well.

5.5.3 Recasts versus metalinguistic feedback

We will now give an account of studies that specifically compare metalinguistic feedback with recasts. The majority of this research was conducted in oral face-to-face situations and only a smaller part was conducted with a written chat interface. So far, to our knowledge, there is no study that compares the two feedback types in the context of a ICALL application, where the learner receives the feedback from an artificial agent.

Rezaei and Derakhshan (2011), Sheen (2007), Sheen (2010a), Lyster (2004), and Ellis et al. (2006) all investigated feedback in oral group discussions, only Carroll and Swain (1993) investigated in the context of oral one-on-one interaction. Of the three studies that investigate in the context of written chat interaction, two were implemented as one-on-one interaction (Sauro, 2009; Razagifard and Rahimpour, 2010) and one as a group discussion (Loewen and Erlam, 2006). The overall trend that emerges from these studies is that metalinguistic feedback seems to be more beneficial than recasts. However, as the following more detailed description will reveal, the realization of the feedback in some of the studies differed to a certain extent from the definitions that we have given above, therefore, any claims about the effectiveness of a particular feedback type need to be well qualified in order to prevent improper, too general conclusions.

Feedback in oral face-to-face situations

Rezaei and Derakhshan (2011) compared the effect of recast and metalinguistic feedback for the acquisition of English conditionals and wish statements. Participants of the study were 60 male participants from three intact English classes in the Iran Language Institute, aged between 15 and 25. They were chosen based on a pretest that ensured that they had no measurable knowledge of the target structures. Classes were randomly assigned to one of three conditions: recasts, metalinguistic feedback, and control, who received no form-related feedback. After an introductory teaching phase, which introduced the target structure to all groups in the same way, the groups had to solve focused tasks (see Section 4.5.2, page 80) to practice the new knowledge. The different feedback was provided during the task-driven in-class interaction and it was addressed either to the whole class or individual students. In a posttest, both feedback groups outperformed the control group significantly and the group that had received metalinguistic feedback achieved significantly higher results than the recast group. Unfortunately, the authors do not give further details about the type of test they employed, nor the length of the feedback episodes.

Sheen (2007) compared recasts with metalinguistic feedback for the acquisition of English articles. Note that, in contrast to the definition given by Lyster and Ranta (1997), she realized metalinguistic feedback as including the correct form. Participants were 80 learners of English enrolled in an American Language Program of a community college in the United States. The students came from various first language backgrounds, were aged between 21 and 50 and had an intermediate level. The study covered two treatment sessions lasting between 30 to 40 minutes in two consecutive weeks, which were conducted within six intact classes. The feedback was provided in the context of a narrative task - students were to retell a story in front of the class. Progression was measured by a pretest before the treatment, a posttest after the treatment, and a delayed posttest five to six weeks after the last treatment. The testing instruments included a speeded dictation test, a writing test (four sequential pictures served as a stimulus to write a coherent story) and an error correction test. Participants who had received metalinguistic feedback outperformed the recast and control group. The recast group did not perform better than the control group in the immediate as well as the delayed posttest. Sheen hypothesizes that the reason for the apparent ineffectiveness of recasts might be due to the shortness of the instruction or to the lack of

salience of the target structure. Sheen (2010a) elaborated on these findings by adding two types of written feedback that correspond to the two oral types of feedback. The written feedback was given in response to narrative texts that had been composed by the learners. It either directly provided the correct form (similar to oral recasts) or gave a metalinguistic explanation plus the correct form. Again, oral recasts did not yield better results than the control condition. All other feedback had a significant effect on learner performance but the effects were not significantly different from each other.

Lyster (2004) compared the effect of recasts and prompts for the acquisition of the grammatical gender of French nouns. Prompts were a feedback category that included metalinguistic feedback, clarification requests, repetitions, and elicitations. They all have in common that they withhold the correct form and try to elicit a learner repair. Participants in this study were 179 students, aged 10 to 11, from eight different classes in an early French immersion program in Canada. The feedback was given in the context of form-focused instruction that included typographically enhanced texts based on the subject-matter curriculum and tasks that asked learners to derive orthographic and phonological regularities that govern grammatical gender in French. The instruction spanned over a period of 5 weeks, comprising 8 to 10 hours in the classroom. Students who received prompts outperformed students that received recasts or no feedback in assigning the correct gender. Learning progress was measured in two written and two oral tests, in which learners had to choose the correct gender, complete a text, name an object and describe pictures respectively. Since this study collapses metalinguistic feedback with other types of feedback that withhold the correct form and prompt for learner repair, one can only conclude that prompting is superior to providing the correct form. It is difficult to draw any conclusion about the effect of explicitness of feedback since Lyster does not reveal how the different types of prompting feedback were distributed.

Carroll and Swain (1993) compared the effect of four different types of feedback for the acquisition of the English dative alternation. Participants were 100 adult low to intermediate learners of English with Spanish L1 background who were enrolled in different courses in Toronto. There was one treatment session, which was preceded by a pretest and followed by an immediate posttest, a delayed posttest was administered one week after the treatment. The target structure was not elicited in a communicative task, but in decontextualized prompts, which asked learners individually to find alternative versions of the prompt, which was presented as text and audio. Learners were told what kind of feedback to expect when they were wrong and the learning goal was to distinguish between verbs that do alternate and verbs that do not. The first group was given a metalinguistic explanation of the dative alternation rules when they proposed a invalid alternation. The second group were just told that they were wrong, the third group was given a recast, and the forth group were asked if they were sure about their response when they made a mistake. A control group received no feedback. The progress as measured by grammatical judgment tasks indicated that all types of feedback resulted in learning gains compared to no feedback. Furthermore, the group that received the metalinguistic explanation of the rules outperformed all other groups, who did not receive an explanation. Since the exercise was not embed-

ded in a meaning-based context and the learners were told what kind of feedback to expect when they made a mistake, the results of this study cannot be directly used to draw conclusions about the effect of feedback in more communicative, task-oriented contexts, in which feedback is normally provided and which are closer to the actual situations of language use.

Ellis et al. (2006) compared the effect of recasts and metalinguistic feedback in face-to-face group discussions for the acquisition of English regular past tense verb forms -ed. Participants were 34 learners of English from three classes in a private language school in New Zealand. Feedback was provided during two half-hour communicative task sessions on two consecutive days. The tasks were designed to encourage the use of the past tense. They contained picture material that served as a prompt for telling a story. The stories were prepared in groups of three and then told within the whole class (of 10 to 12 students) with the instructor giving feedback. Learning gains were estimated with tests for explicit (untimed grammaticality judgment test, metalinguistic knowledge test) and implicit knowledge (oral imitation test). The results of the tests showed that explicit metalinguistic feedback was superior to recasts in promoting accuracy gains for English past tense verbs in both explicit and implicit knowledge.

After we have recapped the results of oral interaction studies, we will now discuss work that examined written interaction.

Feedback in written chat-based interaction

In a replication of the study described above (Ellis et al., 2006), Loewen and Erlam (2006) compared the effect of the two different feedback types in chat-written group discussions instead of face-to-face discussions. Again, learners of English (n=31) received recasts, metalinguistic information or no feedback in response to their errors with English regular past tense. In this study, learning gains were measured with a timed and an untimed grammaticality judgment test. None of the feedback groups showed significant learning gains as measured by these tests. The authors hypothesized that a reason may lie in the lack of immediacy between error and feedback, which makes it harder for the learner to notice a correction. The fact that written group discussions tend to be multi-threaded further increases the gap between error and feedback. Therefore, one-on-one chat conversation might be more effective.

Sauro (2009) examined the effect of corrective feedback in one-on-one chat-written interaction for the acquisition of zero articles for abstract, uncountable nouns. The participants in this study – 23 learners of English enrolled in a Swedish university – were randomly paired with native speakers of English and communicated via text-chat. The treatment included two sessions of 20 minutes on two separate days within one week. The goal of the chat session was to collaboratively write small essays about one of two topics – Swedish culture and global warming. In order to create contexts for the use of zero articles, the learners were given a list of 10 abstract nouns that they had to use in the composition task. Participants were randomly assigned to one of the experiment groups (recast or metalinguistic feedback) or the control group who did not receive any feedback. Learning gains were measured with an acceptability judgment test in a pretest-posttest-delayed posttest design. The test results showed that in direct comparison, neither feedback type was more effective than the other for immediate or

sustained gains in the target knowledge. However, the metalinguistic feedback group showed significantly higher gains between pretest and immediate posttest than the control group who had received no feedback. A problem with this study is that the occasions for feedback arose relatively rarely during the task – on average each session included only two to three feedback episodes.

Another study that examined feedback in one-on-one chat interaction was conducted by Razagifard and Rahimpour (2010). They compared the effect of recasts and metalinguistic feedback for the acquisition of past tense for 30 beginner level learners of English in Iran. The feedback was given in the context of a story completion and a picture description task. Feedback groups outperformed the control group (who received no feedback) in a grammatical judgment test and a metalinguistic knowledge test, but no significant difference between the two types of feedback was found. A fill-in-the-blank test yielded no difference between control and feedback groups. However, this study is questionable for its lack of pretest that should have ensured the comparability of the groups.

Conclusion

The studies that investigate the difference between recasts and metalinguistic feedback either find no difference or an advantage for metalinguistic feedback. It is interesting, that the studies that examined interaction with a type-written interface found no difference between the two feedback types, while the studies that examined oral face-to-face interaction found the metalinguistic feedback to have more effect.

It has to be noted that the measures that were used to test the progress on the target structures were, in general, more likely to assess explicit knowledge. Only Ellis et al. (2006); Sheen (2007) and Sheen (2010a) intentionally employ tests to cover implicit knowledge - an oral imitation test and a speeded dictation test. Loewen and Erlam (2006) included a timed grammaticality judgment test, which has later been argued to tap into implicit knowledge (Ellis, 2009b), but they did not discuss the implicit and explicit aspects of testing.

This summary should have made clear that the body of research that explores the effect of implicit and explicit types of feedback in the context of meaningful interaction is still small. In particular, within the field of human-computer interaction and computer assisted language learning, there is a lack of (a) studies that examine the use and effect of recasts and (b) studies that compare recasts with more explicit types of feedback. As we have detailed above, metalinguistic feedback has been explored within the scope of ICALL applications, but not in direct comparison to more implicit types of feedback.

The present study aims at filling this gap, by implementing recasts and metalinguistic feedback within a task-based meaningful interaction between a learner and a ICALL system. The next chapter will describe the methodology of our study in detail.

5.6 Summary

This chapter provided a closer look at feedback as a relevant factor for language acquisition and learning. It started off with a general discussion about the value of feedback by recounting the debates about (a) the theoretical necessity of corrective feedback and (b) the effectiveness and potential disadvantages of corrective feedback in practical language learning contexts. Perhaps in contrast to popular belief there exists no general agreement that feedback is beneficial and necessary for language learning. Objections are based on theoretical, sometimes ideological arguments but also occasionally on empirical evidence. Existing empirical work on feedback in general shows that its effectiveness is dependent on various contextual features, which keeps the debate alive and makes it hard to come to a general verdict on the effectiveness.

After that discussion, the chapter presented a classification of feedback by introducing the most common types of feedback in the language learning classroom and discussing the parameters that distinguish these types. These parameters comprise the explicitness, whether or not the learner is prompted for a modification and the information content of the feedback. The latter can be further divided along whether or not the correction is provided, the location of the error, and the nature of the error is provided.

The chapter then discussed feedback in the context of ICALL applications. Since learners are particularly dependent on the reliability and appropriateness of feedback, special care has to be taken to account for the risk of inappropriate feedback. We further illustrated the relationship between the information content of a particular feedback type and the types and amounts of information that an ICALL system has to model. Reliability and informativeness of feedback both are dependent on higher costs for development, therefore the cost and benefit of feedback has to be carefully balanced.

In the last part of this chapter we presented a detailed review about existing work regarding two particular types of feedback: recasts and metalinguistic feedback. We summarized evidence about the effectiveness of both feedback types individually and in direct comparison. The reviewed studies concern both oral classroom-based feedback and ICALL feedback. In direct comparison, recasts and metalinguistic feedback often yield similar learning gains, sometimes, metalinguistic feedback is superior. However, our presentation also made clear that there is a shortage of research that examines the effect of recasts on its own and in comparison with more explicit types of feedback in ICALL contexts. It is this gap that our study is targeting.

After we have now finished the series of chapters that provided the theoretical background for our study, presented related work, and motivated the exploration of particular issues, we will use the next chapter to illustrate the approach we used to pursue the objectives we started out with.

6 The Approach

6.1 Introduction

The goal of this thesis is to examine how the current state of the art in natural language processing (NLP) and computational linguistics (CL) can be employed to support foreign language learning. Thus, the thesis is situated in the discipline of computer assisted language learning (CALL) and draws on knowledge from the areas of second language acquisition (SLA) and foreign language learning (FLL) on the one hand, and NLP and CL on the other hand. While CALL comprises a wide range of approaches and technologies, we focus here on the subset of those that are usually called *"intelligent"* (ICALL), or more specifically, those that employ a certain amount of linguistic knowledge. In particular, we focus on approaches that provide interaction in the form of a *dialog* and that give *feedback* to the learner about the correctness and/or appropriateness of their productions.

We concentrate on dialog, and feedback within the dialog, as opposed to other possible ICALL applications as, for instance, vocabulary training or enhancement of authentic language material (see Section 2.1) because dialog is a distinguishing feature of human-human interaction and hard to provide by traditional non-interactive media. Such media can only provide examples for dialog interaction as texts, audio, or video snippets and the engagement of the learner is reduced to merely perceiving or consuming the material with no chance to actively participate. Similarly, feedback in traditional static material is usually constrained to the provision of correct solutions for exercises. Since real-time dialog and feedback are usually considered to be the exclusive domain of human tutors, it is even more interesting to provide it through an ICALL application that seeks to emulate human-like skills.

This chapter will describe our approach to pursuing the general research objective of this thesis by explaining our selection of methods, design, and parameters. The choice of the particular ways in which we realize dialogic interaction for language

learning follows mainly from the currently available state of the art on the one hand and the specific SLA questions that we want to examine on the other hand. In the following section (6.2), we will recap the current state of the art and discuss how it provides and constrains the space of possible implementations of dialog systems. We will then relate the implementational perspective to parameters that are relevant from the pedagogical perspective and introduce the focus we take in this thesis in 6.3. Section 6.4 will discuss alternative relevant parameters. Section 6.5 will describe the context of this thesis within the relevant disciplines in more detail and discuss the focus and approach under that perspective. Finally, Section 6.6 will introduce the research design and methodologies.

6.2 Implementing ICALL dialog and feedback

In this section we will discuss the existing constraints for implementing ICALL dialog systems and how they affect the conditions of our study.

At their core, dialog systems for supporting language learning have two goals. One is to provide opportunities for communication, the other is to provide feedback on the learner input. We have argued above in Section 2.2.2 that the complete and reliable analysis and interpretation of unconstrained learner input is beyond the current state of the art (Gamper and Knapp, 2002; Feigenbaum, 2003; Amaral and Meurers, 2011). Consequently, ICALL system developers need to find a compromise between the scope of system on the one hand and the depth and precision of linguistic analysis on the other hand. This trade-off is reflected in the classification of the systems that we presented in Section 3.2 – some focus on formal correctness and grammatical knowledge, while others focus on meaning and communicative interaction. But also for some of the individual systems, this trade-off is manifested in different features or variants that put more emphasis on the one or the other compared to other features or variants.

For the first group, the input of the learner tends to be considerably constrained but the system provides detailed feedback based on a detailed linguistic analysis. Systems in the second group tend to provide more freedom of input but only little or superficial feedback due to limits in the analysis of the input. As an example for the first group recall e-Tutor (page 43), in which precise error feedback can only be provided for exercise types that focus on grammatical forms and constrain the input to filling gaps, building sentences based on prompts, or translating. As examples for the second group, consider MILT and TLTS (pages 48 ff.) which both have a version that provides relatively free input but only attempts a shallow input analysis and gives no feedback on formal correctness.

Even for SPELL and Te Kaitito, (pages 50 ff.), the two systems that attempt to combine both goals by allowing free learner input in a dialog and providing feedback regarding the correctness of forms at the same time, the trade-off is still apparent in the implicit constraints of the topics of the dialog. Te Kaitito, for instance, constrains the domain, vocabulary, and inventory of grammatical forms to a narrow beginner level.

The particular choice that system developers make in view of this trade-off reflects their priorities regarding the pedagogic approach. Systems that allow free input tend

to follow a meaning-focused approach and value communicative competence and fluency. Systems that attempt a detailed analysis of the learner input and give detailed feedback, implement a more form-focused approach with an emphasis on accuracy. However, as we have discussed in our review, current ICALL systems are rarely evaluated in terms of learning gains and as a consequence, cannot be used directly to evaluate the value of a particular pedagogical approach.

On a more general note, it is also interesting from an engineering perspective to estimate the benefits of an application in order to balance them with the costs of development. As with any kind of project, developers have to consider how to achieve the most benefit with the given resources, or, inversely, starting from the target specification, how much resources and effort need to be spent. Most often, resources are limited, which means that goals have to be constrained. In order to make a good choice and find the optimal balance, the benefits need to be estimated. Benefits relate to performance and there are different measures for performance that can be employed depending on the particular application.

While the performance of many NLP/CL algorithms and models can be measured with relatively clear metrics related to precision and recall, it gets more messy when these models are part of a wider application and the notion of performance starts to extend to usability and user experience issues. A good example for that are the efforts related to finding performance measures for dialog systems. Walker et al. (1997), for instance, propose a metric that combines user satisfaction, task success, and dialog cost. For evaluating ICALL applications, an obvious criterion should be the learning gains they help to achieve. Such measures can then be geared to the more specific skills an application was built to train.

For this study, we attempt to evaluate the effectiveness (in terms of learning gains) of different systems with a view on (a) the trade-off between scope and precision and (b) the general level of complexity and sophistication. While the current state of the art constrains the potential field of instances that we can implement and evaluate, our choice of particular instances is also strongly based on the consideration of the SLA issues we want to examine (see Section 6.3 below). Furthermore, the details of our implementation and the experiment we conduct arise from considerations for the study design (Section 6.6), practical and theoretical considerations regarding the content matter of the instruction (Section 7.1), and the availability of supporting tools and resources (Section 8.1).

Scope and precision of an ICALL dialog system are manifested in different properties of the system. For the sake of this study, we consider the freedom of input and the nature of the feedback as parameters that define particular positions in the trade-off space. We will discuss alternative parameters and variants further below (Section 6.4).

In the discussion of the role of conversational interaction (Section 4.5.1) we argued that feedback is crucial for language learning. As we have discussed in Section 5.3.2, one important criterion to characterize feedback is the information that it contains. The informational value and precision of corrective feedback is crucially dependent on the analysis and interpretation of the input. Related to the feedback is the range of expected learner utterances that the system can interpret. We have discussed above (Section 2.2.3) that constraining the input is one way to deal with the current limits

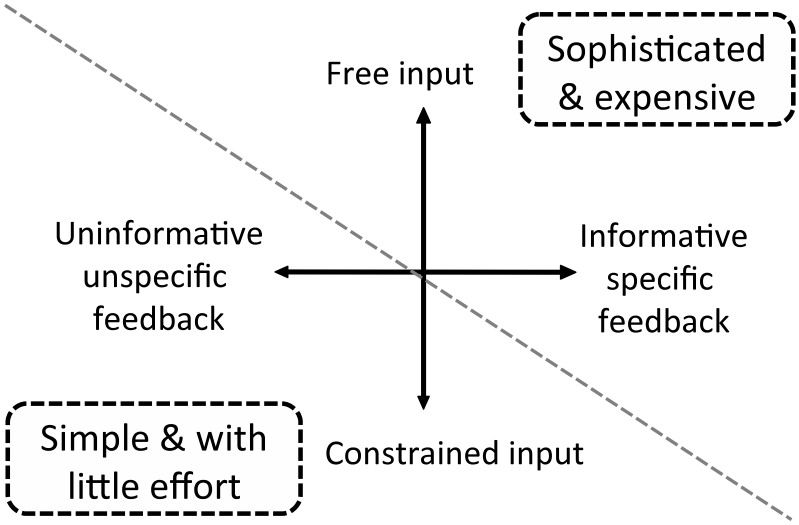


Figure 6.1 – Implementational aspects: Dimensions for sophistication and computational effort of CALL applications

of the processing capabilities of ICALL systems. The freedom of input that a system provides is directly related to the flexibility, naturalness and the similarity to a human teacher. However, it is not clear if such flexibility necessarily increases the utility as a learning tool and has an impact on the learning gains.

Figure 6.1 (repeated from Section 1.1) illustrates the relations between the range of these two parameters and the cost of development. The expenses to develop applications which allow relatively free input and provide relatively informative and specific feedback is higher than the expenses to develop their counterparts with relatively constrained input and uninformative and unspecific feedback. The overall effort arises from the combination of the two parameters and many systems value one over the other. The top right area of the diagram symbolizes systems at or beyond the border of the current state of the art. In the following two sections we illustrate the space of parameters in more detail by drawing on our previous discussion about constraining input and parameters of feedback.

6.2.1 Informativity of feedback

In general, the more informative a certain type of corrective feedback is, the more knowledge needs to be modeled within a system that can provide such feedback. As we have discussed above in 5.3.2 and Section 5.4 the information content of corrective feedback can be characterized in terms of whether or not the feedback contains (a) the correct form, (b) the location of the error, and (c) an explanation of the linguistic nature of the error. The overall informativity of different types of feedback ranges from containing none of these items to all of them. Table 6.1 summarizes our preceding delineation and sorts different types of feedback in terms of their information content. Note that some of the feedback types identified by Lyster and Ranta (1997) come in variants that differ with regard to their informativity. Where applicable, we provide examples to distinguish these variants. In addition to the feedback that occurs in lan-

	Information			
	Correction	Location	Explanation	Feedback type
1	□	□	□	▸ Clarification Request (*Pardon?; I don't understand.*) ▸ Elicitation (*How do you say that?*) ▸ Repetition (of entire erroneous utterance) ▸ Binary feedback for constrained drills (*Right! Wrong!*)
2	□	■	□	▸ Clarification Request (*What did you mean with X?*) ▸ Elicitation (of a particular part) ▸ Repetition (of erroneous part)
3	■	□*	□	▸ Clarification Request (*Did you mean X?*) ▸ Recast ▸ Explicit correction (of the entire utterance) **Learner can infer location of error by comparing own utterance with system feedback, but system does not need to know location.*
4	■	■	□	▸ Recast (embedded in new content) ▸ Explicit correction (of only the erroneous part)
5	■**	■	■	▸ Metalinguistic feedback. ***The correct form may or may not be provided to the learner, but system needs to know it.*

Table 6.1 – Information content of different types of feedback. Feedback can contain the correct form, the location of the error, and a metalinguistic explanation. Filled boxes (■) indicate that the information is present, empty boxes (□) indicate that the information is not present.

guage learning dialog situations, we add to the table another type of feedback that is only relevant for constrained language exercises: feedback that informs the learners whether their response was correct or not. This is relevant for the further setup of our study which we will discuss below.

When we define informativity of feedback, we start from the information that is encoded in the feedback. This information needs to be distinguished from the information that the learner perceives, and from the information that the ICALL system must model. The information about the location of an error may only be implicit in the feedback and dependent on the learner to discover it by comparing the feedback with their original utterance. In these cases (row 3 in Table 6.1) the system can provide the feedback without having explicit information about the location either. Inversely, in some realizations of metalinguistic feedback, the correct form is not provided by the system, which, however, needs a model of the error and the correct form to be able to provide an explanation (see row 5 of Table 6.1).

6.2.2 Freedom of input

It is evident that there is a relation between the level of complexity of a system and the breadth of learner input it allows and handles in an appropriate manner. The more freedom and flexibility a learner has to form utterances, the more sophisticated the system needs to be in order to react appropriately to this unrestricted learner input. In Section 2.2.3 we discussed constraining input as a strategy to deal with the limitations of available resources for language processing and described a range of examples.

We now generalize these examples and introduce a broad classification of ways to constrain input. Table 6.2 enumerates possible constraints for ICALL applications. At the highest end of the scale is a system that allows for completely unconstrained input, similar to a human conversational partner (1). While it is relatively easy to build systems that allow unconstrained input and, at least in the beginning of a conversation, may appear to reply in a sensible manner (see the chat bots described above in Section 3.2.3), the lack of a linguistically informed backbone makes these system unsuitable to give much useful feedback.

Going down the scale, constraints can be imposed through the task scenario which limits the contents and vocabulary while the learner is still free to choose the linguistic means for achieving the task objective (2). Further down, constraints can be set through providing task materials, like list of words or list of pictures that are to be used (3). More constraints can be implemented through more restricted task types, as for instance, translation or dictation tasks, which leave little freedom for creativity for the learner (4). At the end of the scale are activities that constrain the input of the learner such that they can merely choose a word or an suffix to produce as a response to a prompt in a form-focused drill or of a set of multiple choice responses (5).

At a higher level, constraints can be classified as either limiting linguistic forms (syntactic or morphological structures) or meaning (vocabulary, content). The structure-based constraints often entail content-restrictions, and are therefore, in general, more restrictive. Exceptions to this tendency may be tasks in which a structure is practiced with vocabulary freely chosen by the learners.

	Characterization of freedom and constraints
1	input completely unconstrained, every topic possible
	Constraints on content, vocabulary, meaning
2	implict constraint through task scenario, topic is constrained, but linguistic means are free
3	further constraints through task material in form of list of words or list of pictures, otherwise free input
	Constraints on linguistic forms and structures
4	constraints through task type and prompts, e.g., translation, dictation, learner production not free
5	input limited to responses to drill activities, gap filling, ordering words, or multiple choice

Table 6.2 – Freedom of input and constraints, a rough characterization

This is a coarse-grained classification of constraints. A more detailed characterization could be achieved by measuring the extent of possible learner input more rigorously. However, such detail is beyond the scope of this study. Even this coarse classification of ways to constrain input spans a considerable space of options to explore and examine.

In theory it may be desirable to conduct a fine-grained examination of instances in this two-dimensional space and to evaluate their impact on the learning gains achievable by interacting in a dialog along these parameters. However, since the capture of learning gains is severely limited and cannot be automated easily since it requires human subjects, there is a need to limit the range of instances and focus on only a few.

Our selection of instances is informed by relevant issues in the field of SLA. In the following section, we will first recapitulate the relevant pedagogic issues and identify the corresponding parameters. By relating these to the implementational parameters, we will isolate the instances within that multi-dimensional space that we chose to examine in more detail. Thus, in this thesis, we not only relate to relevant issues from SLA, but also contribute to it by conducting a study that generates new knowledge with a focus on instruction and learning in a human-computer context.

6.3 Relating to the pedagogic perspective

In the previous chapters we have presented two crucial issues in the field of SLA. On the one hand, there is the issue of how much emphasis to put on either form or meaning in instruction (Section 4.2). On the other hand, there is the dichotomy of implicit versus explicit forms of learning, knowledge, and instruction (Section 4.3).

Figure 6.2 illustrates this two-dimensional space of parameters. The two dichotomies are interrelated in that meaning-oriented approaches tend to impart linguistic knowledge more implicitly, while form-oriented approaches are often more explicit. How-

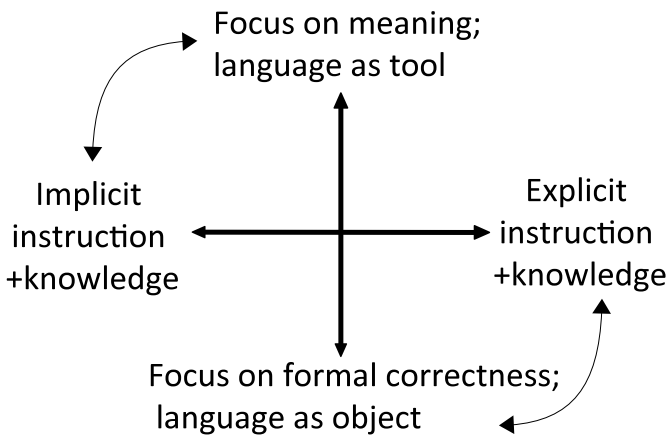


Figure 6.2 – Pedagogical aspects: Parameters in language instruction and learning

ever, as we have discussed in the previous chapters, this is not an absolute correlation and there are many approaches that fall in-between the extremes. Similar to the space of implementational parameters, the space of pedagogic parameters is wide and has many possible instances. Again, we need to make a selection of a small set of instances, in order to be able to collect meaningful learner data, following the conventions according to which SLA studies are conducted.

6.3.1 Explicit and implicit feedback

Section 4.3.2 discussed explicit and implicit forms of instruction. In implicit types of instruction, the instructor tries to *attract* the learner's attention to the form, for instance by making the form more salient, but never directly discusses it. In explicit instruction, the instructor *directs* the learner's attention to the form by discussing the form and putting it into focus during the lesson.

One of the most obvious areas in which explicitness can be varied in the setting of dialog interaction is the type of feedback. Although other factors can have an influence on the explicit/implicit dichotomy – for instance properties of the targeted forms or features of the meaning-providing tasks – it is not feasible in our context to vary them in a controlled manner. In Section 5.3 we introduced the different types of feedback available to the instructor: explicit correction, recast, clarification request, metalinguistic feedback, elicitation, repetition, translation. We argued in Section 5.5 for examining *recasts* and *metalinguistic feedback* because they are prototypical representers of implicit and explicit instruction respectively and because they are also better suited to be realized in a type-written ICALL dialog system than other feedback types.

Recast feedback is the least obtrusive and most natural way to provide FOCUS-ON-FORM, but as we have discussed earlier in Section 5.5.1, the very implicit and unobtrusive nature of recasts puts them at risk for going unnoticed. Metalinguistic feedback, on the other hand, is more explicit and more obtrusive but although it does interrupt the task-level conversation, the interruption is intended to be short. The feedback does not include a general elaborate explanations of the form, instead it only gives

brief hints pointing to the nature of the error.

We picked these two types of feedback as two instances from the scale of pedagogic parameter explicitness. From the implementational perspective, these two instances represent different degrees of informativity of feedback. Recasts require the correct form and, to some extent, knowledge about the location of the error. Metalinguistic feedback, on the other hand, requires knowledge about the correct form, knowledge about the location of the error, and metalinguistic knowledge to explain the error and prompt for a correction.

6.3.2 Meaning, form, and freedom of input

In order to explore the two other scales of parameters – meaning-form from the pedagogic perspective and freedom of input from the implementational perspective – we add another instance of instruction. In this variant, the learner's input is extremely constrained and at the same time very focused on formal aspects of language and very little on meaning. The input is constrained by a prompt that requires the learner to fill a gap in a prefabricated sentence or to bring a given set of words into the correct order to make a sentence. The feedback is binary and indicates whether the learner's input/response was correct. Opposed to that, in the other two conditions that compare recast and metalinguistic feedback, the input of the learner within a task-oriented dialog is relatively free. The input is implicitly constrained by the nature of the task and the prompt material provided. This means that the learners can produce what they want, but are expected to keep their contributions relevant and appropriate for a real-world task. The system gives feedback in response to errors regarding certain linguistic forms.

Figure 6.3 shows the three instances within the two-dimensional space of implementational parameters. Even though we have argued that feedback informativity and constraining input cannot be characterized as simple linear scales, we will use such a simplification for the sake of illustration. The x-axis indicates the informativity of feedback and the y-axis shows the constraints on the input, based on the orders introduced in Table 6.1 and 6.2.

The three instances can be ordered according to the effort that is required to implement them. The constrained input with binary feedback at the lower left is the least expensive to implement. The free input with recast feedback comes second and the free input with metalinguistic feedback is the most expensive to develop. Note that the current state of the art would not allow the creation of instances that combine completely free input with the most informative type of feedback, that is, instances placed in the top right site of the diagram.

In relation to the pedagogic parameters, we can bring the three instances in the following orders, which are also encoded by the shades of the star-shaped icons that encode the position of the instances in the parameter space in Figure 6.3. With regard to the meaning-form dichotomy, constrained input is the least meaning-focused, followed by the free input with metalinguistic feedback, while the free-recast instance is the most meaning-focused. Sorting the instances in terms of explicitness results in the same order: constrained input is the least explicit, followed by the free-metalinguistic feedback instance, while the free-recast instance is the most explicit. We will describe

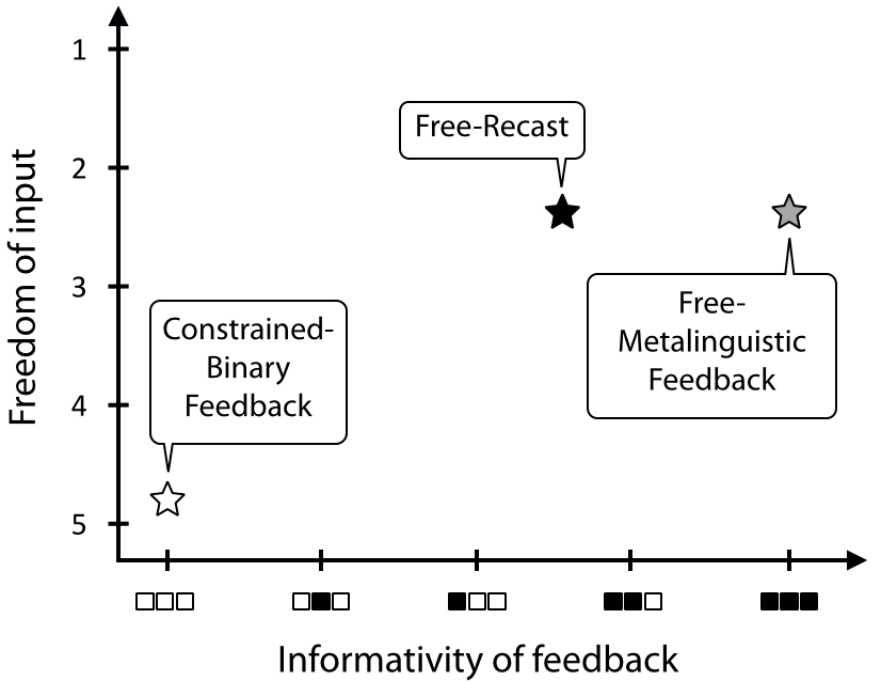


Figure 6.3 – The three instances that we compare, presented in relation to implementational parameters. The relation to the pedagogic parameters is encoded by the gray scale of the icon: the darker an icon, the more focus on meaning it implements and the more implicit it is.

more details of the three instances and discuss additional choices and constraints in Section 6.6.

6.3.3 Relations between pedagogic and implementational parameters

After we have introduced the three instances in the four-dimensional combined space of implementational and pedagogic parameters we conclude with some remarks about the relationship between the two parameter spaces. This characterization is more general and independent from the instances. It is evident that there is no direct and clear correlation between the SLA-related pedagogic and implementational parameter space. They relate to each other in different ways, as discussed in the following paragraphs.

Freedom of input First of all, free input is usually associated with more implicit and more meaning-oriented approaches to instruction, while constrained input tends to serve more explicit and form-focused approaches.

Feedback and explicitness The relationship between informativity of feedback and degree of explicitness is somewhat less clear. From one perspective, it seems that explicitness correlates with a higher degree of informativity while implicitness is related to a lower degree of informativity, because explicit means that information is clearly expressed and observable, and thus accessible, whereas implicitness entails that information is not readily apparent but only potentially inferable and thus perhaps less

accessible and hidden by subtlety, vagueness and ambiguity (see Section 4.3 on page 64). However, if we disregard whether a learner is actually able to infer the information contained in implicit feedback and if we consider just the information content that is potentially inferable under the most favorable circumstances, there is no clear relationship.

Recall that feedback can contain three different types of information, the correct form, the location of the error, and an explanation. Feedback that does not contain any of this information can still be very explicit; consider for instance, the feedback expressed in the utterance: "This is incorrect". At the same time, a frown or a slight hesitation is an example of more implicit feedback with an equal dearth of information. On the other hand, a linguistic explanation of an error is difficult to pass on in an implicit way. However, as we have seen with the example of recasts, the location of an error together with the correct form can be provided in an implicit manner.

Feedback and Form/Meaning Feedback can concern both formal aspects of language as well as meaning or the combination of them. In meaning-oriented conversation, feedback tends to occur only if the intention of the speaker could not be inferred. Other errors regarding the form that are not crucial to getting across the meaning are likely to be ignored. Only in the context of learning, where learner and/or teacher are interested in formal correctness will feedback take into consideration the forms. Thus, the informativity of feedback is not directly correlated with a focus on meaning or form. However, to the extent that emphasis on formal aspects tends to occur in situations where someone intends to learn the language and someone else assumes the role of a teacher, the teacher might assume that more informative feedback is more efficient or even expected. In a meaning-oriented context on the other hand, additional information in feedback may be omitted for reasons of efficiency.

6.4 Alternative parameters

While freedom of input and informativity of feedback are naturally relevant parameters in the domain of modeling dialog that supports language learning, they are by no means the only ones. There are other parameters that are related to the complexity and sophistication of an application and to pedagogic issues. We will briefly discuss them in the following and argue why we did not consider them for the current study.

The first group of parameters are primarily relevant from the pedagogic perspective, while the second group of parameters are more related to features of the dialog system and its complexity.

6.4.1 Parameters related to learning

Speech versus type written

While the default mode of human dialog appears to be speech, type-written real-time synchronous conversations enabled by internet relay chat and instant message services have their place as an alternative to voice-based direct or remote forms of communication. It seems obvious that speech-based dialog systems require significantly higher

effort for interpretation and production. The direct comparison of the learning effects provided by speech versus type-written dialog systems would have been an interesting research question that also relates to issues currently researched in the field of computer-mediated communication (see Section 2.4). However, in the scope of this thesis, we forgo speech because the required effort and the additional difficulties and challenges related to automatic speech recognition would have shifted the focus too much away from the actual goal of this study.

Production and comprehension

A meaningful distinction related to language skills is production versus comprehension. Dialog comprises both. However, our focus here is on production and we consider comprehension or perception only at a coarse and superficial level by asking learners to indicate what they noticed. Apart from that, we start from the assumption that learners understand all or most of the utterances of the system and in case they do not, that reformulations by the system will help. Clearly, this assumption is at best simplistic and at worst inaccurate, but to all intents and purposes, the adopted dialog models presuppose comprehension and only in some cases offer reformulations where the learner's reaction suggests a non- or misunderstanding. Apart from that, different levels or even lack of learner comprehension is not explicitly modeled.

Input enhancement

Another potential parameter that is relevant from the pedagogic perspective is input enhancement, as a way of drawing attention to certain features of the input, as we have discussed in Section 4.2.4. However, from the perspective of effort to implement and also from the perspective of interaction, it appears less relevant. We will briefly return to that issue further below in Section 6.6.3, when discussing the concrete experimental parameters.

Nature of linguistic knowledge

Another choice in the context of pedagogic dialog is the nature of linguistic knowledge that is to be imparted. Participating in a dialog has in principle the potential to provide knowledge on all levels of language, starting from pronunciation and/or orthography, over morphological and syntactical knowledge, semantic, pragmatics, and finally, across all of these, vocabulary. In this thesis, we focus on certain grammatical structures, that are both expressed morphologically and syntactically, as well as the vocabulary and phrases that are useful for certain practical tasks. These structures and words/phrases, of course, involve semantics, in the sense that they encode certain meanings, but the details of that are not in focus. Pronunciation is disregarded as a consequence of opting for type-written interaction. Orthography is only relevant in the sense that the system provides examples of correct orthography, however, it tolerates misspellings up to a certain point and gives no explicit feedback regarding those mistakes. Pragmatics, understood as the interpretation of meaning with regard to the non-linguistic context of the communication only matters implicitly in the way that

the dialog relates to non-linguistic material that provides constraints and stimuli for the task.

6.4.2 Parameters related to dialog

After the discussion of alternative parameters related to pedagogy, this section will now discuss briefly another set of possible extensions to a regular ICALL dialog system. These all advance the complexity of development and potentially increase the user experience and learning gains but are only indirectly related to pedagogic issues. These enhancements did appear before in the presentation of existing systems in Section 3.2.

Multilingual dialog

Multilingual dialog is based on the model of a *foreign* language learning context, in which, as opposed to a *second* language learning context, the existing (often native) language of the learners is frequently used for providing explanations (see also Section 5.2.3 and below). A multilingual dialog system can be designed to provide explanations or other feedback in the first language of the learner or a language in which the learner is more proficient than the target foreign language. An example for that is the Te Kaitito system that we presented above in Section 3.2.3 (Knott and Vlugter, 2008; Vlugter et al., 2009). Related research questions could be framed around the issue of whether resorting to another language has an impact on the efficiency and sustainability of learning.

Multiparty dialog

Another extension would be to design a system able to model multi-party dialogs, including more than one learner and/or more than one artificial agent that take part in the dialog. This could be used for imparting linguistic knowledge related to personal pronouns as in Te Kaitito (Knott and Vlugter, 2008; Vlugter et al., 2009) or in general to provide a richer, more complex setting for dialog, which might also involve reasoning about the active and passive participants in the conversation (Traum and Rickel, 2002). While such extensions are without doubt appealing, the direct advantages for language learning, except for the case of the use of personal pronouns, are less clear.

Contextual information

The last two extensions regard the context of the dialog. Starting from the assumption that the dialog is based on a task or has some goal related to the external world, the degree of complexity can be influenced by the nature of the contextual representation. Tasks can be derived from actual real world contexts or they can have somewhat simplified, abstracted prompts. As an example, consider the task of ordering a meal in a restaurant. The simple version would provide a short abbreviated and fixed menu as a task prompt. A more natural and authentic context might be provided by an actual menu sourced from the real world. An even more flexible context could be provided by using an arbitrary menu that might be retrieved online randomly every time the

dialog is started. The last version would require a component able to parse the menu (or another input stimulus for another task) to make it available for the system, but the increased complexity is rewarded with more flexibility and authenticity. In this case, the higher level of complexity is only indirectly related to NLP/CL methods.

Virtual reality

Authenticity can increase the immersive aspect of a dialog. Arguably an even more immersive and engaging context is an animated three dimensional world, in which the context and possibly also the artificial agents that serve as the dialog partner are graphically represented. The pinnacle of that idea would be a virtual reality environment where the learner interacts with animated, more or less realistic agents (Traum and Rickel, 2002; Harless et al., 2003; Johnson and Valente, 2009). Such a context requires a huge amount of additional modeling, including non-verbal modalities like gestures or gaze. Such complexity is beyond the scope of this thesis, but it would nevertheless provide an interesting premise for examining the learning effects compared to less advanced dialog contexts. It would be of particular interest to attempt to isolate the confounding effect of higher engagement and enjoyment from using the more complex context.

6.5 The context of this thesis

Disciplines and research areas

The work described in this thesis is situated at the intersection of three different fields of research: natural language processing and computational linguistics, second language acquisition and foreign language learning, and intelligent computer-assisted language learning. In this section we will describe the relationships between these three areas and how this study connects them.

Natural language processing and computational linguistics examine how natural human languages can be processed by computers with the goal to model linguistic knowledge in machines and thus make them capable to produce and understand human languages. It thus builds "artifacts that usefully process and produce language, either in bulk or in a dialog setting" (Schubert, 2014). **Second language acquisition and foreign language learning** examine how humans learn second or further languages and which conditions support the acquisition process. They are related to the research area of **first language acquisition** which researches how humans acquire their native language, but there is convincing evidence that acquisition processes for languages learned later in live differ in important ways from the processes for infants (Section 4.2.2, Schachter (1996)). The discipline of **ICALL** attempts to develop computer applications that support language learning, making use of some form of advanced or "intelligent" knowledge. Between these three disciplines, we can find the following relations.

SLA/FLL and ICALL have a mutual relationship. Findings from SLA/FLL can inform and support the design and development of ICALL applications. In turn, ICALL

applications can be used to contribute knowledge to SLA/FLL. This can be done such that ICALL provides tools and environments to examine and test SLA theories. Alternatively, ICALL applications that were developed and employed with more practical goals in mind and not with an explicit intent to use them for research can be tested later and evaluated and inform SLA/FLL research.

NLP/CL and ICALL serve each other in the following ways. NLP/CL knowledge has a crucial role for ICALL in that it informs and supports the development and implementation of ICALL applications. In turn, needs and requirements that arise in ICALL can drive and motivate the work on NLP/CL theory. For instance, the need to treat erroneous learner language motivated work on modeling and diagnosing learner errors.

NLP/CL and SLA/FLL have a twofold relationship. On the one hand, they are connected indirectly through ICALL. In that indirect way, NLP/CL contributes to SLA/FLL by providing knowledge to build ICALL tools that test and optimize conditions for SLA/FLL. Vice versa, SLA/FLL research suggests challenges for NLP/CL to tackle in order to develop ICALL tools.

In contrast, beyond that indirect connection, there are also more direct relations that arise without a role for ICALL. NLP/CL technology can be used to analyze learner language data in an automated way and thus inform about regularities of language acquisition that are infeasible to gather through manual inspection. Advances in the field of *computational language learning*, also known as *grammar induction* can also provide insight about human (second) language acquisition, to the extent that the computational models are adequate in modeling human acquisition (Clark and Lappin, 2011). Knowledge about SLA/FLL processes on the other hand could be used in NLP/CL for developing applications that emulate the performance at certain learner stages.

Related to this tripartite relationship, Chapelle (2001) conceptualizes the field of **computer-assisted SLA research (CASLR)** and identifies its two main objectives. One is to assess the effect of instruction, the other is to discover and reason about learner's knowledge and learning strategies with the help of computers. We consider the contributions that ICALL and NLP can make to assist SLA as part of CASLR.

Approach and contributions

In this thesis, we explore how foreign language learning can be supported through a task-based interactive dialog system that relies on models and processes provided by NLP/CL. The approach of this thesis is to explore the potential space of ICALL dialog implementations and harness them to examine relevant SLA questions. Thus, we contribute to all three involved disciplines. Within the scope of ICALL, we develop a dialog system. For SLA/FLL we contribute new knowledge by transferring findings that were produced in the context of human-human interaction to human-computer interaction and examine to what extent they hold in the new context. This work contributes to the disciplines of NLP and CL by creating a framework for exploring how basic state-of-the-art technology can be employed to examine and compare different

parameters for language instruction. Thus, it attempts to show that dialog-based instruction can induce learning gains and which parameters are more effective. In this thesis, we create a space of potential usage of NLP/CL applied to the goal of foreign language instruction. In this manner, we generate practical experience for a specific application context of NLP/CL.

Using knowledge and methods from all three areas, this study presents an example of how to examine SLA questions in the context of human-computer interaction. It thus also contributes to CASRL. Beyond that, our approach can be used as a basis for developing more comprehensive frameworks to examine further SLA issues.

As described in the previous chapter, the SLA issues we want to explore have so far been targeted mostly in traditional human-only contexts. Most of them were not conducted from the perspective of ICALL and the computer as a conversational partner. At the same time, although the number of ICALL-systems that engage the learner in conversational interaction has been growing in recent years, ICALL-developers usually do not take an SLA-perspective when evaluating their systems. One notable exception is the work described by Petersen (2010), who compares the effect of recasts in oral human-human interaction and ICALL type-written interaction. Our study tries to contribute to this as yet small body of research that integrates the ICALL and SLA perspectives.

This thesis did not set out to advance the state of the art for any of the specific technological conditions for implementing dialog for language learning. As such, the goal was not to find a more reliable or more comprehensive approach to diagnose errors or a more flexible and powerful dialog management. Instead, the contribution of this thesis is set up a framework in which existing technology is employed to explore dialog- and feedback-based ICALL guided by SLA research issues. Thus, we gather experience and create new knowledge about how NLP/CL can be employed both for practical applications and at the same time as research tools.

Starting with the general goal of exploring language learning through dialog systems, there are different possible approaches. For one, it is conceivable that we could answer this question simply by combining all existing research in a meta-study. However, this relies on a sufficient body of existing research. As we have seen, only a few of these systems were subjected to a detailed study about the learning they afford. In theory, it would have been an option to employ existing systems and to re-evaluate them in the necessary ways. However, such a re-evaluation is mostly infeasible in practice, since these systems are mostly not accessible, except for off-the-shelf commercial products and a few chat bots. Also re-engineering them is impossible due to insufficient documentation and inaccessible resources. Furthermore, the fact that systems have been implemented for different languages makes direct comparison difficult. Therefore it seems necessary to implement a system specific to our purposes.

We have, however, examined the information about previous and current approaches to ICALL systems and used this as a background and source to motivate the specific parameters that we examine. In the remainder of this chapter we present the design of our research approach starting from the parameter space that we have discussed above.

6.6 Research design

Based on the three instances from the two-times-two dimensional space of implementational and pedagogical parameters that we introduced above, we will now describe the research design in more detail. This comprises a formulation of the research questions from the SLA perspective and an explanation of the methodological choices.

6.6.1 Research questions

From the SLA point of view, the purpose of this study is to examine and compare the effect of different types of computer-delivered instruction that vary with respect to (a) the importance they attach to either formal aspects of the language or the underlying meaning and communicative purposes and (b) how implicit or explicit they are.

The three instances were realized as variants of a text-based dialog system. Learners of German were recruited to engage individually with the system using a desktop computer. Their language skills were tested before and after interacting with the system. The interactive communication was framed within a meaning-based task that the participants had to solve. The task was devised such that it provided an opportunity to make use of specific linguistic target forms.

In Section 6.3, we have characterized the three instances of dialog-based instruction as free-input with recast feedback, free input with metalinguistic feedback and constrained input with binary feedback. In formulating the research questions, we will construe the three conditions with an eye to the SLA terminology we used in Section 4.2 and 4.3

First, in Section 4.2, we described the approaches that differ with respect to the focus they put on meaning or form. On the one hand, there is the accuracy-oriented FOCUS-ON-FORMS approach that focuses on forms in isolation, providing no or only very limited meaningful context. Opposed to that is the meaning- and fluency-oriented FOCUS-ON-MEANING approach. Finally, a FOCUS-ON-FORM approach tries to integrate meaning and forms by drawing learners' attention to linguistic forms as they arise within primarily meaning-based interaction. Using these terms, we formulate the first SLA-focused research as follows:

> SLA Research Question 1:
> Is there a difference in effectiveness between the effects of computer-based FOCUS-ON-FORM and FOCUS-ON-FORMS instruction?

Second, with the purpose of further examining different options within the FOCUS-ON-FORM approach along the implicit-explicit dichotomy (Section 4.3), we then investigate the effect of different types of feedback given to learners in response to their erroneous utterances. Feedback as a mechanism of drawing or directing learners' attention to formal aspects of language can vary with respect to its explicitness and obtrusiveness regarding the meaning-based conversation. *Recasts* are employed as an implicit, unobtrusive way to provide correct forms while keeping the primary focus on meaning. *Metalinguistic feedback* is employed as an explicit way to incidentally focus on forms during the conversation. The second question asks about the effectiveness of feedback that varies with regard to explicitness:

SLA Research Question 2:
Is there a difference in effectiveness between computer-delivered recasts and metalinguistic feedback?

In the following, we will discuss the experimental methods we used to answer these questions and specify more details of the implementation.

6.6.2 Methodological choices

This section describes and elaborates the motivation for the methodology used for answering the research questions. It therefore involves a characterization of the experimental design, including the selection and randomization of participants, as well as the elicitation of data.

Experimental design

When the goal is to compare different treatments, the common approach within SLA research is to operationalize the treatments as independent variables, and examine the effect on another, dependent variable. This approach is called *experimental*. In general, the objective of an experimental design is to determine whether there is a causal relationship between the variables in order to evaluate the effect of a certain treatment. By controlling all potentially interfering factors carefully, the experimental design tries to raise confidence that the variation in the independent variable is the reason for the variation in the dependent variable. In contrast to that, in correlational research, researchers explore relationships between existing variables that they do not control. While correlational research is concerned with co-occurrence, experimental research seeks to determine whether there is a causal relationship.

Between-subjects design

When the goal is to compare different treatment conditions, there are in general two options. One is that each subject experiences only one treatment condition - *between-subjects design*. In the alternative, each subject experiences all of the treatment conditions, but in a different order. The latter is known as a *crossed design*.[1] The advantage of a crossed design is that it requires fewer subjects and that it is not as sensitive to subject-individual differences, which loose their potential to become a confounding factor. However, a crossed design is not always feasible, depending on the nature of the treatment. The most important obstacle to varying the conditions of treatments within subjects is when the treatment has a lasting effect, thus inducing carry-over effects. Since we expect our treatment to have an effect that lasts over a certain amount of time, we cannot vary the treatment within subjects. Given that we examined two different target structures and task scenarios, another way to save subjects would have

[1] It is also known as *within-subject* or *repeated measures* design, because the treatment is varied within each subject and because measures are collected repeatedly. However, these two labels confound the distinction between the measurements of different treatment conditions on the one hand, and the repeated measurements under the same condition across time with the goal to examine temporal effects of the treatment on the other hand.

been to combine the type of instruction with one of the two target structures and to treat each subject with two of such combinations. However, this was not possible for two reasons. One was that the amount of time during which we had access to the participants was limited. The other reason was that the two target structures slightly differed regarding the proficiency level they were appropriate for.

Comparison group design

Between-group designs differ according to whether they include a true control group that does not receive any treatment (control group versus comparison group design). In general, the inclusion of a true control group is desirable in order to exclude any effects stemming from exposure to the tests, maturation, or disregarded external factors. This is especially important in our context since the German-speaking environment potentially provides considerable outside exposure to the German language. However, given the limited number of subjects we had access to, we opted for a comparison group design, rating the goal of comparing treatment conditions higher than the objective of evaluating the effect of a treatment as such. Thus, any conclusions regarding the effect of the treatment in itself need to be considered in light of the limitations mentioned above. As a matter of fact, the predominant reason for not including a true control group when designing the experiment was the ethical concern that we could not use the self-paying students' valuable course time for something without apparent value to them.

Randomization

One of the essential criteria of the experimental design is the random assignment of participants to comparison groups. Randomization ensures that each participant has the equal and independent chance of being selected for a group. The goal is to render groups that are statistically equal such that differences in the results are not the result of extraneous factors or pre-existing differences. Randomization controls both known and unknown variables. It converts unknown or unknowable systematic differences between group members into random quantities that follow probability distributions.

Since the participants were recruited from intact classes that differed in their overall level due to assignment to classes based on a placement test, the study employed a *randomized block design*. In such a design, the complete sample is divided into relatively homogeneous blocks and a fixed fraction of each block is randomly assigned to each control group. The underlying assumption is that the variability in each class is less than in the entire sample. Thus, by introducing a block for each class, we controlled for the assumed differences between classes.

Pretest-posttest design

The goal of this study is to examine the effects of a treatment, in the sense that the treatment causes a change in some measure related to language skills. This requires a comparison of pre-treatment and post-treatment performance. Although there are research designs that use post-treatment only, such a design is in general not desirable

because it cannot ensure that the experiment groups are comparable, i.e. have a similar level before the treatment. The main reason for choosing a posttest only design is that a pretest might have revealed the purpose of the study, which could have corrupted the results. We tried to avoid this problem by concealing the target structures through inclusion of distractor items in the tests (see Section 7.3.1 and Section 7.3.2).

Because we are interested in the accumulated time-related effect of our treatment, we used a *repeated-measures design*, repeating measures over the course of time (not over different types of treatment, see Footnote 1 above), and administered a posttest after each of two treatment sessions. In addition, to examine the long-term effects of the treatment, we included a delayed posttest that took place five weeks after the last treatment. The inclusion of delayed tests is important because it allows to evaluate the sustainability of a treatment (Long, 1991). The extent of the treatment and the number of treatment sessions were restricted by the time we could get access to the classes. While it would have been desirable to have more and longer sessions, unfortunately this was not possible.

Classroom

Traditionally, second language research distinguishes between classroom-based research and research conducted in controlled laboratory settings. Although our treatment does not require a classroom setting per se, and theoretically we could have administered it individually in laboratory settings, we used intact classes for the sake of convenience. It would have required much more logistical effort and expenditure of time to conduct the treatment with each participant individually instead of simultaneously. For each class that was employed, participants were randomly assigned to experiment groups. This was possible due to the one-on-one nature of the instruction: Each participant worked individually on one computer. However, this setup required that all conditions need to be similar to a degree that differences should not become obvious to participants seated next to each other. This circumstance would have made it somewhat problematic for a control condition that consisted of no treatment or only a dummy treatment.

Note that the common disadvantage of class-based research, namely that randomization is impossible, does not apply here, because we are able to vary the independent variable within classes. Furthermore, the observation that feedback seems to be more effective in laboratory settings than it is in the classroom (Nicholas et al., 2001, discussed above in Section 5.2.2) is unlikely to have much effect here, since the feedback is given individually.

Choice of language

We used German structures for two reasons. First, as native speakers we had the required knowledge to devise the tasks and system interaction. Second, of those languages for which we had the required expertise, German was the one for which we were likely to recruit the largest possible number of participants within our context. However, since this study was conducted in Germany all participants were in a second-language learning context. As a result, they were likely to progress and mature

independently from course content due to outside exposure. In comparison, learners in a foreign-language context, i.e. learning a foreign language while living in a country where this language is not spoken, have less outside exposure and thus, in general, progress more slowly.

6.6.3 Parameters of the experimental treatment

After we have discussed general parameters that concern the conduction of the experiment, we will now further narrow down the conditions for the actual instruction. In Section 6.3, we have characterized the three instances as free-input with recast feedback, free input with metalinguistic feedback and constrained input with binary feedback. In the research questions, we described them in terms of FOCUS-ON-FORM and FOCUS-ON-FORMS. FOCUS-ON-FORM and FOCUS-ON-FORMS are rather general characterizations referring to various kinds of instruction techniques that differ in important aspects, as we have discussed above in Section 4.2.1 and 4.2.3. However, for this study, we choose specific instances from the set of possible realizations of instruction. For instance, the FOCUS-ON-FORM characterization does not specify how explicit or implicit the instruction can be. We vary this property by the type of feedback, but there are other ways in which it could be varied too.

For FOCUS-ON-FORMS instruction, there are different kinds of controlled form-related activities, and there is some variability with regard to how much meaningful context is provided. Similarly, for FOCUS-ON-FORM instruction there are different ways to provide the meaning-based communicative activity which gives rise to a focus on form. In the remainder of this section, we will first discuss and justify the parameters we adopted for the FOCUS-ON-FORM instruction. Apart from the type of feedback which we already discussed above as an important determiner for the implicitness of the instruction (Section 6.3.1), these comprise the degree of planning that is involved in the instruction, and the assumptions and preconditions. In the end, we will describe the conditions for the constrained input FOCUS-ON-FORMS instruction.

Incidental versus planned

In general, when implementing FOCUS-ON-FORM pedagogy, the instructor can take a reactive or a proactive stance. As we have explained in more detail in Section 4.2.4, following the reactive approach, the instructor observes problems with forms as they become apparent and provides an immediate response. The proactive approach usually involves an a-priori need analysis or merely a curriculum-driven decision on which forms to teach, and the subsequent creation of meaningful contexts, i.e. *tasks*, which require learners to use the problematic form. Since the reactive approach is not feasible for a controlled study, we employ a proactive approach. To evaluate the progress of learners would require the performance of tests on a wide range of forms, which would require more time than we would have had in the context of our study. Furthermore, a reactive approach would have required us to cover a larger range of forms. In order to keep the treatment computer-based, we would have needed an all-purpose system that could handle a wide range of errors. A wider coverage may be viable from the perspective of error-diagnosis, given the recent advances in the treatment and pars-

ing of learner errors in German (Foth et al., 2004; Boyd, 2012). However, modeling the corresponding unconstrained dialog would have been likely to raise additional challenges, in particular the implementation of the pedagogical objectives and strategies that human teachers would adopt in similarly unconstrained conversations.

In the context of planned FOCUS-ON-FORM, we have also discussed *input enhancement* as proactive method to draw learners' attention to formal aspects of the language in an unobtrusive way by manipulating the input that learners receive (Section 4.2.4). Even though the type-written interface of our dialog system lends itself conveniently to include the visual enhancement of forms, we did not use this as a distinguishing parameter. However, this may be a worthwhile parameter to examine in future work.

Explicit introduction of forms as preparation

In Section 4.2.4 we discussed different manners to integrate meaning and form in a FOCUS-ON-FORM approach. Apart from a simultaneous or sequential integration, one method of integration is to precede FOCUS-ON-FORM activities with explicit teaching of the forms (Lightbown, 1998; DeKeyser, 1998). DeKeyser argues on the basis of skill acquisition theory that explicit procedural knowledge is the prerequisite of implicit automatic knowledge. Lightbown, in arguing that "brief focus on form in context is not the right time for explanations" (p.194), seems to imply that many linguistic structures indeed require an explicit metalinguistic presentation at some point, because FOCUS-ON-FORM alone is not sufficient or effective for inducing grammatical knowledge. However, Doughty and Williams (1998c) argue that the inclusion of distinct explicit teaching of forms as a preparation to FOCUS-ON-FORM activities cannot count as proper FOCUS-ON-FORM. By excluding it from their further considerations, they avoid discussing the necessity of such preparation. We neglected this question too and did not attempt to include explicit preparatory instruction for the forms in any of our experiment conditions. Note in particular that we do not aim to evaluate whether or not FOCUS-ON-FORM alone without preceding explicit instruction is sufficient. However, we assumed that all participants had received some kind of an explicit instruction at one point in their previous studies, but not necessarily in their current course and not all in the same form, given the different learning histories of the participants. This assumption was confirmed by interviews with the teachers who were responsible for the courses. We rely on some previous, most likely explicit, instruction of forms, because we assume that the treatment provided by our system alone is not sufficient to introduce entirely new knowledge.

Since we conceive of the system as a practice tool that relies on a existing knowledge, we did not aim to find subjects with a complete lack of knowledge of the target structures. Even though we consider it possible that the treatment provided by the system can have an effect on learners that had no previous exposure to the target forms, we assume that this would require a longer and more intensive treatment which was infeasible in the context of our study. In general, studies that assume zero knowledge of the target structure are very rare, Ellis (2010) assumes that this is due to the difficulty of finding a linguistic structure that is entirely new to a group of learners. However, in order to reduce the problem and minimize any influence by previous knowledge, it is possible to exclude participants who exhibit existing knowledge in a pretest (Long

et al., 1998).

Constrained input condition

After having described in detail the parameters for the two FOCUS-ON-FORM treatment conditions, which we developed with the goal of comparing the effect of implicit with explicit feedback, we now characterize the FOCUS-ON-FORMS condition. This condition allows us to compare the FOCUS-ON-FORMS approach with the FOCUS-ON-FORM approach. The main feature of the FOCUS-ON-FORMS condition is that the forms have priority while the meaningful context is reduced. When designing the FOCUS-ON-FORMS condition, the underlying objective was to make it as similar as possible to the other conditions in order to avoid any additional variance. In particular, we wanted all conditions to be performed on the computer, because this allowed us to run different conditions simultaneously within one class. If the conditions had been more different (e.g., one on the computer and the other not), it would have been more obvious to the participants that they were subjected to different conditions.

In addition, we wanted to ensure that there was no effect due to the inherent attraction of the medium that participants engaged with. This requirement ruled out any paper-based exercises.

Apart from implementing all conditions on the computer, we also tried to keep the interface as similar as possible. Therefore, we used the task scenarios (described below in Section 7.2) that prompt the interaction between learner and computer in the FOCUS-ON-FORM conditions to prepare a dialog and use this dialog to generate the form-focused prompts. Participants were asked to manipulate grammatical forms that would then become part of the scripted dialog. This means that the grammar exercises were embedded in an overall meaningful context. However, as opposed to participants of the FOCUS-ON-FORM groups, participants were not free to choose their own linguistic means. Moreover, the focus on form was established pre-emptively in advance and not incidentally as a response to erroneous input as typical for FOCUS-ON-FORM instruction. While the FOCUS-ON-FORMS condition could have been implemented with much less meaningful context, we aimed at rendering the conditions relatively similar.

6.7 Summary

This chapter laid out the approach we take to explore the potential of NLP-based ICALL for language learning. Our main premise is to focus on a small selection of instances and compare them with an in-depth SLA-oriented evaluation approach. We based the selection of instances from the perspectives of (1) pedagogic SLA-related concerns, and (2) implementational and technological concerns. In each of the perspectives two parameters play a part. From the pedagogic perspective, the parameters come out of (a) the continuum between implicit and explicit instruction and (b) the range between focus on meaning and focus on form. From the implementational perspective, the parameters arise from the scopes of (c) feedback informativity and (d) the freedom of input for the learner. These four parameters span a multi-dimensional

space. We motivated the choice of three instances from within that space and justified the disregard of possible alternative parameters. We then discussed the nature of each of the three involved research areas NLP/CL, SLA/FLL, and ICALL and their relationships. Based on this discussion, we positioned our approach within that context and characterized our contributions. We argued that our study contributes to all three areas in different ways. It develops a new ICALL system that serves to answer SLA-motivated questions and thereby generates new application-oriented knowledge regarding the practical use of NLP/CL methods for research purposes. We finished this chapter by formulating the SLA research questions and explaining and justifying the research design that we adopted. This served as a basis for the detailed description of the experiment that we will provide in the following chapter.

7 The Experiment

This chapter describes the details of the experiment we conducted to compare the different conditions and their effect. In Section 7.1, it introduces general considerations and criteria for selecting the target structures and then describes the structures and their properties in more detail. Section 7.2 specifies the tasks that serve as a meaning-based background for the instruction and describes the behavior of the dialog system and the interaction it affords. Section 7.3 discusses the range of tests that we used to assess the development of language skills. Finally, Section 7.4 describes the procedures and details of the data collection.

7.1 The target structures

Research about acquisition of German as a second language (GSL) has focused on a wide range of grammatical phenomena, while the areas of phonology or lexical acquisition have been relatively disregarded (Eckerth et al., 2009). One of the most widely researched topics among grammatical phenomena is word order (Clahsen, 1984; Ellis, 1989), others are case marking (Kempe and MacWhinney, 1998; Liamkina, 2008), gender (Rogers, 1987; Spinner and Juffs, 2008; Menzel, 2004; Mika, 2005), tense (Timmermann, 2005; Schumacher, 2005), modal particles (Rösler, 1982; Möllering, 2004), negation (Weinert, 1994; Meisel, 1997), and agreement between subjects and verbs (Rogers, 1984). Eckerth et al. (2009) provide an overview of the more recent GSL research conducted between 2002 and 2008.

The structures chosen for this study are *dative prepositional phrases* and *causal subordinate clauses*. A number of theoretical and practical issues guided this choice. First, it was necessary to devise a plausible task scenario in which the target structures were likely to be used. As we discussed in Section 4.5.2, not all structures are equally elicitable. For a preselection of potential structures we consulted relevant textbooks for German which were developed compliant with the Common European Framework of

Reference for Languages (Council of Europe, 2001; Trim et al., 2001; Glaboniat et al., 2005) and included real-life scenarios for the application of grammatical structures. An indication that the target structure was indeed problematic for learners was another criterion for the selection. We therefore consulted with teachers of German as a foreign language to determine the grammatical phenomena that were known to be difficult for the majority of their students.

A further practical concern was that it was feasible to test the acquisition of the target structures. A high degree of syncretism between distinct forms would make it challenging to test one form. This applies, for instance, to the accusative case in German, since most of the accusative determiner forms are identical with the unmarked nominative forms (see details below in Section 7.1.1).

The two structures we selected are of a different linguistic nature: Dative noun phrases are morphological, while the word order in subordinate clauses is a syntactical phenomenon. Further, as we will describe in more detail below, the two structures differ in aspects that influence their teachability as explained in Section 4.4. Differences between the structures were also a criterion for the choice.

After choosing the structures according to the criteria mentioned above, we conducted a pilot study to confirm that the structures could indeed be elicited within the tasks we had designed.

7.1.1 Dative case in prepositional phrases

German case system

Case is understood as a "grammatical category of inflected words which serves to indicate their syntactic function in a sentence." (Bussmann, 1998, page 62). Case marking of nominals is typically **governed** by the constituents that take the nominals as their complements. **Government** is conceived as "the lexeme-specific property of verbs, adjectives, prepositions, or nouns that determines the morphological realization (especially case) of dependent elements" (Bussmann, 1998, page 193). In German, the most prevalent instances of government are verbs and prepositions governing the case of their complement nominals.

German has four cases: nominative, genitive, dative, and accusative. In noun phrases, case is marked morphologically primarily by the determiner and sometimes in addition by a suffix on the noun. Table 7.1 provides the forms of the German definite article (corresponding to the English "the") for the three genders masculine, feminine, neuter and the two numbers, singular and plural.

The declension paradigm in the table illustrates the high degree of syncretism of the German case marking system. The forms have merged supposedly due to sound change. Since the articles (and other determiners) do not only mark the case but also gender and number, there are 24 different positions in the paradigm (4 cases * 3 genders * 2 numbers). However, there are only 8 different forms, each of which can have between 2 and 8 interpretations (Schwind, 1995). For instance, "der" realizes the singular masculine nominative as well as the singular feminine genitive and dative and plural genitive for all genders.

In summary, there is no one-to-one mapping between form and meaning, but case

	Singular			Plural
	Masculine	Feminine	Neuter	all genders
Nominative	der	die	das	die
Accusative	den	die	das	die
Genitive	des	der	des	der
Dative	dem	der	dem	den

Table 7.1 – Determiners and Cases in German NP

marking is conflated with gender and number marking (Spinner and Juffs, 2008), which increases the difficulty for L2 learners to interpret and produce the ambiguous forms.

Prepositions

Prepositional phrases consist of a preposition and its nominal complement. Prepositions in German can govern genitive, dative, or accusative case, with some prepositions governing exactly one case and others governing two cases.[1] Besides the alternation of dative and genitive, which is stylistically motivated, the alternation between dative and accusative is primarily based on semantic differences (Pittner and Berman, 2008; Eisenberg, 1999).

Prepositions that govern both dative and accusative, also known as "two-way prepositions" (Folsom, 1981), have a spatial meaning:

in	'in, into'
an	'at, on (up against)'
auf	'on, (down) on(to)'
über	'above, over, across'
unter	'under, below, beneath'
vor	'before, in front of'
hinter	'behind'
neben	'beside, next to'
zwischen	'between'

In general, these prepositions govern the dative case to describe a **location** ("when the place *in which* is denoted") and the accusative case to describe a **direction** ("when the direction *towards* or *into* an object is expressed" (Curme, 1970, page 378)). Folsom (1984) uses the terms "intralocal" and "translocal" to distinguish the two meanings. Eisenberg (1999) notes that the local and directional meaning of the prepositions do not differ regarding the spatial parameters, but that the difference is only temporal: When using the dative as in "der Bus an der Ostsee" ('the bus at the baltic sea') the spatial configuration described by the preposition "an" - 'at', holds at speech time, whereas when using accusative as in "der Bus an die Ostsee" ('the bus (heading) to/towards the baltic sea'), the spatial configuration will hold at a time after the reference time. Thus, the difference between location and direction can be defined in terms of differ-

[1] "entlang" - *along* is the only German preposition that governs all three cases

ent temporal references. The explanation given in German learning classrooms and text books, however, is usually phrased in terms of the location/direction contrast, despite the fact that, as Folsom (1981) and Schröder (1978) note, this dichotomy is a simplification that needs to be further elaborated for specific cases.

In addition to the two-way prepositions, there are prepositions with a spatial meaning that exclusively govern dative case. Among them are *bei* ('at, by, near') and *gegenüber* ('opposite') which indicate a location. There are also *zu* and *zu* in combination with *bis* ('to, towards') which indicate a direction, but in contrast to the meaning distinction for two-way prepositions exclusively govern dative case.

Prepositional phrases

Prepositional phrases have three different syntactic functions according to Eisenberg (1999): complements (1), adverbials (2), and attributes (3), (4)[2].

(1) Helga hofft auf den Durchbruch.
Helga hopes for the breakthrough.

(2) Ilse frühstückt in der Küche.
Ilse has breakfast in the kitchen.

(3) Helgas Hoffnung auf den Durchbruch
Helga's hope for the breakthrough

(4) Ilses Frühstück in der Küche
Ilse's breakfast in the kitchen

As complements – also referred to as prepositional objects – prepositional phrases are governed lexically by the verb. The verb determines the preposition as well as the case. In (1), the verb "hoffen" governs the preposition "auf" and the accusative case. As adverbials, prepositional phrases further qualify the event described in the clause. They can either refer to the verb or the whole clause, but they are not obligatory and can, in principle, qualify any clause. They are thus not governed by the verb. In (2), the prepositional phrase "in der Küche" ('in the kitchen') specifies the location where Ilse is having breakfast, but "frühstücken" ('to have breakfast') does not require a local specification. In the adverbial usage the meaning of the preposition is usually concrete, while the preposition in a prepositional object has often lost its lexical meaning through a process of abstraction (Eisenberg, 1999).

When prepositional phrases work as prepositional attributes, the relationship between the attribute and the nominal that it further specifies can be similar either to that of prepositional objects or to that of adverbials. In (3) the relationship resembles prepositional objects, in that the prepositional phrase is obligatory and governed by the nominal. "Hoffen" ('hope') is the nominalization of "hoffen" ('to hope') and as

[2]Example (2) is from (Eisenberg, 1999, page 293).

such governs the same prepositional object as the verb. Opposed to that, consider (4), which is similar to the adverbial relationship, i.e. the prepositional phrase is not governed by the nominal and not obligatory.

Note that distinguishing between prepositional objects and prepositional adverbs is sometimes controversial and not always as obvious as in the examples. There are borderline cases which cannot be easily determined, considering that the criteria to discern them are also subject of debate (see Eisenberg (1999) for details).

Regarding the relationship between the verb and its prepositional object, we can distinguish two situations. In one, the verb has a concrete local meaning and uses a local prepositional object to realize this meaning as in (5).

(5) Er wohnt in dem Haus.
He lives in the house.

In the other, the concrete meaning of the preposition related to the verb has gone through a process of abstraction and the government is merely syntactically motivated, as in (1). For these instances, case government is a formal feature that is determined lexically and thus arbitrarily. The verbs and the prepositional objects they govern have to be learned item by item by the L2 learner, because there are no readily available syntactic or semantic criteria (Eisenberg, 1999). On the other hand, for the more concrete and less abstract meanings it is possible to state semantic rules to determine the case, as we have seen for the distinction between locational and directional usage of two-way prepositions. Within the scope of this study, the targeted prepositions will mostly have a concrete spatial meaning.

Learnability

Morphological case markers in German determiners are not particularly salient. A clue for that claim is found in patterns of first language acquisition, which show that German pre-nominal case markers are acquired much later than suffixed verb inflections, and also later than suffixed case markers in other languages (Slobin, 1973; Szagun, 1997). These patterns led Slobin to posit a processing strategy that he hypothesized must be at work in first language acquisition: "Pay attention to the end of words!". If we understand this strategy and the observed acquisition patterns as a predictor for salience, we can argue that case marking in determiners is not salient. Of course, the processing strategies in second language acquisition differ to a certain degree from first language acquisition, but there is some confirming evidence for second language learners too – Diehl, Pistorius and Dietl (2002) showed that francophone learners of German acquire case marking relatively late compared to verb inflections. Another argument for the lack of salience is that case markers are usually unstressed and in some cases hard to distinguish, e.g., *dem* versus *den* (Szagun, 1997).

In addition to the complexity of the morphological case marking system with its high degree of syncretism and the low salience of morphological markers, a further difficulty for the L2 acquisition is the fact that correct case marking is often semantically redundant. Incorrect case marking is seldom essential for conveying the meaning (see Section 4.4.3), because the semantics of a verb often expresses sufficient informa-

tion. Consider for instance the verbs "legen" ('to put, to lay") and "liegen" ('to lie, to be located, to rest'). "legen" governs a prepositional object in accusative case to indicate the target zone of the put-action, which is directional in nature, see (6). Opposed to that, "liegen" governs a prepositional object in dative case to denote the static location of the rest-event as in (7).

(6) Das Buch liegt auf dem Tisch.
The book lies on [the table]$_{dat}$.

(7) Ich lege das Buch auf den Tisch.
I put the book on [the table]$_{acc}$.

(8) and (9) are derivations of the examples above with incorrect case marking. We believe that these erroneous examples are comprehensible to native speakers when communicating with non-native speakers, therefore they will not cause a communication breakdown.

(8) *Das Buch liegt auf den Tisch.
The book lies on [the table]$_{acc}$.

(9) *Ich lege das Buch auf dem Tisch.
I put the book on [the table]$_{dat}$.

The only instances in which the interpretation depends on correct case marking are those where a directional prepositional object (governing accusative case) can alternate with a static locational adverbial using the same preposition (governing dative case). Consider for instance (10), where the prepositional object refers to the goal of the movement and (11), where the prepositional adverbial indicates the location in which the movement is situated:

(10) Er rennt (fährt, springt) hinter das Haus.
He runs (drives, jumps) behind [the house]$_{acc}$.

(11) Er rennt (fährt, springt) hinter dem Haus.
He runs (drives, jumps) behind [the house]$_{dat}$.

Here, the case marking on the determiner is crucial for conveying the meaning. In theory, this kind of minimal pair is possible for all verbs governing a directional prepositional object, which are mostly verbs indicating a movement. However, given that a typical target of a directional action is often distinct from a plausible location of that action, we assume that these instances are fairly infrequent. As we will show in Section 7.2.1 these problematic instances are also not likely to occur in our task scenario.

Regarding scope and reliability (cf. Section 4.4.2) – how frequent is the structure and how many exceptions are there to the linguistic rule – marking of dative case in prepositional phrases can be considered a reliable rule that is wide in scope. Dative prepositional phrases are common in oral and written German. Compared to

accusative and genitive, dative prepositions are the most frequent (Folsom, 1984). All prepositions require case-marking of the noun phrase they take, and for many, the case they govern is unique and fixed. However, two-way prepositions, which can take both dative and accusative have a considerable frequency. Folsom (1984) cites corpus studies in which frequency counts range from 39 to 50 percent of all prepositions. In these instances, the preposition is not sufficient to indicate the required case. Instead, the case is either governed by the specific verb or it has to be inferred by examining if the prepositional phrase indicates a location or a direction.

In the discussion of salience above, we already have hinted at orders of acquisition and cited the study by Diehl et al. (2002). They observed that case marking is acquired relatively late in a foreign language learning context, compared to the acquisition of the verb forms and word order. They attribute this to the complexity of the case morphology and its functions as well as the limited communicative value. They further note that case marking appears in prepositional phrases before it appears in nominal phrases, however, target-like case marking in PPs seems to occur only after successful acquisition of case marking in NPs.

Contractions An additional source of difficulty is the fact that some prepositions can be contracted with the definite article into one word. Contraction is only possible for unstressed articles. Some contractions are restricted to colloquial use, whereas others are part of standard German. Below is a list of standard German contractions for definite articles in dative case :

an	+	dem	→	am	'at, on (up against)'
bei	+	dem	→	beim	'at, by, near'
in	+	dem	→	im	'in, into'
von	+	dem	→	vom	'of, from'
zu	+	dem	→	zum	'to'
zu	+	der	→	zur	'to'

As we will show in Section 7.2.1, our task focuses on the use of dative prepositional phrases in a giving directions scenario. The spatial two-way prepositions mentioned above are expected to be used in their static locational meaning (governing dative case) to anchor landmarks in the path descriptions. In addition to these prepositions, we also try to elicit prepositions with exclusively static locational meaning like *bei* ('at, by, near') and *gegenüber* 'opposite' and the directional preposition *zu* and *zu* in combination with *bis – bis zu* ('to, towards') all of which exclusively govern the dative case.

The correct production of dative prepositional phrases requires two distinct pieces of knowledge: (1) to know that the employed preposition governs the dative case (in general or in a specific semantic context), and (2) to know how the dative case is marked. Furthermore, correct dative case marking also requires knowledge about the gender of the noun to be marked. It follows that there can be several causes for a failure to produce an accurate dative prepositional phrase. Error diagnosis and feedback provision may need to consider these.

7.1.2 Word order in subordinate clauses

German word order

The word order in German is relatively free, however, the word order rules are rather complex and pose considerable difficulties for learners of German. As opposed to English or French, parts of the verb cluster can be separated and as such build a bracket - *Satzklammer* (sentence bracket) (Pittner and Berman, 2008). The sentence bracket and the positions before, within, and after (Vorfeld – pre-field, Mittelfeld – middle field, Nachfeld – post-field) comprise the so-called *topological field model* which is used to describe the complex constraints governing German word order. In German, usually three types of clauses are distinguished according to the position of the finite verb (Verbstellungstypen): verb-initial (V1), verb-second (V2), and verb-final (VF) position (Eisenberg, 1999; Pittner and Berman, 2008). Examples for each of the types are provided by (12) - (16), where the finite verb is underlined.

(12) Ich arbeite in der Bibliothek. (V2)
I work at the library.

(13) Ich habe in der Bibliothek gearbeitet. (V2)
I have worked at the library.

(14) Arbeitest du in der Bibliothek? (V1)
Do you work at the library?

(15) Arbeite in der Bibliothek! (V1)
Work at the library!

(16) ... weil ich in der Bibliothek arbeite. (VF)
... because I work at the library.

For the verb-initial and verb-second clauses, there is a strong relation between pragmatic function and the verb position. For instance, polar questions (14) and imperative sentences (15) are usually realized as sentences with verb-initial position, while declarative sentences (12), (13) are usually realized as verb-second sentences. In contrast, verb-final clauses (16) are not related to a particular pragmatic function – their defining property is their subordination to a main clause. In both verb-initial and verb-second clauses, the finite verb constitutes the left bracket. The difference is that the Vorfeld is empty in verb-initial clauses while it is not in verb-second clauses.

The prototypical word order of subordinate clauses is verb-final position together with an introductory word at the first position of the clause, which constitutes the left bracket. The introductory word can be a subordinating conjunction, an interrogative pronoun, or a relative pronoun. Clauses in this form are also called “Spannsatz” (’span clause’), referring to the introductory word and the finite verb in final position

which together span the clause (Eisenberg, 1999). In (16), the introductory word is the conjunction "weil" ('because'), the finite verb is "arbeite"('work'$_{1per-sg}$).

Default order Given the frequency of declarative sentences, the verb-second position is arguably the most frequent one (Pittner and Berman, 2008) and from a pragmatic point-of-view, is therefore often considered as the default position, (cf. Eisenberg (1999) for references). However, from the perspective of generative grammarians, the verb-final position is considered the default, unmarked one, because, in contrast to the other positions, all parts of the verb cluster are placed together at the end. There are no discontinuities like in the other position types. In the non-finite position types the verb cluster is discontinuous with the finite verb constituting the left bracket and other parts of the verb cluster build the right bracket (Bierwisch, 1963; Pittner and Berman, 2008). Another argument for the verb-final position as default is that in infinitive constructions the verb follows all other constituents - "der Versuch, im Haus einen neuen Leiter zu finden" ('the attempt to find a new head in-house') (Bierwisch, 1963, page 35). Similarly, the format of lexical citations of verbs and their complements places the verb in the end after all complements, e.g., "jemandem etwas geben" (as opposed to English: "to give somebody something") (Bierwisch, 1963, page 35). Clahsen and Muysken (1986) cite the fact that in clauses with a modal or auxiliary verb the lexical verb is in final position (see (13)) as another argument for the final position being the default.

Functions of subordinate clauses

Subordinate clauses have different functions. They can be complements (subject or object) that are governed by the verb of the main clause. They can be attributes that further specify a noun, usually realized as a relative clause. Besides these, an important function of subordinate clauses is that of an adverbial. As adverbial clauses they are independent of the verb in the main clause, but they further specify the proposition in the main clause. The relation to the main clause is determined by the meaning of the subordinating conjunction that introduces the adverbial clause (Eisenberg, 1999). Conjunctions can express temporal, causal, instrumental, conditional, final, adversative, concessive, and consecutive relations between the main clause and its subordinate clause. For the scope of this study we focus on causal clauses introduced by the subordinating conjunction *weil*. The proposition in the weil-clause contains the reason or cause for the proposition given in the main clause. The reason for choosing the causal clause is that it is relatively easy to elicit in a natural task, compared to other subordinate clauses, as we will describe in Section 7.2.2.

Learnability

The word order of subordinate clauses has been shown to be problematic for learners of German (Rogers, 1982). Clahsen and Muysken (1986) observed that adult learners of German go through a phase of using subject-verb-object (SVO) word order in subordinate clauses, probably overgeneralizing the observed canonical main clause word order (which (Pienemann, 1989, page 55) claims to be "psychologically the simplest

way of marking underlying grammatical and sentence-semantic relations"). Summarizing the results of the ZISA (Zweitspracherwerb italienischer, spanischer und portugisischer Arbeiter) project (Clahsen, 1984), which is concerned with natural (non-instructed) L2 learners of German, Ellis (1989) indicates that among all word order rules, the verb-end rule is acquired the latest. Ellis then shows that instructed learners followed the same order of acquisition for word order rules as natural learners, regardless of the sequence in which the rules were introduced and the amount of emphasis given on these rules. In contrast to L2 learners, there is no evidence of L1 German learners producing word order errors in subordinate clauses (Clahsen and Muysken, 1986).

However, recently, the universality of the acquisitional sequence of word order shown by Clahsen (1984) and Ellis (1989) was challenged by conflicting findings. For instance, Diehl et al. (2002) observed that subordinate clause word order is acquired before subject-verb inversion by adolescent native French speakers. Lund (2004), on the other hand, could not find a fixed order for the acquisition of inversion and verb-final order for adult English native speakers.

The relationship between form and function of subordinate word order can be considered as complex (for a general discussion of complexity see Section 4.4.3). The problem arises from the fact that the function of an adverbial clause is coded by the meaning of the conjunction. The repositioning of the finite verb does not carry meaning, but it is a purely formal requirement. It is semantically redundant, given that an erroneous positioning would not change the meaning nor make the whole sentence incomprehensible. Additional complexity for causal clauses comes from the fact that there is a coordinating conjunction (*denn*) with the same causal meaning. Since coordinating conjunctions like *denn* introduce main clauses, the finite verb should be in second position. Consider examples (17) and (18) which illustrate the difference.

(17) Ich kann nicht, weil ich arbeiten muss.
I can't because I must work.

(18) Ich kann nicht, denn ich muss arbeiten.
I can't because I must work.

Thus, a learner who arrives at the hypothesis that the clause-final position of the verb is related to the causal meaning in a weil-clause could be falsely deriving a subordinate word order for clauses introduced by *denn*.

Regarding the regularity of subordinate clause word order, it can be considered a reliable rule with wide scope (Section 4.4.2). However, there are two important restrictions: One is the lower frequency of subordination in oral language compared to written language, the other is the trend in oral German to use *weil* as a coordinating conjunction.

The tendency of spoken German to prefer coordination over subordination affects the scope of the subordinate word order rule. Corpus studies have repeatedly indicated a different distribution of main and subordinate clauses in written versus spoken German (Engel, 1974). Within the spectrum of spoken German, the frequency of subordination increases with the formality of the context and the extent to which the utter-

Figure 7.1 – System Interface

ances were prepared. Spontaneous and informal speech shows the lowest proportion of subordinate clauses (Elmauer and Müller, 1974; Weijenberg, 1980). One possible reason to avoid subordinate structures in spontaneous, conversational spoken language is that they are more complex. The higher complexity requires more processing efforts than speakers can normally afford in such situations (Chafe and Danielewicz, 1987).

Related to that, there is a tendency of subordinate clauses in oral language to adapt to the word order of main clauses (Günthner, 1996, 2008). Haag (1985) interprets this as a sign of economy. This trend pertains specifically to the use of *weil* as a coordinating conjunction (Gohl and Günthner, 1999), a fact that effects the reliability of the the rule. Although learners of German might be unlikely to encounter *weil* as a coordinating conjunction within the classroom or while attending to written material, it is possible that they encounter it in informal interaction with native speakers or while consuming authentic audio(-visual) media.

The **salience** of word order in subordinate clauses somewhat hard to characterize. In comparison to case marking which is realized by relatively short suffixes of the determiner, subordination involves the repositioning of whole words. This might be easier to notice and thus more salient. However, since we do not know of any study that examines the noticeability of word order phenomena in German, these assumptions remain somewhat speculative.

7.2 Tasks and interaction

This section describes general concerns regarding the instructional activities that we employed for the experiment. According to the FOCUS-ON-FORM approach, the attention to forms should be embedded in a communication-driven, meaning-based context. As we described in Section 4.5.2, *tasks* provide the opportunity to use language in situations that are similar to real life. While engaging in a task, learners attend

to meaning rather than form (Nunan, 1989). However, forms should not be entirely disregarded if accuracy in production is an objective (cf. Section 4.2.3). The purpose of *focused tasks* then is to increase the likelihood that certain forms are used without enforcing them explicitly. In this section, we will describe the focused tasks that we designed to implement the FOCUS-ON-FORM instruction and the constrained FOCUS-ON-FORMS variants that we derived from them.

Recall that we examine three different experimental conditions:

1. free input with recast feedback (FOCUS-ON-FORM)
2. free input with metalinguistic feedback (FOCUS-ON-FORM)
3. constrained input (FOCUS-ON-FORMS)

We will describe all of them in detail below. We start with a general characterization of the interaction with the dialog system that serves as a partner to engage in the experimental activities. After that, we illustrate the task scenarios and properties of the task-related interaction. This includes a presentation of the expected utterances by the learners and the system's strategy (a) to elicit the target forms and (b) to provide feedback.

In all three experimental conditions the learner interacts with a computer by means of giving textual input in response to a textual system prompt, and, in turn, receiving a response and a new prompt by the system. The computer interface to realize prompts and responses and to receive learner input also includes additional information about the tasks. Figure 7.1 illustrates the system interface for one of the tasks. On the left hand side the task is described and additional material is provided. The right hand side contains the dialog history on the top (*Dialog-Verlauf*) and the input area (*Eingabe*) along with additional control buttons on the bottom.

The system initiates the interaction by providing a prompt. The dialog is modeled with a consideration of adjacency pairs that constitute the local structures of dialog and grounding mechanisms that we discussed in Section 3.1.1. The system attempts to demonstrate and check its understanding of the leaner utterance by providing a rephrase of parts of the learner utterance.

The dialog is managed based on a finite state-based dialog model. The finite-state model is enhanced with a few global state variables that further control some of the state transitions. In the two free input FOCUS-ON-FORM conditions, learners are allowed and required to freely formulate their dialog contributions. In the constrained FOCUS-ON-FORMS condition, learners are constrained to provide the target form. The feedback in this condition explicitly states whether or not the supplied form was correct. For the two FOCUS-ON-FORM conditions, the system provides recasts or metalinguistic feedback respectively.

The following two sections describe each of the two task scenarios in more detail and specify the dialog model that the system adopts.

7.2.1 Giving directions task

Giving directions usually involves instructions on how to navigate with reference to given landmarks. The usage of spatial and directional prepositions in this context is

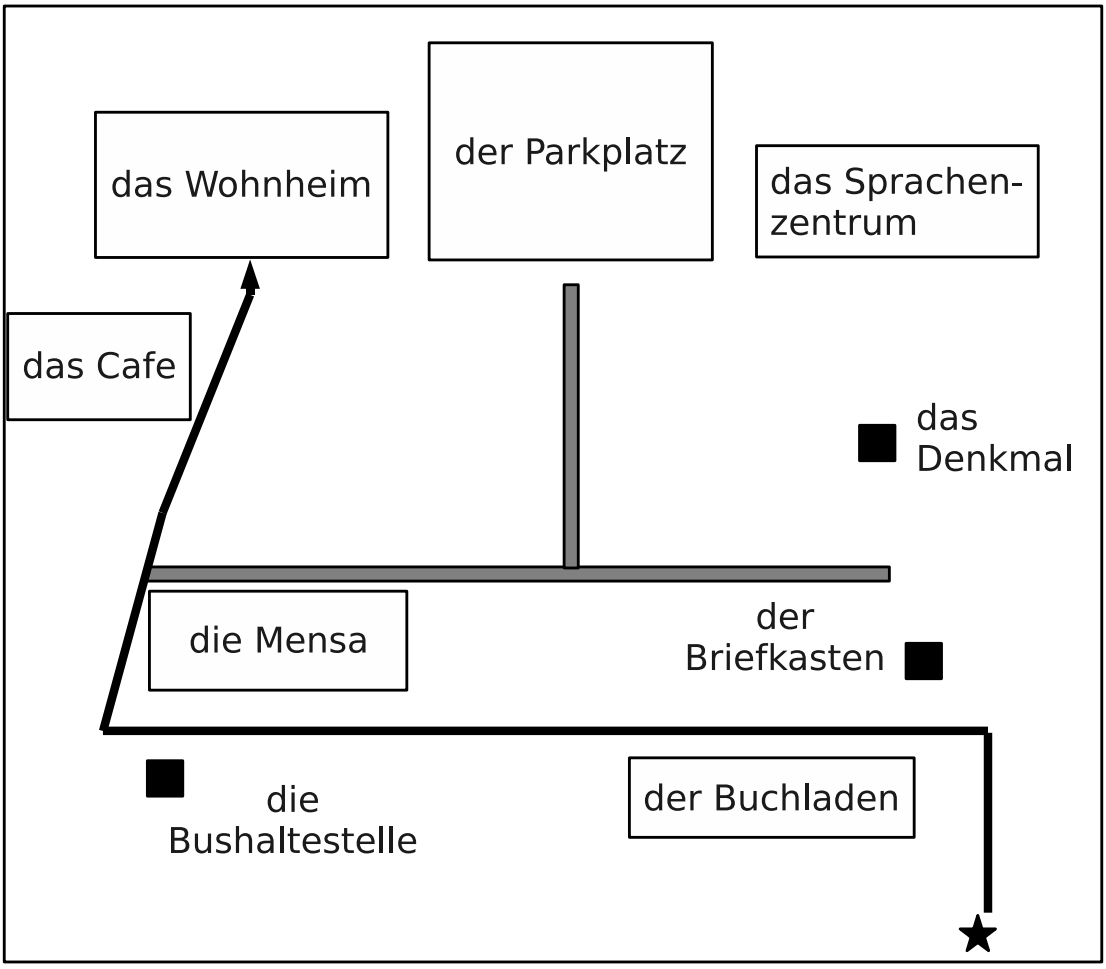


Figure 7.2 – Task material from giving directions: the map

natural. As we explained in Section 7.1.1, spatial prepositions that govern both dative and accusative case govern the dative case when referring to locations. In addition the local preposition *bei* ('at','by','near') exclusively governs the dative case as does the directional preposition *zu* ('to'). This makes "directions giving" a suitable task scenario for eliciting dative prepositional phrases.

For this task, participants were asked to give directions and describe a route: They were given a simplified map of a fictitious campus or city, with buildings and landmarks and the route that they were supposed to describe. Figure 7.2 shows an example of the actual material given to the participants. The task was phrased like this:

> "You are at the university campus. Your task is to give directions. Someone stops you and asks you for directions. You are at the point indicated by the star at the bottom. Provide directions according to the marked route."

Given the task criterion that learners should be free in their choice of linguistic means, the task description did not contain any explicit hint to use prepositional phrases or pay attention to dative case.

We placed landmarks of different gender at turning points along the route and

close to the target landmark with the goal to create opportunities for the learners to refer to those landmarks. The landmarks were labeled with their gender in order to ensure that learners have access to the correct gender information. This was supposed to eliminate insufficient gender knowledge as an error cause.

We did not include any streets or crossroads in the map in order to prevent any reference to those. Pilot testing with more realistic maps had shown that learners prefer to refer to crossings instead of other landmarks. We wanted to prevent this, because, although they can be used in dative prepositional phrase, crossings and streets are both of feminine gender and thus participants would not have had enough opportunities to build dative PP with masculine and neuter nouns.

The route includes two points of direction change. At each of these points, there are two landmarks and the target is also placed close to two other landmarks. We balanced the landmarks regarding their gender such that the two landmarks at one of the points of direction change have both feminine gender, and both masculine at the other point. All landmarks close to the target have neuter gender. We expected the participants to refer to at least one of the landmarks at each turning point and to one landmark close to the target. We further expected them to refer to the landmarks by using a dative prepositional phrase as for example in (19) (underlined).

(19) Gehen Sie <u>hinter dem Cafe</u> nach links.
Turn left, past the coffee-shop!

A pilot study confirmed that the learners largely do refer to landmarks in contexts that require dative case. However, two types of avoidance strategies occurred, but only rarely. One strategy was to refer to landmarks in perceptional statements, e.g. “Dann sehen Sie den Buchladen.”(“Then you see the book shop”). Perceptional verbs usually govern an accusative object which refers to the phenomenon that is perceived. The other strategy was a non-standard way to use two-way prepositions in a directional sense, in which they govern accusative case, e.g. “Gehen Sie hinter das Cafe!” (“Go behind the coffee shop”). While this usage is not formally incorrect, it is unusual in a directions giving scenario. Given that both of these avoidance behaviors occurred very rarely, we did not implement any remedial measures to suppress them.

For the sake of variety, each treatment session in the experiment consisted of two different variants of the task, which differed with regard to the route but contained the same landmark configuration.

System strategy

We start by highlighting the important aspects of the system strategy here, but the complete dialog model is given further down. The core feature of the dialog system is to give corrective feedback on erroneous dative prepositional phrases produced by the learner. Apart from this, the system glosses over any other errors in the learner production. If the learner avoids dative prepositional phrases, the system tries to elicit them and provides examples for them. It also has a strategy to extend the dialog if the learner tries to provide the complete route description in one turn in the beginning. It is important to note that the dialog model does not check the validity of the given route

description. We assumed that such a consistency check was not necessary. Even if the learners had given false directions intentionally or unintentionally (which rarely happened), the grammar-related purpose of the task – the practice of dative prepositional phrases – was still served.

Feedback

If the learner utterance contains an incorrectly realized dative prepositional phrase, the system gives corrective feedback. As we have explained in Section 7.1.1, the dative case is required with prepositions that govern the dative case in general (e.g., zu, bis zu – 'to') or by a two-way preposition used for describing a location as in directions like "take a left at the cafeteria" or "the coffee-shop is in front of the dormitory".

In the recast condition the system provides implicit feedback by reformulating (*recasting*) the learner's utterance (or parts thereof). Recasts were implemented in a manner so as to have them carry the additional meaning of an *acknowledge* grounding act, as in (20). **S** and **L** mark system and learner turns respectively. The bold emphasis did not appear in system output and is used here only to indicate the incorrect form and its correction via recast.

(20) **L:** Gehen Sie vor **das** Cafe nach links.
'Turn left, in front of the coffee-shop'
S: Okay, [vor **dem** Cafe nach links,]$_{RECAST}$
[und dann?]$_{PROMPT}$
'Okay, left in front of the coffee-shop, and then?'

In the metalinguistic feedback condition, the system explicitly states that there is an error, points to the location of the error and elicits a correction by the learner, as shown in (21). In case the learner does not succeed in correcting the error, the system gives a further hint, as in (22).

(21) **L:** Gehen Sie vor **das** Cafe nach links.
'Turn left, in front of the coffee-shop'
S: ['das' in 'das Cafe' ist nicht richtig.]$_{ML-FB}$
'das in 'das Cafe' is not correct.'
[Bitte noch einmal!]$_{PROMPT}$
'Please try again!'

(22) **L:** vor **den** Cafe nach links.
'in front of the coffee-shop'
S: ['den' in 'den Cafe' ist auch nicht richtig.]$_{ML-FB}$
'den in 'den Cafe' is not correct either.'
[Nimm Dativ!]$_{PROMPT}$
'Use the dative!'

Note that both recast and the particular type of metalinguistic feedback that we pro-

vide are not dependent on the cause of the error. That means that the system is unconcerned about whether the learners do not know that the preposition they used governs the dative case or if they do not know how to mark the dative case. The feedback merely addresses the failure to realize the dative marking. For recasting, the error source is not relevant. While metalinguistic feedback could be more precise and informative with error source information, the interaction is too short to reliably estimate the source of the error. Therefore we keep the metalinguistic feedback so general that it is suited for all error sources. Admittedly, the second hint in (22) may be confusing if the learner intended to produce a dative prepositional phrase but the learner can still infer that their attempt was not successful.

Another common error that appeared in that task is the attempt to use *bis* ('to, towards') as a directional preposition. In that context *bis* can only be used in combination with *zu*. In the recast condition, the system adds *zu*. In the metalinguistic feedback condition, the system demands that *zu* is added or *zu* is used alone (23).

(23) **L:** Gehen Sie **bis** dem Cafe.
'Go to the coffee-shop'
S: ['bis' kann hier nicht allein stehen. Nehmen Sie **bis zu** oder nur **zu**.]$_{ML-FB}$
"bis' cannot stand alone here. Use 'bis zu' oder only 'zu'.'

If in the metalinguistic feedback condition an error could not be identified, the system falls back to providing a recast. Section 8.1 will explain how this was implemented. In both conditions correct learner utterances are also recast by the system to signal its understanding. For the repetitive recasts the relevant information is incorporated in the confirmative grounding move of the system, just like in the corrective recast illustrated by (20).

Eliciting and providing the target form

If the learner does not use the target form, the system either tries to elicit it or provides an example for using it. If the learner gives the order to change the direction but does not refer to a landmark, the system requests a clarification by asking where to change the direction. As an answer we expect a prepositional phrase that specifies the location of change. Consider (24):

(24) **L:** Gehen Sie nach links!
'go left!'
S: [Wo soll ich nach links abbiegen?]$_{ELICIT}$
'where do I turn left?'

If the learner refers to landmarks in perceptional statements, or if the utterance cannot be parsed and only a reference to a landmark can be detected, the system requests a clarification and suggests that the landmark was referred to as a intermediate goal. It thereby provides an example of the target form as in (25).

(25) **L:** Dann sehen Sie den Buchladen.

'Then you see the book shop.'

S: Ich verstehe nicht. [Soll ich bis zu dem Buchladen gehen?]$_{CL-REQUEST}$
'I don't understand. Should I go up to the book shop?'

Ensuring minimal dialog length We anticipated that some learners might give a complete route description in one turn at the start of the dialog. We identify such attempts by counting the number of commands to change the direction. If the learner's turn includes more than one change of direction, we deem it as too long. The system asks the learner to slow down and only the part of the utterance that refers to the first change of direction is further analyzed and responded to. Consider (26) as an example:

(26) **L:** Gehen Sie erst hinter dem Buchladen links, dann bei der Mensa rechts, dann geradeaus bis zum Wohnheim.
'First take a left after the book shop, then take a right at the cafeteria, then straight forward to the dormitory.'

S: Nochmal langsam.
'Wait a second!'

S: Also erst hinter dem Buchladen nach links.
'So, first to the left after the book shop.'

S: Und dann?
'And then?'

Dialog model for the directions giving scenario

Figure 7.3 illustrates the overall dialog model for the directions giving scenario. Rectangles refer to system actions. The grey diamond shapes indicate case distinctions on properties of the learner utterance. For each case, an edge goes out from the decision node and leads to a system action or another decision case. Note that the presentation is slightly simplified in order to highlight the important aspects of the model. The system initiates the conversation by asking for the way to a goal landmark according to the map [0]. The first case distinction regards the length of the learner response and triggers the slow down response in case of a too long utterance [1].

In general, the learner utterance is classified as one of the following four moves or combinations thereof. The learner can indicate a straight forward move, a change of direction, the arrival at a subgoal, or a reference to the goal landmark. If the learner utterance is interpreted as forward move or a reference to a subgoal, the system acknowledges [3] these and elicits a continuation [9]. If the learner refers to the target landmark, we assume that the description is complete. However, if the learner has not produced any direction change so far in this dialog, the system will acknowledge the target landmark reference but ask for further directions [2]. If the learner had produced at least one change of direction before, the system acknowledges the target landmark reference and finishes the dialog [4]. Information about whether the learner has indicated a direction change previously is stored in an additional state variable. If the learner produces a direction change that includes a correct prepositional phrase, the

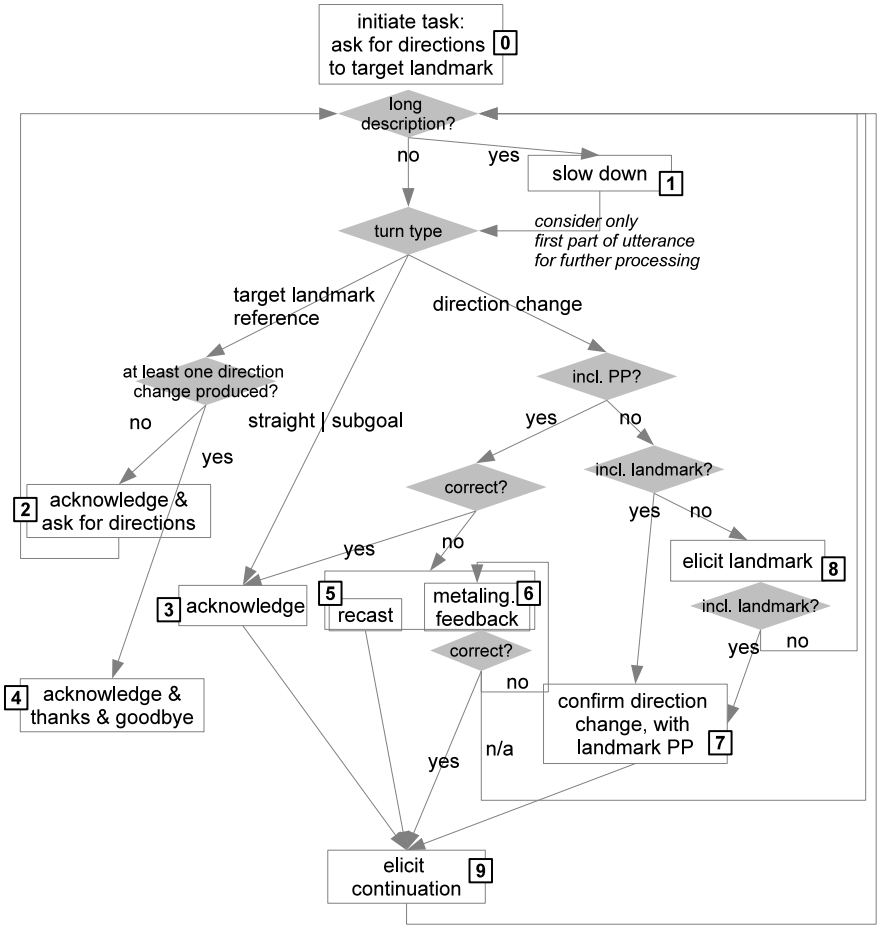


Figure 7.3 – Dialog model for directions giving scenario. Rectangles refer to system actions. The gray diamond shapes indicate case distinctions on properties of the learner utterance. For each case, an edge goes out from the decision node and leads to a system action or another decision case.

system acknowledges [3] and elicits a continuation [9]. If the direction change contains an incorrect prepositional phrase, the system gives corrective feedback, depending on the experiment condition. In the recast condition, the system produces a recast [5] and then proceeds to elicit a further description [9]. In the metalinguistic feedback condition [6], the system expects a correction. If the learner succeeds with the correction, the system proceeds to elicit a further description [9]. If the learner fails at the correction, the system provides another metalinguistic hint or repeats the previous until the learner either succeeds or attempts to produce something unrelated. In the latter case, the system abandons the correction attempt and proceeds to interpret the learner utterance as in the default case.

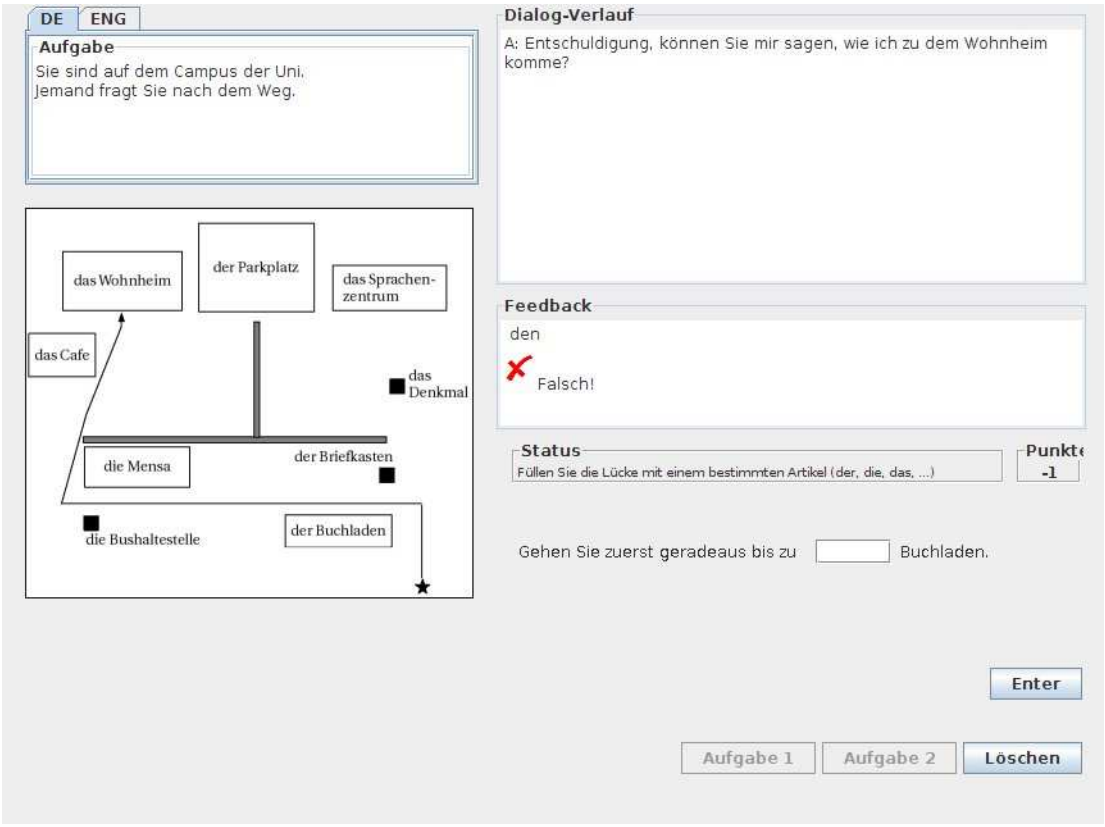


Figure 7.4 – System interface for the FOCUS-ON-FORMS condition with constrained input: the learner has to fill a gap and gets explicit feedback

Constrained input

In the FOCUS-ON-FORMS condition activity, unlike in the free FOCUS-ON-FORM conditions, the learners have no opportunity to freely produce an utterance and to choose the linguistic means they judge necessary. Instead, the exercise focuses strongly on the target structure: The learner is required to fill a gap in a pre-scripted dialog turn as in the example below. The learner is told that the gap is to be filled with a definite article:

(27) **S:** Wie komme ich zur Mensa?
How do I get to the cafeteria?
L: Gehen Sie hinter ☐ Cafe nach links.
Turn left past the coffee-shop

The learner is allowed three attempts to produce the correct form. If an invalid form is supplied, the system signals it with a message 'That was wrong!'. The feedback is displayed in a designated feedback area. After the third unsuccessful attempt the system appends the correct utterance to the dialog history. The system then generates its next turn based on the dialog model. Figure 7.4 depicts the actual interface. In addition to the verbal feedback, the system provides a score. Each correct response increases the score by one, while each incorrect response leads to subtracting of one point. We introduced the score feature with the goal to encourage the learner to try hard to be correct. There are in total three utterances with gaps to be filled to be as similar as possible with the free input conditions.

<table>
<tr><th>Zeit</th><th>Montag</th><th>Dienstag</th><th>Mittwoch</th><th>Donnerstag</th><th>Freitag</th></tr>
<tr><td>9</td><td rowspan="2">Seminar</td><td rowspan="2">Japanisch-Kurs</td><td rowspan="2">Vorlesung</td><td rowspan="3">Arbeit</td><td></td></tr>
<tr><td>10</td><td></td></tr>
<tr><td>11</td><td rowspan="3">Vorlesung</td><td></td><td rowspan="2"></td><td></td></tr>
<tr><td>12</td><td></td><td rowspan="3">Demonstration gegen Krieg</td><td></td></tr>
<tr><td>13</td><td rowspan="3">Lernen mit Florian</td><td></td><td></td></tr>
<tr><td>14</td><td rowspan="3">Opa im Krankenhaus besuchen</td><td></td><td></td><td></td></tr>
<tr><td>15</td><td></td><td></td><td rowspan="4">Wandern (wenn kein Regen)</td></tr>
<tr><td>16</td><td rowspan="5">Arbeit</td><td></td><td></td></tr>
<tr><td>17</td><td rowspan="2">Klavier-Unterricht</td><td></td><td></td></tr>
<tr><td>18</td><td rowspan="3">Joggen (wenn gutes Wetter)</td><td rowspan="3">Wohnung putzen (wenn nötig)</td></tr>
<tr><td>19</td><td rowspan="2">Vivaldi-Konzert</td></tr>
<tr><td>20</td></tr>
</table>

Figure 7.5 – Task Material for Making Appointments: The Agenda

7.2.2 Appointment task

In order to elicit causal clauses, we created a task that induces participants to provide reasons and justifications. The most obvious way to elicit a reason is to ask "Why?". A more subtle way is to create a situation in which the interlocutor is likely to provide a justification on their own accord. As we discussed above in Section 3.1.1, Levinson (1983) argues that some conversational actions are preferred over others. For instance, the preferred response to a request or an offer is acceptance, whereas a refusal is dispreferred. Dispreferred responses "are typically delivered [...] with some account of why the preferred second cannot be performed" (Levinson, 1983, page 307). So, if the task conditions force the interlocutors to refuse an offer, they can be motivated indirectly to give a justification for their refusal.

A suitable task scenario for this is to arrange an appointment. We try to induce refusals by proposing a time that is in conflict with the schedule of the learners. The learners received that schedule as part of the task. An example of such a schedule is given in Figure 7.5. The task was phrased as follows:

> "This is your day planner. Your task is to make an appointment. You and a fellow student are working on a project together. He wants to meet you to work on it. In the dialog he will propose possible times. Agree or disagree giving him information from your schedule. Give as much information as possible."

The entries in the schedule serve as material for the expected refusals. They are either expressed by a noun (e.g., *Arbeit* ('work')) or by a verb phrase, with the verb in infinitive form (e.g., *Wohnung putzen* ('to clean flat')). Furthermore, the task provides opportunities to produce conditional clauses. Conditional clauses are also subordinate and thus require the same word order as causal clauses. We try to elicit them by providing schedule entries that are tied to conditions, for instance: "hiking, if no rain".

Again, like in the directions scenario, the task description does not contain any explicit hint to use or pay attention to causal or conditional clauses. This is because learners should be free in their choice of linguistic means.

Each treatment session consists of two different variants of the task, which differed with regard to the character adopted by the system. In one variant, it was a friend, in the other variant, it was a boss or supervisor. This difference involves a different level of politeness, reflected most notably in the use of pronouns to address the dialog. As the superior character, the system addressed the learner with the polite form, which in German is realized with the pronouns for third person plural – *Sie*. In the role of the friend character, the system used the familiar form of address, i.e., the second person singular – *du*. Since this difference was mainly introduced for the sake of variety, the system did not provide feedback if the learner violated these conventions in their own utterances.

The system proposes appointment times known to be occupied on the learner's schedule and expects the learner to refuse the proposal and give a reason. However, although it may be expected to provide an excuse, this is not obligatory, and neither is it obligatory to phrase it as a subordinate clause. So we expected that it would be harder to elicit subordinate clauses than, for instance, to elicit dative prepositional

phrases in the directions scenario, because they are not as essential for the task. Our assumption was indeed confirmed by the pilot study we conducted. However, this study also gave evidence that learners could be primed to use this structure by giving them examples of the structure within the dialog. Thus, we designed a system strategy that provided examples of the target structures either as part of its own refusals or as recasts of learners' refusals that did not realize the target structure.

System strategy

Similar to the directions scenario, the main purpose of the system used for the FOCUS-ON-FORM condition is to maintain a task-driven conversation with the learner and give corrective feedback if errors occur that relate to the word order in subordinate clauses. Any other errors are ignored. Because the causal clauses are not as essential for the task as dative prepositional phrases in the directions scenario, the dialog model contains a wider range of strategies to elicit them and provide examples for them.

The system behavior can be summarized in the following way: The system proposes five time slots that are in conflict with the learner schedule. There is no check if the given reason for refusal is consistent with the items in the given schedule. However, if the learner unjustly accepts a system proposal that is in conflict with the schedule, the system revokes its proposal and continue the dialog. The system refuses any learner proposals. If the learner does not offer a proposal, the system explicitly asks for one. After the system has proposed five impossible times, it concludes the dialog by proposing a time that is compatible with the learner schedule. We will now illustrate the relevant system moves with examples.

For the appointment scenario we only implemented the FOCUS-ON-FORMS condition and the FOCUS-ON-FORM with recast. As we will explain later in Section 7.4, we decided to not test the FOCUS-ON-FORM with metalinguistic feedback condition with the second scenario due to the scarcity of participants.

Feedback

The FOCUS-ON-FORM system provides implicit feedback in case of learner errors in the target form by recasting the learner's utterance. Similar to the directions scenario, recasts are phrased as implicit confirmations, as by **S**, below, which corrects the error made by **L**, by correcting the word order but realizing it in a second person form:

(28) **L:** *Nein, ich kann nicht, weil ich muss arbeiten.
'No, I can't because I have to work.'
S: Ah, du kannst nicht, weil du arbeiten musst.
'Ah, you can't because you have to work.'

Just as in the directions task scenario, the system always confirms the refusal reasons by uttering a reformulation like above, even if the learner's utterance contained no error. In this case, the recast serves as a confirmative grounding move that signals to the learner that the system has understood the refusal. The uniform response to correct and erroneous utterances also allows a simplification of the error diagnosis approach

as we will explain in Section 8.1, since we do not need to recognize errors as long as we arrive at a valid interpretation.

Eliciting and providing the target form

Besides the corrective feedback, the dialog model encodes three strategies to elicit causal clauses if the learner does not use them spontaneously:
(**A**) If the learner gives a reason for refusal but does not produce a subordinate clause, the system recasts the refusal into a subordinate clause. In addition, the conjunction **weil** is set in bold font in order to increase its salience. However, no other form of input enhancement is used to make the learner aware of the crucial characteristic of the target form, namely the clause-final position of the finite verb. Consider (29) as an example:

(29) **L:** Nein, ich kann nicht, ich muss arbeiten.
'No, I can't, I have to work.'
S: Ah, du kannst nicht, **weil** du arbeiten musst.
'Ah, you can't **because** you have to work.'

(**B**) If the learner fails to give a reason in their refusal the system asks for one explicitly:

(30) **L:** Nein, am Montag um 15 Uhr kann ich nicht.
'No, I can't make it on Monday at 3.'
S: Warum kannst du denn nicht?
'Why can't you make it?'

(**C**) In order to present examples of a causal clause not as part of a recast, but as part of an original refusal, the system will refuse all learner-initiated proposals with a reason formulated as a causal clause:

(31) **L:** Ich kann am Montag um 4.
'I have time on Monday at 4.'
S: Da kann ich nicht, **weil** ich arbeiten muss.
'I can't **because** I have to work.'

If the learner does not propose a time on their own account, the system asks the learner what day and time would suit them. It thereby elicits a proposal, only to then refuse it.

Note that the feedback and eliciting efforts are targeted on causal clauses only. Although the schedule items provide opportunities to produce conditional clauses, the system does not give corrective feedback for them, nor does it try to elicit them. This is because the focus of the task and instruction is on causal clauses, while the conditional clauses serve merely as a bonus to provide more varied stimuli and an opportunity to produce other subordinate clauses.

Dialog model for the appointments scenario

Figure 7.6 depicts the dialog model for the appointments scenario as a finite-state automaton. As an extension to the state representation, two counter variables store the number of *weil*-clauses that the learner attempted to produce (x) and the number of proposals the system made (y). The system initiates the dialog with an introduction [1] and the first proposal for an appointment which is in conflict with the learner's agenda [2]. If the learner accepts the proposal even though it is at odds with their agenda, the system retracts the proposal [3] and proceeds. In case of a rejection, the learner gives a reason or not. If they don't provide a reason, the system tries to elicit one [4]. If the justification of the learner is not given in the form of a *weil*-clause, the system recasts the learner's reason as a *weil*-clause in form of confirmation [5] and proceeds. If the justification includes a *weil*-clause, the system response depends on the correctness of this clause. If it is correct, the system acknowledges the justification [6] and proceeds, if the there is an error, the system gives a corrective recast [5]. If the learner proposed an alternative time slot along with their refusal, the system rejects and uses a *weil*-clause example for the justification [7]. If the learner did not propose another time slot, the system either tries to elicit a proposal from the learner [8] or produces another conflicting proposal [2], or closes the dialog if it already uttered 5 proposals [9]. The system elicits a proposal if it already prompted the learner with two proposals and the learner did not attempt to use a *weil*-clause (y=2 & x=0), or if it produced more than two proposals in the entire dialog and there was no *weil*-clause produced by either the learner or the system since the last system proposal. If the learner in response proposes a time slot, the system rejects the proposal using a *weil*-clause [10]. Otherwise, the system issues another proposal [2].

Constrained input

As in the FOCUS-ON-FORMS condition in the directions scenario, the exercise focuses on the form and does not allow free learner input. Considering the nature of target form – word order – the exercise is to put a set of randomly ordered words into the correct order. Again, the thusly created utterance becomes part of an enfolding appointment negotiation dialog that is displayed in the history area of the interface, as in (32):

(32) **S:** Kannst du am Montag um 10 Uhr?
Are you available on Monday at 10am?
L: Nein, ich kann nicht, weil (arbeiten) [muss] [ich]
No, I can't because I have to work.

See Figure 7.7 for a screenshot of the actual system. The learner can either type the words in the respective blank fields or move them via drag and drop. As in the directions scenario, the learner has three trials. If they fail the third, the system appends the correct solution to the dialog and continues with the next prompt. In total there are five prompts presented, to be as similar as possible with the free input condition.

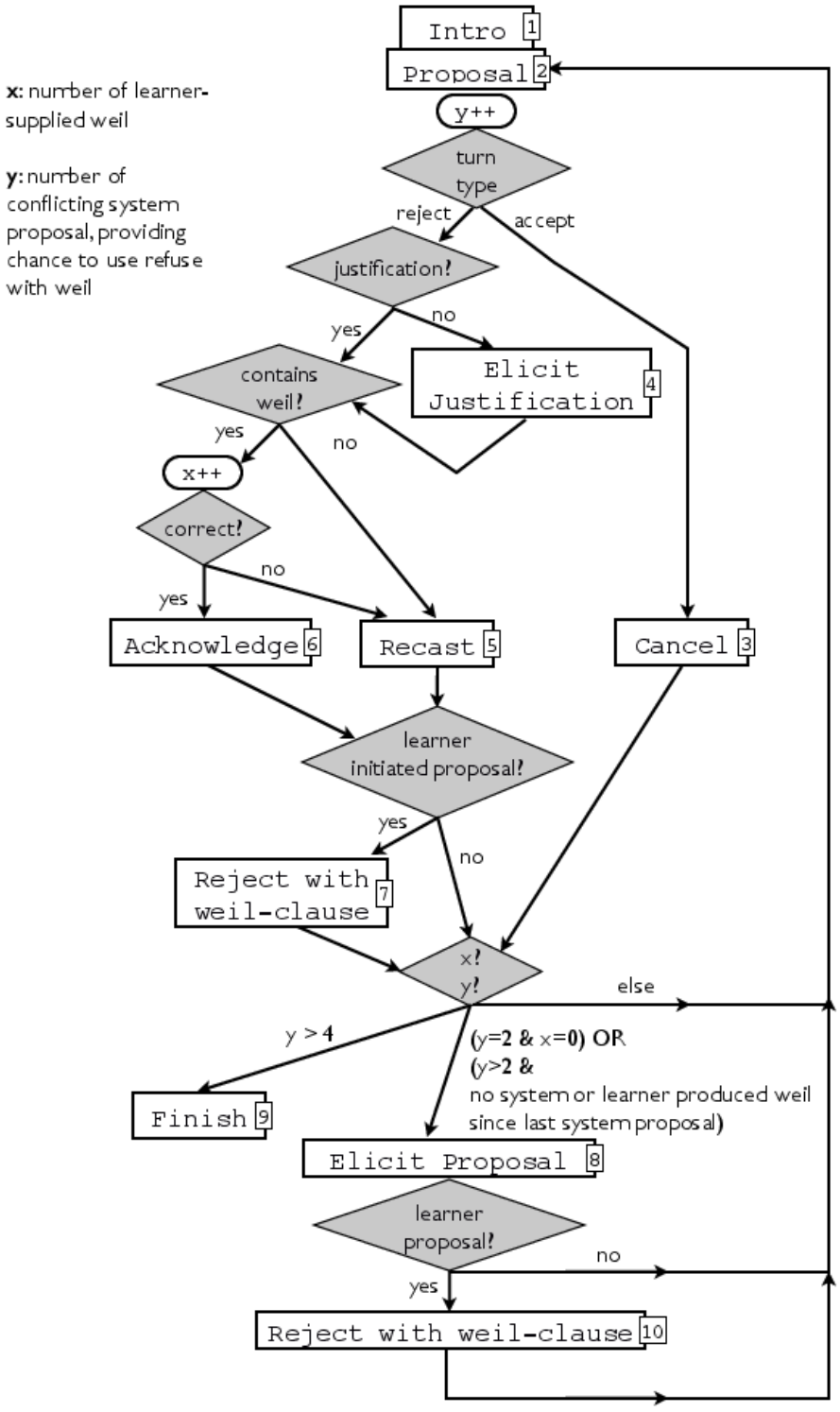


Figure 7.6 – Dialog model for appointments scenario. Rectangles refer to system actions. The gray diamond shapes indicate case distinctions on properties of the learner utterance. For each case, an edge goes out from the decision node and leads to a system action or another decision case. The capsule-shaped boxes indicate an incrementation of the counter variables.

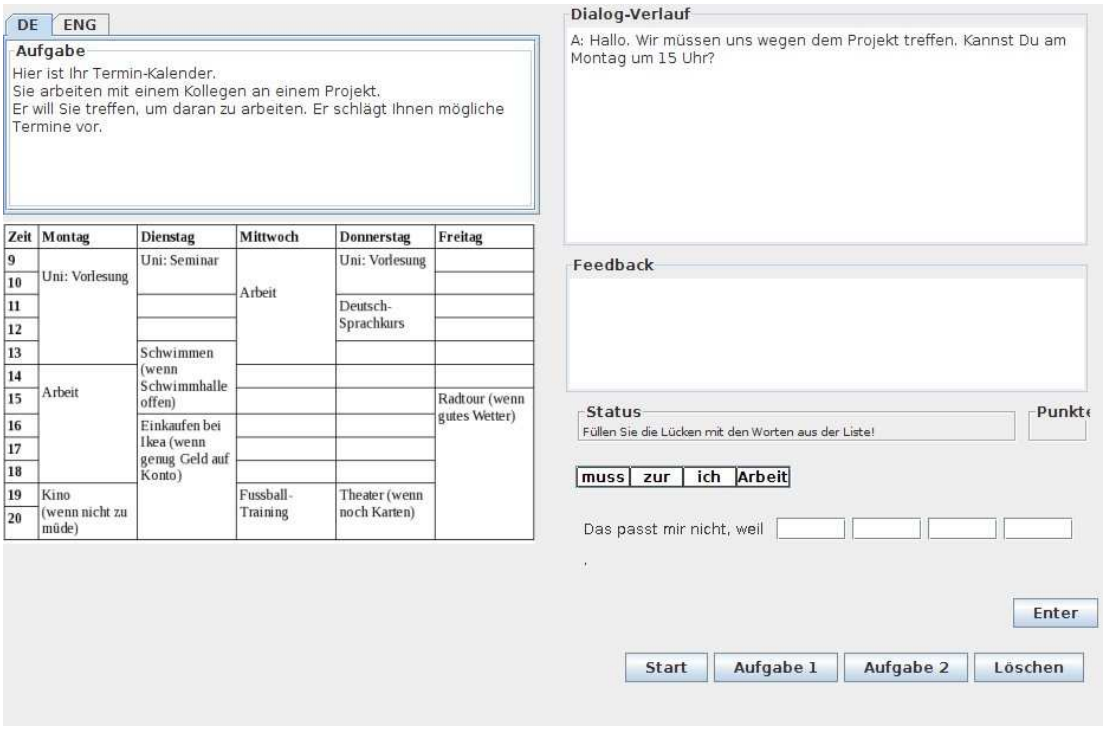


Figure 7.7 – System interface for the FOCUS-ON-FORMS condition with constrained input: the learner has to arrange the given words into the correct order and gets explicit feedback

7.3 Assessment of linguistic development

The goal of this study is to compare the effect of different types of instruction on the development of language skills. A key question then is how to assess the development of language skills.

As we have argued in more detail in Section 4.3, the contrast between explicit and implicit is an important dimension to characterize learning, knowledge, and instruction of language. It is generally assumed that second language learning and acquisition result in both explicit and implicit knowledge (Ellis et al., 2009), and that language instruction should not only engender explicit metalinguistic knowledge, but implicit knowledge as well. It is therefore desirable to assess the development of both types of knowledge. This is especially relevant for our study, since we compare types of instruction that differ in their degree of explicitness and implicitness, which possibly have different effects on the respective types of knowledge. Drawing on previous research (as discussed in more detail in Section 4.3.6), we selected one test for each of the two types of knowledge: a *sentence construction* test with no time constraint that targets explicit knowledge, and a *grammaticality judgment* test with a time limit that targets implicit knowledge of the target structure.

In addition to the tests that focus on learners' control of the target structure, we also want to assess the development of communicative skills with the task scenario that was used to practice the forms. Therefore, we collected samples of spontaneous language use by asking participants to perform a task analogous to the instruction task, but in a spoken dialog with a peer. The focus of this test is to measure the overall

fluency in oral production. Such a holistic approach is often disregarded as it goes beyond relatively controlled and straightforward form-related measures and focuses on language in actual use (Doughty, 2003).

It requires more effort both for administration as well as analysis. Considering this effort, we only analyzed the results for a subset of participants. However, we think it is worth to evaluate language skills in this manner because it allows us to address an important argument for meaning-based instruction: the ability to use language within a meaningful context as a tool, as opposed to having abstract knowledge about language rules.

In the following sections we discuss the tests and measures in more detail. The first two sections describe the two tests that focus on the target structures, the third section gives an account of the measures used to assess the oral communicative skills.

7.3.1 Implicit knowledge: timed grammaticality judgment test

In Section 4.3.6 we described the features of measures that test implicit knowledge. Measures for implicit knowledge:

- induce learners to respond according to their 'feel' rather than to conscious rules
- do not require metalinguistic knowledge
- focus on meaning rather than form
- limit the response time

Our selection is in particular based on the last criterion. In a grammaticality judgment test, learners are asked to indicate whether or not a given item is grammatically correct. Such a task admittedly draws attention to form because it does not serve a communicative goal and the stimuli are presented in isolated contexts. However, when a time limit is enforced for the judgment task, it can be used as a measure for implicit knowledge (Han and Ellis, 1998; Ellis, 2006, 2009b, and more details in Section 4.3.6). This is based on the rationale that a time limit forces learners to rely on their implicit knowledge, since they do not have enough time to access their explicit knowledge (Ellis, 2006). However, it is not exactly clear how to determine the appropriate length of the time limit. While it should prohibit the use of explicit knowledge, it should allow enough time to process the item semantically (Loewen, 2009).

The limits that have previously been used in timed grammaticality judgment tests range from 1.8 to 10 seconds per item (Bialystok, 1979; Ellis, 2006; Han and Ellis, 1998; Mandell, 1999). Only Ellis (2006) explicates how he devises his time limit. He trialed native speakers and used their average response time as a basis for the limit he imposed on L2 learners. For the learners he added 20% of the average time, which resulted in limits between 1.8 and 6.24 seconds per item. However, this limit would, in some cases, prevent the slow native speakers from succeeding. We therefore chose a more generous time limit of 10 seconds per item, based on a trial with native speakers. We use twice the time the slowest native speaker had used. We do not assign limits for each test item individually because, unlike the items in Ellis (2006), our items are similar in length and difficulty.

Note that the judgment test only involves the comprehension, but not the production of the target features. As a consequence it cannot make any prediction about the production performance of the learner. Even though it measures implicit knowledge, it cannot predict if learners are able to use this implicit knowledge to produce language accurately and fluently. A common test for implicit knowledge that involves production is the *elicited oral imitation test* (e.g., Erlam (2006)). For this test, learners are presented with an auditive stimulus (a sentence, clause, or word), which they have to repeat orally after a short delay. This procedure requires that a learner is tested individually or, if more learners are tested simultaneously, sound proof facilities are needed. Since we did not have access to such equipment this test was unfortunately not feasible for us.

We mentioned above that tests are likely to assess implicit knowledge if they involve language use that serves a communicative purpose and is unplanned and unfocused on forms. However, the problem with such tests based on free production is that they are difficult to rate. In particular, it cannot be reliably predicted if learners will use the target structure (Erlam, 2006).

Considering these problems, we did not use the free production test as a source for the evaluation of accuracy. Nevertheless, we do employ an oral communicative task, described below in Section 7.3.3, which elicits meaning-focused and relatively uncontrolled language use. We use it to test general fluency, but not the accuracy of the target structures.

In summary, while the timed grammaticality judgment test does not fulfill the criterion to focus on meaning (rather than form) it matches the other three criteria for implicit knowledge test: The available time is limited and as a result, learners cannot make use of their metalinguistic knowledge or conscious rules, but have to rely on their feel (Ellis, 2009b). While alternative measures, in particular elicited oral imitation and free production, fulfill all four criteria, they are more difficult to administer or rate.

For the timed grammaticality judgment test, we created four versions to be administered at four times of assessment (T1, T2, T3, T4). The versions differed in the combinations of prepositions, noun phrases, and verbs, but were otherwise comparable with regard to the lexical items used. The assignment of a test version to a test time was randomly varied for each participant in order to compensate for any unintended differences between test versions. Within each test, items were presented in random order. The tests were prepared and administered using the *Webexp Experimental Software*[3]. Each correctly judged item was scored at 1 point, each item that was incorrectly or not judged at all was scored at 0.

Test items for dative prepositional phrases

The test items included six different prepositions with a spatial meaning. Four of them were two-way prepositions: *auf* ('on', *hinter* ('behind'), *neben* ('next to'), *vor* ('in front of')). In the items, they were used in a context to describe a static location (as opposed to a direction). This was realized by verbs that describe a state like *stand*, *lay*, and *sit*. The other two prepositions were *zu* ('to') and *bei* ('at'), they exclusively govern dative

[3] http://www.hcrc.ed.ac.uk/web_exp/

case.

As part of the prepositional phrases, the items included nouns of the three genders in equal proportions. A problem with testing case marking is the dependence on gender knowledge. The grammatical gender of a noun is essential for realizing the correct case inflection of the determiner. In order to minimize insufficient gender knowledge as an error source, we used common feminine and masculine nouns whose grammatical gender matches the semantic gender, e.g. mother, man, son, etc. For neuter nouns we chose frequent nouns that are likely to be taught at the beginner's level, e.g. *Kind* ('child'). However, it was not possible to verify beforehand if all learners had indeed learned this gender information already.

The test included 9 grammatical and 9 ungrammatical test items and 7 grammatical and 7 ungrammatical distractor items. Unfortunately, we had to exclude one of the ungrammatical test items for the evaluation, because we had overlooked a spelling error.

Test items for subordinate clauses

The test items included subordinate clauses of different complexity. The complexity varied according to the amount of additional material present in the clause, e.g. objects, modal verbs, negations or additional modifiers. The test included 6 grammatical, 6 ungrammatical test items based on causal clauses with the conjunction *weil* ('because'). It further contained 9 grammatical and 9 ungrammatical distractor items. Two of these were conditional clauses with the conjunction *wenn ('if')* and another two were subordinate clauses connected by the conjunction *dass* ('that'). We had originally planned to analyze the performance on them further to gather information about the generality of the knowledge and the ability of the learners to transfer rules to other subordinate clauses. However, as it turned out later, their number was too small to get to a reliable analysis.

7.3.2 Explicit knowledge: sentence construction test

The features of tests for explicit knowledge arise as counterparts to the features of tests for implicit knowledge (summarized above). More precisely, learners are encouraged to use metalinguistic knowledge and search for rules in order to respond to the test stimuli. Further, the test draws attention to linguistic form and it gives learners sufficient time to access their explicit knowledge. The sentence construction test that we employ, meets especially the latter two criteria. There is no time-limit on the test items and the items do not serve a communicative purpose. Participants are asked to complete sentences given the beginning of a sentence and a set of unordered uninflected phrases or words. Full noun phrases were given along with gender information, as in the example below:

(33) **Item:** Das Pferd (stehen, die Kuh, vor)
Solution: Das Pferd steht vor der Kuh.
The horse stands in front of the cow.

(34) **Item:** Ich habe keine Zeit (weil, arbeiten, müssen, ich)

Solution: Ich habe keine Zeit, weil ich arbeiten muss.
I don't have time, because I must work.

This test does not explicitly require for metalinguistic knowledge. Consider, for instance, an grammaticality judgment test, that not only demands the learner to judge but also to indicate the error if an item was considered ungrammatical. However, we did not use this test, because we feared overlap with the timed grammaticality judgment test (see Section 7.3.1), that could result in boredom and an increased practice effect.

Another test which focuses even more on linguistic form is a selected response test, in which the participant has to choose a correct answer from a given set. Here we feared that the test would have been too similar to the FOCUS-ON-FORMS treatment condition (see Section 7.2), which could have favored the FOCUS-ON-FORMS group and induce boredom.

As with the timed grammaticality judgment test, we created four different versions of the tests that were assigned to participants randomly across the four different time points. The items were presented randomly.

Test items for dative

The test consisted of eight test items containing six different prepositions – four two-way prepositions – *hinter* ('behind'), *neben* ('next to'), *vor* ('in front of'), and *zwischen*, ('between') – and two others that only govern dative case: *bei* ('at') and *zu* ('to'). As with the judgment test, the nouns were equally distributed across gender. In addition, the test contained four distractor items. Although there are more, we consider these to be the most relevant prepositions for the directions scenario. Note that the used prepositions differ slightly between the two tests types for practical reasons: For instance, although 'between' is a relevant preposition, we did not use it in the judgment test, because it requires two noun phrases that have to be judged at the same time which makes it hard to attribute on which the judgment was based.

Similar to the judgment test, we used frequent and prototypical verbs that indicate a state (as opposed to a movement), like *stand*, *lay*, and *sit*. Except for the preposition *zu* ('to'), which was used in three test items, all others were used in only one item. We gave this prominence to *zu* because it is the only preposition with a directional meaning. Each item was scored one point if the prepositional phrase was built correctly. The item with the preposition 'between' was scored at one point for each correct noun phrase. All other form errors were neglected. Similar to the task material, we provided gender information for nouns to rule out any errors originating in missing or incorrect knowledge about the gender.

Test items for subordinate clauses

The test consisted of 6 test items for causal conditional clauses with the subordinating conjunction *weil* ('because'). Similar to the judgment test we also included two items with different subordinating conjunctions *dass* ('that') and *wenn* ('if, when'), as well as one item with the coordinating conjunction *denn* ('because') and three three other

distractor items. The items were scored one point if the word order was correct. All other form errors were neglected.

7.3.3 Communicative skills

The tests described above assess the control of the target structures. However, they cannot assess how the learner uses language in communicative meaning-based contexts. A common method to characterize learner language in more general terms, without the focus on specific target structures, is to elicit samples of free production and analyze them in terms of *fluency, complexity* and *accuracy* (Larsen-Freeman, 2006). For this study we specifically look at fluency, because we suspect that this aspect can be influenced by the freedom of language use and the role of meaningful context during the instruction. We do not analyze the overall accuracy, because we do not expect that it is influenced by the type of instruction. For the same reason we do not take into account the complexity exhibited in the sample of spontaneous speech.

In order to elicit samples of spontaneous, oral language use with a communicative purpose, we posed a task that is analogous to the task in the instruction and asked learners to complete this task in pairs. The ensuing conversations were recorded and edited for further evaluation. This task was completed at three instances: as a pretest before the treatment, as a posttest after the first treatment and as a delayed posttest five weeks after the second treatment. An additional posttest similar to the other tests directly after the second treatment was not possible due to time constrained (further explicated in Section 7.4 below).

For the directions giving task each partner receives two different maps (adapted and simplified from Map Task (Anderson et al., 1991)). Each map contains five different landmarks labeled with names and gender information. One of each participant's maps contains a route to describe, which is missing in the other partner's map. The maps are identical otherwise.

Each participant is the directions giver once and the receiver at the other time. The giver has to describe the route indicated in their map, while the receiver has to draw the described route in their corresponding map.

For the appointment task, each partner receives a schedule of a week which contains several free slots and several slots filled with different plans or commitments. To ensure a sufficiently long negotiation phase, the two partners' agendas were synchronized such that were only two slots free for both of them. As part of the task instruction, each participant was given a leisure activity (e.g., going to the cinema or museum, having dinner, etc.) and the order to convince the other partner to meet for this activity and find a suitable time for both. Thus, each dyad had to find and agree on two time slots.

The recorded conversations were edited to remove irrelevant, non-task-related conversation. In a second step, each conversation was split into two samples, each of which primarily contained the utterances of one speaker only. Apart from short confirmations and clarification questions, longer utterances by the other partner were excluded. If the conversation was unusually long, we cut off the sample at 90 seconds. This resulted in samples with a length between 30 and 90 seconds. The purpose of

cutting very long conversations was to achieve a more homogeneous sample length and not to overstrain the raters that were supposed to evaluate the samples later.

By fluency we understand the ability to talk rapidly, coherently, smoothly and without hesitation or reformulation (see Kormos and Dénes (2004) for a more extensive review of definitions for fluency.) For the assessment of fluency we use two approaches. One is based on human perception, the other is based on temporal measurements. We took this two-fold approach to achieve a more comprehensive view. The following two sections describe the details of the two approaches.

Holistic rating of fluency

The first approach for evaluating the development of fluency relies on human perception. Although human perception is subjective and susceptible to subtle disturbing but irrelevant influences, it is an important source of evaluation. After all, language is usually produced for humans who perceive and judge fluency.

We employed three different raters and asked them to order each participant's samples according to the degree of fluency. This is the rating instruction, translated from German:

> "You have to rate how well the learners speak German. Can they express themselves clearly and within an appropriate time limit? How fluent and efficient are they? If possible, do not consider pronunciation and grammatical accuracy!"

Since we are aware that fluency, in a broader sense, refers to global proficiency and thereby includes pronunciation and accuracy (Lennon, 1990), we explicitly asked the raters to disregard these aspects. We do not expect the treatment to have an effect on pronunciation and we evaluate accuracy with the measures described above. Therefore the rating instruction targets the narrower sense of fluency and tries to reduce confounding effects as much as possible.

The rating procedure follows the *visual sort and rate* (VSR) method (Granqvist, 2003). In VSR, stimuli are visualized as movable icons. The rater can click on an icon to listen to the stimulus (as often as required) and then drag the icon on a two-dimensional pane to indicate a ranking. We only need one dimension in our context, since judgments are only based on one holistic criterion. At each step, raters were presented with the three samples of one participant (or only two, if the participant did not take part at the delayed posttest) and asked to build a rank order between those. The order of presentation was randomized, both over the participants as well as over the samples of each individual participant. The samples were bundled into sets of approximately 15 minutes total speaking time. In each session raters rated up to three of those sets (depending on their time constraints), which took them between 50 to 135 minutes. They were told to stop when noticing signs of fatigue. Raters were allowed to put different samples on the same rank, if they could not perceive a difference in fluency. In addition, raters were asked to comment on any difficulties they had.

We employed three raters with a background in teaching German as a foreign language (GFL) and with differing amount of experience in judging learners' performance. Two of them were experienced teachers, the third one was a fourth year stu-

dent of GFL with some minor rating experience. In order to estimate the consistency and reliability of each rater (i.e., intra-rater reliability), we let them re-rate a subset of participants.

Temporal measures of speech related to fluency

Considering the subjectivity of human ratings, we sought to complement them with presumably more objective measures. We therefore chose temporal measures of the participants' speech that have been shown to correlate with fluency perceptions. Kormos and Dénes (2004) examined correlations between human judgments of fluency and temporal properties of learners' speech. For learners of English as a second language, they found that the best predictors for the perception of fluency by humans were speech rate, the mean length of runs, the mean length of pauses, the phonation time ratio, and the number of stressed words per minute. In order to obtain those temporal measures we transcribed and annotated the speech samples. For the transcription we identified words and syllables. In addition, the annotation scheme included the phonation time, pause time and also *filled pauses*, which are non-lexical voiced fillers like for instance, "uh", "er", "mmh".

The annotation provided the following measures:

- mean length of runs
- mean length of pauses
- speech rate
- phonation-time ratio

The mean length of runs is calculated as the average number of syllables produced in utterances between pauses longer than 0.25 seconds (following Kormos and Dénes (2004)). Since it is not clear if filled pauses should be considered as part of a run, we calculated two measures for the mean length of runs: one includes syllables of filled pauses, the other excludes them.

The mean length of pauses refers to the average length of all pauses above 0.25 seconds. Even though Kormos and Dénes (2004) use a limit of 0.2 seconds for pauses, we chose the 0.25 seconds limit because it is more consistent with the criteria to calculate the mean length of runs.

For the speech rate, we distinguish three variants - one is the number of words per seconds, the other two relate to the number of syllables per second. Given the unclear status of filled pauses, as mentioned above, for one measure we count them as syllables, for the other we do not.

The phonation-time ratio refers to the percentage of time spent speaking given the time taken to produce the entire speech sample (i.e., speaking time divided by speaking time plus the rest of the time Towell et al. (1996)). Again, the status of filled pauses is not clear, Kormos and Dénes (2004), for instance, do not specify whether they count filled pauses as phonation time. Therefore, we calculated three variants of phonation

time ratios, differing as to how filled pauses are considered. The first measure disregards filled pauses altogether, the second counts them as phonation time, and the third does not consider them as speaking time.

The annotation did not include information about the stress of words because this feature is not straight forward to identify and also not corresponding clearly to a single acoustic parameter.

In summary, we examine development on the following nine measures:

- mean lengths
  - of pauses in seconds
  - of runs in number of syllables (including filled pauses)
  - of runs in number of syllables (excluding filled pauses)
- speech rate
  - syllables per second (including filled pauses)
  - syllables per second (excluding filled pauses)
  - words per second
- phonation-time ratio
  - disregarding filled pauses
  - counting filled pauses as phonation
  - counting filled pauses as silence

Figure 7.8 – Experiment procedure and timeline

7.4 Procedures

This section describes the details of the experimental procedures. We describe the setup for the data collection, the different data sets, and discuss problems related to participant dropout. In the end we provide details about the participants.

Setup

As we have discussed in Section 6.6.2, the testing and treatment sessions were integrated into lessons of intact classes. Over the course of the experiment, learners would come to the computer equipped classroom at our faculty to have their lesson. Figure 7.8 illustrates the timeline of the experiment. The complete experiment consisted of three sessions, spanning over six weeks, including two treatment sessions. As we have further explained above (page 129), we used a *repeated-measures design* including a pretest before the first treatment, one posttest after each of the two treatments, and a delayed posttest. The first two sessions took place in two successive weeks, each of them contained one treatment and some tests. The last session took place five weeks after, it only contained the delayed posttest.

Note that the tests for the oral skills were administered in different intervals than the tests that focused on the target structures. In particular, it was not possible to administer the oral tests directly after the treatment. Since the learners needed different amounts of time to work on the grammar tests and the treatment, they would not be ready at the same time afterwards to be matched up in pairs for the oral test. For this reason, each of the three sessions started with the oral test. Based on this constraint, the arrangement was as follows (cf. Figure 7.8): The first session started with pretests for oral skills and the target structures. It proceeded with the first treatment session and ended with the first posttest for the target structures. The second session started with a posttest for oral skills, it then provided the second treatment, and it ended with the second posttest for the target structures and a questionnaire collecting biographical data as well as usability assessments. The last session contained the delayed posttest for the oral skills and the target structures.

The interval for the delayed posttest was set at five weeks for practical reasons. We are not aware of any thorough discussion about suitable intervals for delayed posttests and most studies seem to choose the interval in an ad-hoc manner, usually driven by practical constraints of the experiment context. Five weeks however, seems to be well within the common range of delayed posttest intervals, as it was used for instance by Spada and Lightbown (1993). The courses met over a semester of about three months. For the first couple of weeks the students and teachers were supposed to get to know each other. This was also the time for students to switch courses if the initial assignment based on the placement test turned out to be inadequate. The last weeks of the course were dedicated to different exams. Considering additional holidays, this left a core of about 6 weeks available for this study.

Collected data

Figure 7.9 gives an overview about the different data sets we collected for each of the two target structures dative case in prepositional phrases (DatPP) and word order in subordinate clauses (SubC). It shows that for both structures we conducted treatments that implemented the FOCUS-ON-FORMS approach (Constrained) and the FOCUS-ON-FORM approach with recast feedback (Free-Recast). Only for DatPP we have an experiment group who was treated with metalinguistic feedback within the FOCUS-ON-FORM approach (Free-Metalinguistic Feedback). The reason for this limitation is that

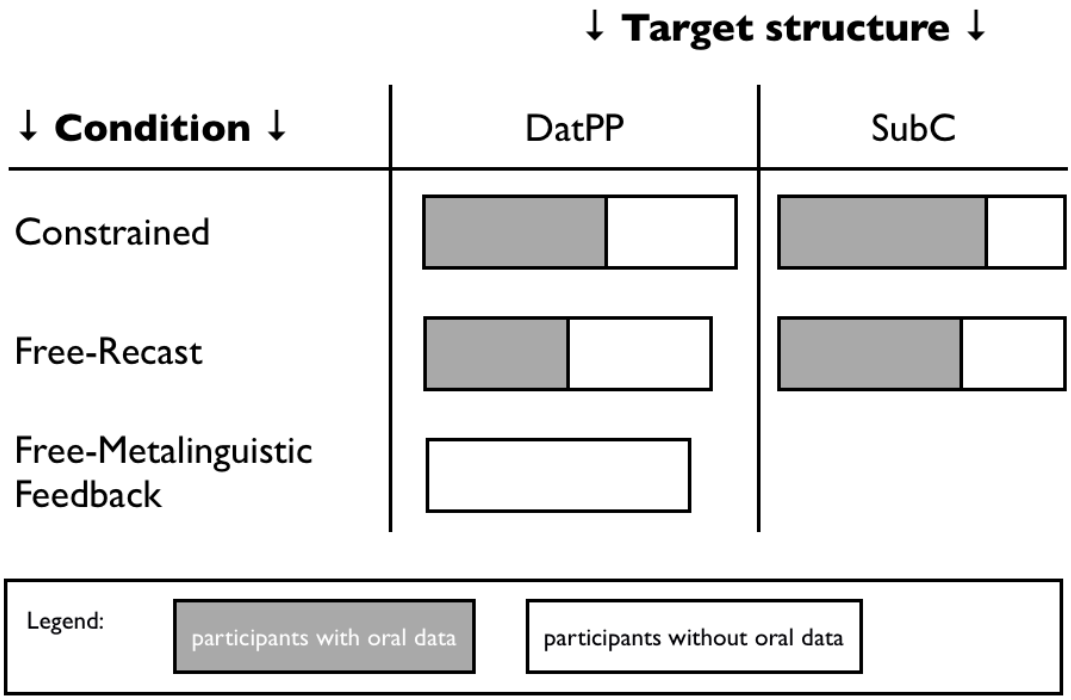


Figure 7.9 – Overview of experiment sample groups for each target structure indicating whether oral skill development was analyzed

we started the series of experiments by comparing only the constrained with the recast condition on the two target structures. Only later did we add the third condition (Free-Metaling). The decision to apply this condition only to the DatPP structure was driven by the scarcity of available participants. Given that we had so few participants we decided to focus on one structure only, in order to collect enough data to be able to draw meaningful conclusions.

Taking into account the considerable expense and effort of analyzing the oral communicative test data, we decided to only inspect part of the data. The shaded areas in Figure 7.9 indicate the set of participants for whom we collected data on their communicative skills. It shows that these data were only analyzed for a subset of participants, and in particular, not for the metalinguistic condition.

Table 7.2 gives a more detailed account of the number of participants for each condition and the time of data collection, sorted along the different collection periods. It also lists for each of the three sessions (s1, s2, s3) how many learners attended. We collected the data over the course of three semesters, using eight different courses in total. In the first semester (Dec 2009 / Jan 2010) we only collected data for the FOCUS-ON-FORM with recast feedback and the the FOCUS-ON-FORMS (constrained) condition. In the second semester we added the third condition – FOCUS-ON-FORM with metalinguistic feedback, applied to only one of the target structures – DatPP. Unfortunately, for the second semester (May/June 2010) we could not conduct the delayed posttest session within the course time. Therefore we asked participants to take part in the third session outside of their course time and paid them a compensation. However, as

	Dative			Subord. clause	
	Recast	Metaling	Constrained	Recast	Constrained
	s1 s2 s3	s1 s2 s3	s1 s2 s3	s1 s2 s3	s1 s2 s3
Dec 2009 / Jan 2010	11 8 6		10 10 7	10 10 8	9 9 8
May / June 2010	10 10 5	20 15 6	9 4 4		
Fall 2010		5 5 4	1 1 1	3 3 3	3 3 3
Total	22 18 11	25 20 10	20 15 12	13 13 11	12 12 11

Table 7.2 – Participant counts for each target structure and experimental group and time of data collection.

Table 7.2 illustrates, our recruitment showed only limited success. For instance, from the 15 students in the metalinguistic feedback condition who had taken part in the first two sessions, only six volunteered to take part at the delayed posttest session. Given that the amount of collected data was so limited, we decided to collect additional data during a third semester (Fall 2010). Since we had no access to class time at all for this semester, we recruited additional participants who were treated on an individual basis. They were also paid for their participation. The additional participants attended the same type of courses, so they were comparable to the first participants.

In general, it is evident that this study suffered from considerable participant mortality. Even for the first semester, when all three sessions were conducted within the course, the dropout rate between the first session and last session exceeded 50 percent in the most adverse cases. One cause for this was probably the fact that attendance at courses was not obligatory.

Participants

The majority of participants were foreign students enrolled at Saarland University, either on a temporary exchange for one or two semesters, or for the complete study. A minority were doctoral students, post-docs, or externals not related to the university. Participants had a varied first language background (18 different languages), but with a considerable majority of Spanish natives speakers (32 of 73). There were 3 Italian, 4 French, and 3 Romanian and 1 Catalan and 1 Portuguese, 8 English, 2 Czech, and 1 Bulgarian, 1 Russian, 1 Belarusian, 1 Georgian, 1 Turkish, 1 Arabic, 4 Chinese, 2 Korean, 1 Indonesian, and 1 Ewe native speaker. 3 participants failed to provide their native language. The average age of participants was 25.5 years (median 24), ranging from 19 to 46 years.

7.5 Summary

This chapter characterized the details of the experiment we conducted. Section 7.1 motivated the choice of the particular target structures based on (a) their suitability to be elicited in a meaning-based task, (b) the attested difficulty of their acquisition, and (c)

the feasibility to test them. We then went on to provide a detailed linguistic characterization of the selected structures dative prepositional phrases and subordinate clauses. This also included a discussion of the features that determine learnability.

Section 7.2 described the focused tasks that we designed to elicit the target structures. The task to elicit dative prepositional phrases is to give directions according to a given simplified map. Prepositional phrases are supposed to be used for referencing landmarks as anchor points for direction changes or subgoals. The task to elicit subordinate *weil*-clauses is to arrange an appointment, which requires to provide justifications for refusing a proposal. This section explained the interaction between learner and system and the particular system strategies to elicit the structures and provide feedback. For each of the two tasks we specified a dialog model that the system follows. For the constrained FOCUS-ON-FORMS conditions, which use a prescripted dialog as a context to prompt isolated forms, the prompts and feedback were described.

Section 7.3 presented the tests for measuring the development of language skills. We argued that for a comprehensive assessment both implicit and explicit knowledge need to be considered. We then motivated the use of the timed grammaticality judgment test as a measure for implicit knowledge, arguing that a time pressure prevents learners to access their metalinguistic knowledge and forces them to use their feel. We further reasoned that a sentence construction test is suitable to tap into explicit knowledge because it draws attention to linguistic forms and provides enough time to access explicit knowledge. For both tests, we presented the set of test items that we used. While the two tests target the command of the grammatical target forms, we also sought to evaluate the communicative skills, in particular because task-based FOCUS-ON-FORM instruction is claimed to promote the the real-time contextualized application of language skills. We described the collection of speech samples by engaging pairs of learners in a spoken dialog modeled after the tasks of the ICALL treatment. These samples are then analyzed with regard to the fluency that the participants exhibit, using two complementary measures. One is ask teachers to rate the perceived fluency, the other is to extract temporal measures from the speech sample.

Finally, Section 7.4 described the procedures to conduct the study and details of the data collection including a characterization of the contextual conditions and the timing.

Before we give a detailed account of the collected data and results in Chapter 9, we will describe the most important details about the design and implementation of the ICALL system that we used for the instructional treatment and provide a brief evaluation of the system, both in terms of its performance and user ratings.

8 The System

We have characterized the types of interaction that the system provides to the learners in Section 7.2. This chapter first describes the underlying mechanisms and the basics of their implementation in Section 8.1. It then reports on the performance of the system with regard to the instructional goals in Section 8.2. Finally, it provides an evaluation of the learners interacting with the system in Section 8.3.

8.1 Design and implementation

The system communicates with the learner in written mode, similar to instant messaging interfaces. In contrast to oral communication, the learner has access to a transcript of the interaction. The system takes the initiative by asking questions. The task is supplemented with additional graphical material, described above in Section 7.2. This material as well as an explanation of the task, is integrated into the graphical interface.[1]

The three different experimental conditions are realized by three different system variants have been implemented based on the same system architecture. In the description in this section, we concentrate on the components required for the free production activities; the constrained production activity is a simplified variant.

The system maintains a dialog with the learner by following the dialog strategies outlined above in Section 7.2. This involves interpreting the learner's input, responding to the learner by selecting a communicative goal according to the dialog model and the pedagogical strategy, and realizing the goal as a surface string. A particular requirement for the learning context is that the system has to recognize errors in the learner input and generate feedback on them.

Figure 8.1 shows the system's architecture: the modules (rectangle boxes), the resources they employ (boxes at the bottom tier with rounded corners), and the units of information that are passed between them (labels along the arrows). The graphical

[1] This description is based on the description given in Wilske and Wolska (2011)

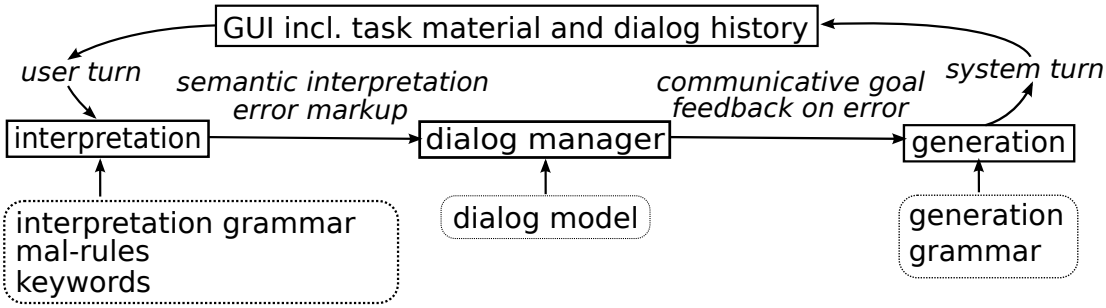


Figure 8.1 – The system architecture: Rectangles with solid lines indicate modules, rectangles with rounded corners and dashed lines at the bottom refer to resources, arrows indicate flow of information.

user interface (GUI) displays the utterances of the system and allows the learner to input their utterances. The productions of the learner are interpreted and passed to the dialog manager, which, based on the dialog model outputs a communicative goal, possibly including error feedback. The communicative goal is then generated and passed to the GUI. The information flow in this architecture is based on a pipeline model. The initiative is assigned to the system which commences the dialog by providing the first prompt. The only input modality is text for the learner, the sytem outputs text and provides static graphic material, which, however, does not change during the course of the dialog. The learner input is parsed as a whole as soon as the learner submits it be clicking the enter key or pressing the enter button.

In the following paragraphs we will explain in more detail the different processing steps.

GUI

The user interface is implemented as a web-based Java applet which runs independently of a particular operating system and browser. As we have showed above (Figure 7.1) this includes input area for the learner, the dialog history containing both the utterances of the learner and the system, and illustrative task material. For the constrained conditions, the interface contains a slot to fill the gap with a preposition (dative prepositional phrases, Figure 7.4) or fields to arrange words in a certain order (subordinate clauses Figure 7.7) and an area for explicit feedback about the performance.

The dialog model and manager

The dialog model represents the sequences of possible turn transitions, that is, the alternating turns produced by the learner and the system. It is implemented as a state machine using State Chart XML (SCXML) as an underlying representation (Barnett et al., 2012). We use the Java implementation of Apache SCXML[2]. The Apache framework also provides a dialog execution engine which receives input interpretations and

[2] `http://commons.apache.org/scxml`

triggers system responses according to the model. The actual SCXML dialog models are based on the conceptual models specified above in Section 7.2.

Across the range of possible approaches to dialog modeling, the ones based on finite state machines are the most basic (cf. Section 3.1.4, page 36). While more sophisticated modeling approaches have several advantages, the conditions of the experiment do not require these advantages. For example, frame-based modeling approaches allow for more flexibility and efficiency through their capability to simultaneously incorporate several pieces of information contained in the user utterance. However, in our context, we do not want the learners to take shortcuts, for instance, by providing the complete route description in one utterance. In fact, the dialog model was designed such that learners would be exposed to certain forms and encouraged to produce them repeatedly. Similarly for the appointments scenario, the goal was not, as it may be in a real dialog, to find the mutually available slot in as few turns as possible. On the contrary, the dialog was designed to have a minimal length to provide practice.

As another example of more advanced modeling paradigms, recall the information state update (ISU) approach (page 38). One of its important advantage is the flexibility afforded in handling general dialog phenomena. For any given context, these general phenomena comprise feedback moves like acknowledgment and recasts. However, we did not choose to implement this approach as the development of such models is more complex and with the two highly constrained scenarios that we chose, the additional effort did not seem to pay off. Nevertheless, for a possible extension to different scenarios and tasks, such a more general modeling such as this is likely to be worthwhile. Similarly, while the advantages of more sophisticated representations that can model beliefs and intentions (ISU and plan-based approaches), can be exploited for task-based ICALL dialogs, this was too expensive to build for the scope of the present study.

Interpretation of learner input

In general, interpreting the user input involves mapping a surface string of an utterance to a meaning representation. As typical in small-scale dialog systems, we implemented the system's language model (the set of linguistic expressions it covers) as a context free grammar with semantic tags. For parsing, we use the Java Speech API implementation of the CMU parser which is part of the Sphinx system.[3] The semantic tags encode two types of information: first, the symbolic meaning of utterances, and second, information on violations of grammatical constraints.

Given that the system is built for interaction with non-native speakers of German, a key requirement is to deal with non-target like input. The system comprises two complementary approaches for handling defective input. Recall the distinction between robust and sensitive approaches that we sketched in Section 2.3. On the one hand, a system can be built to ignore certain types of errors, on the other hand, for pedagogical purposes, it is desirable to diagnose errors. The system we implemented for the present study, is built to be robust towards orthographic errors and typos, but it needs to give feedback in response to grammatical errors that are part of the treatment.

[3] `http://cmusphinx.sourceforge.net`

Based on the two approaches, three specific error handling strategies are implemented and applied in a consecutive manner by the system. According to the robustness objective, the first processing step is to handle spelling or typing errors with a fuzzy matching approach for unknown words. Then, in the next step, the system builds on a set of anticipation-based mal-rules which are part of the interpretation grammar to detect and diagnose errors, following the sensitive approach. In a third step, deviant utterances that are not covered by the mal-rules, are interpreted based on extracted keywords, thereby adopting the robust approach. By utilizing mal-rules we apply the validity-based, language licensing diagnosis approach (see Section 2.3).

A possible alternative to this handcrafted and customized implementation is the use of freely available, more general parsers. Since standard, freely available parsers for German are built for native language and usually assume well-formedness, these were not suitable for our system. The MaltParser (Nivre et al., 2007), which has been successfully employed to parse learner errors robustly, does not provide information about errors (Ott and Ziai, 2010). Many approaches for parsing learner German and providing information on errors are only in prototype stage and/or not readily available (Reuer, 2003; Heift and Nicholson, 2001). An exception is the weighted constraint dependency grammar (WCDG) parser presented by Foth et al. (2004) which is robust but also suitable for error diagnosis through the information on constraint violations it provides. This may be a useful resource for future extensions with a wider range of task scenarios that require a larger coverage. However, for the scope of this study, we adopted a custom hand-crafted approach to save the additional costs of integration.

Fuzzy matching for unknown words

In order to ensure robustness with respect to typos and spelling errors the system first identifies unknown words in the input and tries to map them to known words by calculating the Levenshtein distance between the unknown word and known words (Levenshtein, 1966). Note that the set of known words come from the application-specific grammar and not from a general lexicon of German. For replacement with in-vocabulary candidates we consider those words which have a Levenshtein distance within a certain range to a known word, normalized by word length. In our implementation, we set the threshold for the distance at two, and one for words with a number of letters smaller than three. Replacement of unknown, supposedly mis-typed words yields one or more hypotheses which are then matched to the context free grammar.

Error diagnosis with mal-rules

Figure 8.2 presents a fragment of the interpretation grammar for prepositional phrases in the directions scenario, including mal-rules. The rule `<dir-change>` covers the utterance given in (1). If the prepositional phrase `<pp>` is not in the dative case, the semantic tag `non-dat` is returned, indicating that the dative case was required, but was not found. We encoded a set of mal-rules based on informal prior pretesting of the system with beginner learners.

(1) **L:** Gehen Sie vor das Cafe nach links.

```
<dir-change> = Gehen Sie <pp> nach (<left> | <right>)
<pp>         = <pp-DATIVE> {dat} | <pp-NODAT> {non-dat}
<pp-DATIVE>  = <prep> <np-dat>
<pp-NODAT>   = <prep> (<np-nom> | <np-gen> | <np-acc>)
```

Figure 8.2 – A simplified fragment of the interpretation grammar including a mal-rule; {non-dat} is the semantic tag indicating that a dative PP was not used where it was expected.

'Turn left, in front of [the]$_{nom/acc}$ coffee-shop'

The mal-rules approach was only implemented for the prepositional phrases in the directions scenario. Since the task for the subordinate clauses was only examined in the recast condition, the anticipation of errors was not essential. If a learner production could not be parsed with the regular grammar, we went on to the next step. This was possible because the dialog model prescribed the same response for correct and incorrect decline justifications – a recast in the second person (Section 7.2.2).

Keyword spotting

The drawback of the mal-rule approach is that it is usually impossible to anticipate all possible errors that might occur. For such cases, our system also implements a fall-back strategy based on keyword spotting: If no parse is found for an utterance, we create a semantic interpretation based on content words, using a keyword lexicon. The system generates a response utterance based on the interpretation of the recognized keywords, this utterance works as a recast for the learner's utterance. The implicit nature of a recast and the fact that it does not explicitly indicate that the learner's utterance was erroneous comes as an advantage here for cases in which the learner's utterance was actually correct but not covered by the grammar. This means that the keyword spotting strategy and the recast response is used for both input that is neither covered by the standard grammar nor by the mal-rules of the grammar. For the metalinguistic feedback condition this means however that the errors not covered by the mal-rules can only be treated with recast feedback.

Generation of system responses

The system output realization is performed using a template-based approach. The output is produced by generating a dialog move selected according to the dialog model using a context free generation grammar. The grammar associates atomic symbols representing communicative goals with sets of possible realizations. The generation templates contain slots that encode references to landmarks or directions (directions giving task) or activities on the agenda (appointments task) for confirmation moves. For generating metalinguistic feedback, the slots in the templates contain necessary information about the grammatical parameters referred to in the feedback. Slots in the templates are filled using feature-value pairs passed as arguments to the templates along with the communicative goals to be realized. Figure 8.3 provides a simplified fragment of the generation grammar that can realize recasts and metalinguistic feedback in the

```
<confirm-dirChangeWithLandmark> =
Okay, @makePP( -SLOT-prep-SLOT-, -SLOT-landmark-SLOT- )
nach -SLOT-dir-SLOT- [abbiegen].

<metaling-feedback-indicate-error> =
  (<ARTP> [IN <QUOTE-NP>] | DAS ) IST (FALSCH | NICHT RICHTIG);

<ARTP> = -SLOT-ARTICLE-SLOT-

<QUOTE-NP> = @QUOTENP(-SLOT-article-SLOT- -SLOT-errorLandmark-SLOT-)
```

Figure 8.3 – A simplified fragment of the generation grammar with templates for recasting and metalinguistic feedback; slot variables are surrounded by `-SLOT-`, `@QUOTENP`, `@makePP` are macros for generating specific sub-fragments of templates

directions giving scenario. For instance, the entry `<confirm-dirChangeWithLandmark>` realizes the recast in the example (3) in response to (2) (repeated from Example (20)) with the parameters `prep=vor, landmark=Cafe, dir=links` which are filled into the slots. The entry `<metaling-feedback-indicate-error>` is used to generate metalinguistic feedback as in (4) (repeated from Example (21) in Section 7.2.1).

(2) **L:** Gehen Sie vor **das** Cafe nach links.
'Turn left, in front of the coffee-shop'

(3) **S:** Okay, [vor **dem** Cafe nach links,]$_{RECAST}$
'Okay, left in front of the coffee-shop.'

(4) **S:** ['das' in 'das Cafe' ist nicht richtig.]$_{ML-FB}$
'das in 'das Cafe' is not correct.'

Constrained system for FOCUS-ON-FORMS

The system that implemented a constrained version of the task, provided the same material and the prompts were embedded in a dialog similar to the expected dialog of the free input conditions. However, instead of free input, the learner is presented an utterance that has to be completed by either filling a gap with a word or by arranging a set of words into the correct order.

As there is only one correct response, the system only has to compare the response value with the expectation. If the are identical, the system indicates to the learner that the response was correct and moves on to the next prompt that will be part of the evolving dialog until the dialog is completed. If the response does not match the expectation, the system informs the learner that their response was incorrect and lets them try two more times. After the third incorrect response, the system provides the target answer and moves on to the next prompt.

8.2 System performance

In this section, we describe the performance of the system during the interaction with the learners in the free input condition. According to the dialog model outlined above and the objective of the treatment, the system attempted to interpret the learner utterance, and if it recognized an error, responded with a corrective recast or a metalinguistic feedback respectively. Since the learner input was free and uncontrolled, it was expected that the dialog grammar and model could not anticipate each and every possible input. Therefore, a certain amount of non optimal responses from the system were expected. We will analyze the extent to which the system had to cope with unexpected and deviant input and how it reacted. There is no need to further analyze the system performance for the constrained condition, since the interaction was controlled and there was no room for a system failure. We will analyze the system performance for both target structures separately.

8.2.1 Dative prepositional phrases

For the target structure dative prepositional phrases, we recorded in total 75 interaction sessions, divided into 40 sessions of the recast condition and 35 sessions of the metalinguistic feedback condition. Due to technical failures, some of the interactions that had been taken place were not logged, as a result we have logs for 40 participants across both free input conditions, for some of which only one of the two sessions were recorded. In the 75 sessions, there were a total of 3127 utterances, divided between 1076 learner utterances and 2051 system utterances. The higher number of system utterances is due to the fact that the system always initiated and ended the dialog, but mostly because it would often produce two utterances at a time. For instance, the system would recast an erroneous learner utterance and then go on to elicit a continuation. For the learner, in general, this was not possible, since as soon as the learner submitted their production, the system would not accept any further input until it had produced a response.

In order to analyze the performance, we annotated for each system utterance, whether it is in accord with the dialog model as described above or whether it shows some sign of deficiency. In total, within the 2051 system utterances, there were 2076 instances of adequate system behavior and 306 problems. Note that some utterances contained more than just one type of success or failure. For instance, one system utterance may indicate the correct interpretation of a learner utterance and at the same time a successful (or failed) feedback move. Table 8.1 shows a more detailed breakdown of the adequate and the problematic system performance. The successful system responses can be analyzed as indicating a correct interpretation and a correct production. The correct productions can be further divided into adequate responses to erroneous or unexpected input on the one hand, and standard productions and responses to expected or correct learner input on the other hand.

As the table shows, a third of the successful system utterances were adequate interpretations of a learner utterance. About one sixth substituted to appropriate feedback given in response to erroneous or unexpected input of the learner. These include recasts, metalinguistic feedback and clarification requests, which we will further analyze

	count	percent
Successful performance	2076	100
Adequate interpretation of learner utterance	707	34.1
Adequate feedback in response to erroneous or unexpected input	338	16.3
Standard productions	1031	49.7
Problems	306	100
Failed recasts	43	14.1
Failed metalinguistic feedback	21	6.9
Inadequate interpretation	204	66.7
Inadequate productions	28	9.2
Time-outs	10	3.3

Table 8.1 – Breakdown of system performance for dative prepositional phrases in the directions giving scenario

below. The biggest part – one half of successful system utterances were standard productions like initiating and terminating the dialog, eliciting a learner continuation or a correction after previous feedback. These were not directly dependent on the correct interpretation of the preceding learner utterance.

The system failures comprise failed recasts (14%) and failed metalinguistic feedback (7%), as well as other, more general instances of inadequate interpretations (67%) and inadequate productions (9%). A final class of failures comprises the instances in which the system took more than 20 seconds to respond, which made the learner attempt another production in the meantime or start the exercise all over again (3%). The long response time was based on a bug with the interpretation of very long and complex utterances with certain characteristics that had not been discovered in testing before.

The higher number of failed recasts compared to metalinguistic feedback is related to the fact that, overall, the system attempted to produce more recasts than metalinguistic feedback. This is because metalinguistic feedback was produced only in response to a clearly erroneous learner utterance, while recasts were also used in response to learner utterances that were not covered by the interpretation grammar including the mal-rules for error-recognition. This means that the system would produce recasts as a fallback reaction to deviant learner input also in the metalinguistic feedback condition. Furthermore, recasts were also produced in both free input experiment conditions in confirmation moves of apparently well-formed utterances.

8.2.2 Recasts for dative prepositional phrases

We will take a closer look at the productions of recasts now and analyze the types of successful recasts and reasons for failed recasts. Table 8.2 indicates the counts of different recasts and failed recasts. Overall, there were 497 successful recasts and 43 instances of failed recasts. About a third of the successful recasts were produced by the

	count	percent
Successful recasts	497	100
in response to dative prepositional phrase errors	145	29.2
in response to avoidance of dative pp constructions	84	16.9
as repetition of correct input in acknowledgment	187	37.6
as confirmative repetition after metalinguistic feedback	33	6.6
in response to other errors	48	9.7
Failed recasts	43	100
Error in target description was not recast	12	27.9
Erroneous recast for error in target description	16	37.2
Erroneous recast for PP with preposition *zwischen*	12	27.9
Others	3	7.0

Table 8.2 – Distribution of successful and failed recasts

system in response to a learner error in dative prepositional phrases, which was the focus of the treatment. Nearly 17 percent were reactions to the learner avoiding the production of dative prepositional phrases. The biggest proportion of recasts in the system production, 38 percent, were repetitions or reformulations of correct learner utterances which served as acknowledging grounding moves. Another seven percent of recasts were confirmative repetitions after a learner corrected their production in reaction to metalinguistic feedback. Apart from that, 10 percent of recasts came in response to other deviances of the learner production which where not in focus of the treatment. Although the system was not specifically programmed to correct other errors, these errors were recast as a side effect of the policy to use a recast in response to deviant utterances that contained a dative prepositional phrase.

We observed a total of 43 instances in which corrective recasts failed. The main source of problems came from insufficient analysis of learner utterances that referred to the target landmark. According to the dialog model (cf. Section 7.2.1 and Figure 7.3), the system interpreted a reference to a target landmark as a signal to finish the dialog, but only if there was some previous utterance with at least one change of direction before. The part of the interpretation grammar that covered the target describing utterances and in particular the use of dative prepositional phrases therein was designed to be less rigorous than the part of the grammar that interprets a direction change with a landmark reference, which we saw as the main source of dative prepositional phrases. Since we did not expect the use of of dative prepositional phrases when referring to the target landmark and therefore did not enforce such use, the grammar ignored many of the dative errors that occurred there. This resulted in 12 instances of learners' target descriptions that contained erroneous dative prepositional phrases which were not corrected in a recast at all and 16 instances of erroneous recasts. The erroneous recasts failed to adequately reproduce the relation of the target landmark to the anchor landmark produced by the learners. They either referred to a wrong landmark or they confounded the relation.

	count	percent
Successful Metalinguistic Feedback	66	100
Incorrect article in dative PP	41	61.2
- Initial	*29*	*43.3*
- Subsequent after failed trial	*12*	*17.9*
Incorrect contraction with *zu* or *bei*	7	10.4
Missing article in dative PP	2	3.0
Preposition *bis* used without *zu*	16	23.9
Failed Metalinguistic Feedback	21	100
Incorrect article in dative PP	11	52.4
Missing article in dative PP	5	23.8
Contraction with *zu* plus superfluous article	2	9.5
Follow-up after a successful learner correction	3	14.3

Table 8.3 – Distribution of successful and failed metalinguistic feedback

Another relevant source of errors in recasts was the insufficient modeling of the preposition *zwischen* ('between') in the realization of recast – the realization grammar did not cover that *zwischen* governs too arguments. This insufficiency was repaired after it became evident and therefore it only concerned the first round of experiments in 12 instances. Finally, there were three instances of failed recasts that arose from different interpretation problems.

8.2.3 Metalinguistic feedback for dative prepositional phrases

Table 8.3 shows the distribution of different types of successful and failed metalinguistic feedback that occurred during the treatment. In total, there were 66 instances of successful metalinguistic feedback in the 35 sessions of metalinguistic feedback treatment. At the same time there were 21 instances of failed metalinguistic feedback.

The biggest part of the successful metalinguistic feedback related to the primary objective of the treatment, incorrect articles in dative prepositional phrases, which made up 41 instances. Of these, about three quarters were initial feedback right after the learner error and about one quarter was subsequent feedback when the learner's attempt to correct their error in response to previous feedback was still erroneous. Feedback was also given seven times in response to errors regarding contractions with the preposition *zu* ('to') and *bei* ('at'). In two cases, the system had to complain about a missing article. Finally, there were 16 instances in which the preposition *bis* 'till,to' was erroneously used without the preposition *zu* 'to'. This was an error that was not in focus of the instruction, but pervasive enough to require feedback.

The 21 instances of failed feedback are of four different types. In 11 cases, the system failed to recognize an incorrect article; in five cases it did not complain about a missing article. In two instances, it tolerated a contraction with *zu* with a superfluous article. In three cases, the system did not react appropriately towards a valid learner

	count	percent
Successful performance	890	100
Adequate interpretation of learner utterance	282	31.1
Adequate feedback in response to erroneous or unexpected input	187	20.6
Standard productions	437	48.2
Problems	87	100
Inadequate interpretation	81	93.1
- resulting in missed recasts	*4*	*4.6*
Time-outs	4	4.6
Inadequate Inferences	2	2.3

Table 8.4 – Breakdown of system performance for subordinate clauses, appointments scenario

correction. All failures were based on insufficiencies in the interpretation grammar – the deviances in the learner utterances were not recognized. While recasts can be given in response to any deviance, metalinguistic feedback should only be given in response to a clear error – therefore it is necessary to have an error grammar as broad as possible in order to recognize as much errors as possible. Here we meet the limits of anticipation-based error recognition – it is hard to predict all possible errors and usually the error grammar can be broadened in iterative development steps by collecting more learner data. For the current study, more extensive pretesting with learners might have decreased the failure rate further.

8.2.4 Subordinate clauses

For the target structure subordinate clauses in the appointments scenario, we have collected 26 session logs of 15 different participants. These comprised 1278 utterances in total, with 855 system utterances and 381 learner utterances. Similar to the interactions in the directions scenario, there are more system utterances because the system would sometimes produce two utterances right after one another.

Table 8.4 shows the distribution of successful and failed system performance. There were 890 utterances in which the system performed adequately in accordance with the dialog model opposed to 87 problems. Of the adequate system utterances, about a third indicated an adequate interpretation of the learner production. One fifth were adequate feedback in response to erroneous or unexpected learner input. Nearly half were standard productions, that were not directly dependent on the correct interpretation of the preceding learner utterance, like, for instance, initiating and finishing the dialog, proposing time slots, or eliciting learner proposals.

Of the 87 failures, the biggest part (81 instances) related to problems in inadequate interpretation. Of these interpretation failures, four led to a missed recast in response to an erroneous learner attempt to produce a *weil*-clause. The remainder of the failures consist of four instances in which the system did not respond in under a minute, which

made the learner submit another utterance. The long response time was based on a bug with the interpretation of very long and complex utterances with certain characteristics that had not been discovered in testing before. In two instances, the system failed to draw more complex inferences from previous learner utterances and proposed time slots that the learner had indicated as impossible before. The reason for that failure was that dialog model did not include a memory state to keep track of all constraints that the learner had expressed, but given that this only became an issue in two cases, this shortcoming is not very severe.

proposal	36
justification	23
decline	13
accept	9

Table 8.5 – Breakdown of interpretation failures

Table 8.5 gives a more detailed analysis over the interpretation failures. In 36 instances, the learner's proposal for a time slot was not correctly interpreted, in about a half of these cases, the system failed to arrive at any interpretation, in the other half of cases, its interpretation missed some of the details of the proposal. The second most frequent interpretation problem concerned the justifications given by the learner. Of the 23 instances, fortunately, only four resulted in a missed corrective recast. In the other instances, a recast was not necessary as the learner had not produced an erroneous *weil*-clause. In 13 instances, the system did not understand that the learner declined a proposal. Finally, in 9 instances, the system did not understand that the learner accepted a proposal.

opportunities to use weil-clause	212	recasts	185
avoided weil-clause	161	providing recast	143
correct weil-clause	41	repetitive recast	36
incorrect weil-clause	10	corrective recast	6

Table 8.6 – Learners' performance on weil-clauses and distribution of system recasts

Table 8.6 shows the distribution of recast types along with the learner behavior that triggered them. When the learners give a justification for their refusal of a system proposal, they can avoid a *weil*-clause altogether, or, if they use a *weil*-clause it can be correct or erroneous. From the 212 opportunities to phrase a justification as a *weil*-clause, the learners avoided them in 161 cases. There were 41 instances of *weil*-clauses that were correct with regard to the position of the finite verb, and ten instances of erroneous *weil*-clauses in which the finite verb was not placed in the end. In the case of avoidance, the system provided a *weil*-clause as a recast of the justification. In case of correct *weil*-clauses, the system provided an acknowledging grounding more in form of a repetitive recast which rephrased the justification into the second person

and sometimes changed the wording a bit. For the case of incorrect attempts at *weil*-clauses, the recast provided a correct example based on the reason given by the learner. The corrective recast was also rephrased as the second person.

As the table shows, there is a difference between the numbers of each type of learner utterance and the numbers of the particular recast version provided by the system – the difference is based on the failed interpretations of the system.

Compared to the directions giving scenario, it is obvious that the number of incorrect attempts at *weil*-clauses is much lower than the incorrect attempts to produce a dative prepositional phrase. This shows that some grammatical structures are easier to elicit than others in tasks-based interaction scenarios. Because the learners produced relatively few *weil*-clauses, there is much less opportunity for the system to provide corrective feedback. The system however, can still provide lots of examples by incorporating *weil*-clauses within its own productions.

In summary, our performance analysis showed that the system is suitable for the objectives we pursued with our instruction. However, the failure rate suggests that the extent of pre-experimental pilot-testing proved was insufficient. An optimization is required for eventual further experiments. In the remainder of this chapter we will examine how the strengths and shortcomings of the system were reflected in learner appraisals.

8.3 Learners' perception and rating of the system

In this section we briefly present the subjective impressions of the learners working with the system. Learners were asked to fill out a questionnaire after the second session. In it, we asked for biographical data and queried how they perceived the interaction with the system and the system itself. We report results on the following questions:

1. Did you enjoy the interaction with the system?
2. Was the interactive task useful for you?
3. Would you like to use such a system more often to practice German or other foreign languages?
4. How natural did you perceive the dialog with the system?
5. Do you think the system's utterances were coherent and appropriate in the given dialog context?
6. Do you think that the system understood you?

The first three questions were given to all learners, the latter three only to those learners who had interacted in a condition which allowed them to formulate their own input, i.e., free-recast and free-metalinguistic feedback, but not the constrained condition. Responses were allowed on a 4-point Likert scale without a neutral answer, but there was always an option to choose "I don't know".

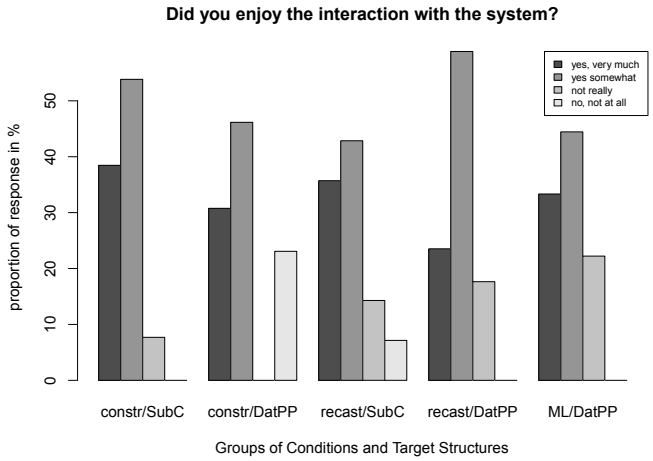


Figure 8.4 – Ratings for enjoyment of system interaction

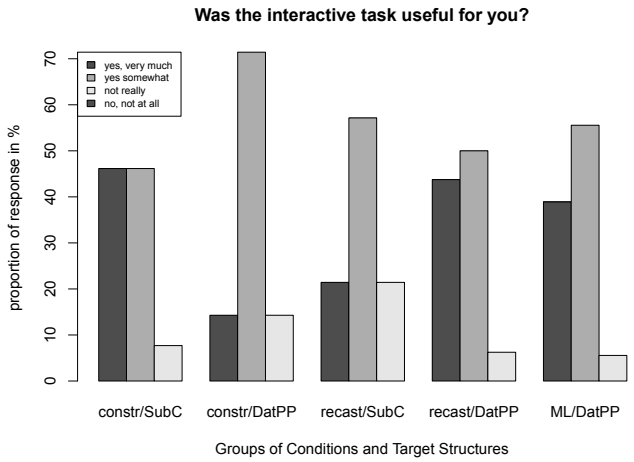


Figure 8.5 – Ratings for perceived usefulness

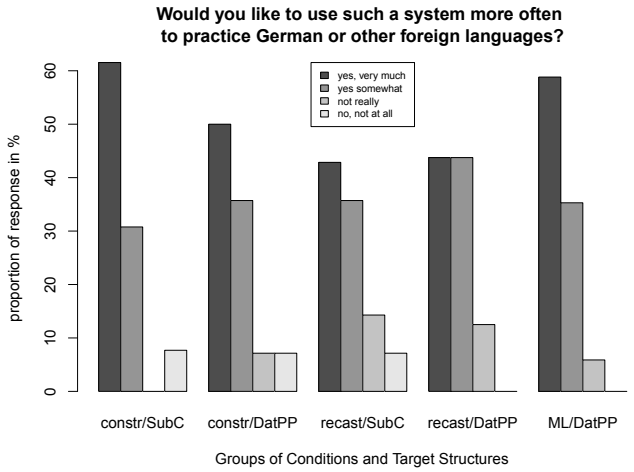


Figure 8.6 – Ratings for likelihood of future usage

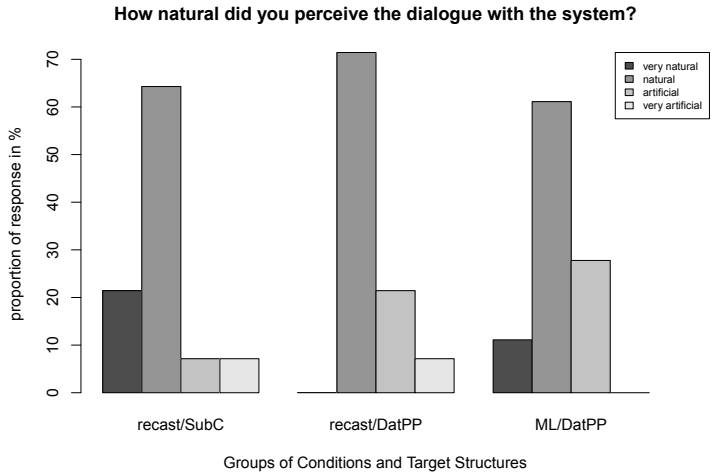


Figure 8.7 – Ratings for naturalness of the system interaction

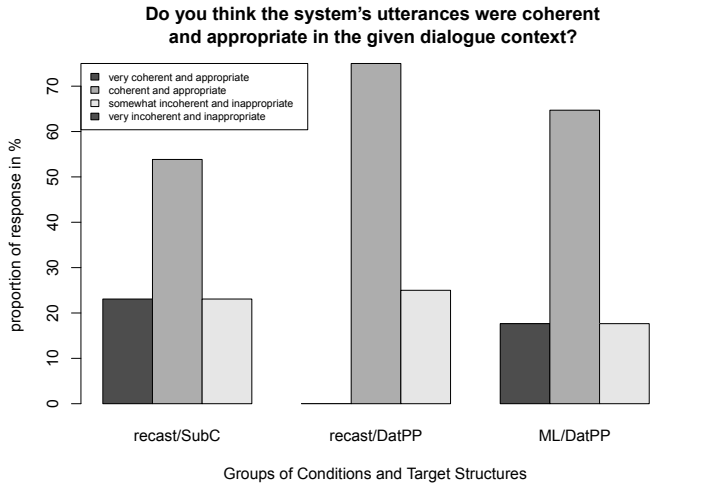


Figure 8.8 – Ratings for coherence and appropriateness of system's utterances

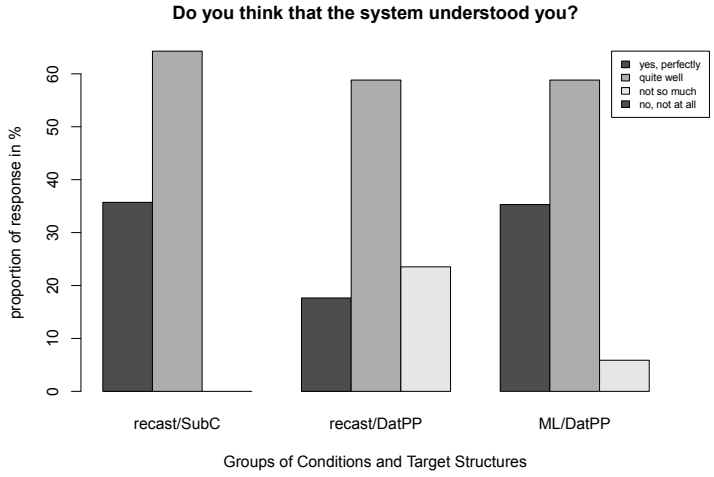


Figure 8.9 – Ratings for comprehension skills of system

Condition	Constrained		Recast		Metaling FB
Structure	SubC	DatPP	SubC	DatPP	DatPP
Question↓ / n→	13	15	14	17	18
1. enjoy	1.69	2.15^{++}	1.93	1.94	1.89
2. useful	1.62	2.00^{+}	2.00	1.62^{+}	1.67
3. future use	1.54	1.71^{+}	1.86	1.69^{+}	1.47^{+}
4. natural			2.00	2.36^{+++}	2.17
5. coherent			2.00^{+}	2.25^{+}	2.00^{+}
6. understanding			1.64	2.06	1.71^{+}

Table 8.7 – Questionnaire results, the number of + indicate the number of "don't know" - responses

The bar plots depicted in Figure 8.4 to 8.9 show the distribution of responses for each question, ordered along the different experimental groups. Table 8.7 summarizes the data for all questions. It indicates the mean value of the response, on a scale from 1 to 4, where 1 is the value of the most positive response and 4 the value of the most negative response. It further indicates the number of participants in each condition for whom responses could be gathered. There was some data missing due to technical problems during the collection process.

A quick look at the first three bar plots (Figure 8.4, 8.5, 8.6) shows that all five condition-structure combinations have very similar results. For the first question, *"Did you enjoy the interaction with the system?"*, it is remarkable that the constrained condition for SubC obtains the least negative replies, and consequently, also the highest average rate for that question (1.69). It is further noticeable that the constrained condition for DatPP and the recast condition for SubC have the most negative replies and are also the only two groups which obtained any of the most negative response (*"no, not at all"*). The constrained condition for DatPP receives the lowest average rating (2.15).

For the second question *"Was the interactive task useful for you?"*, the results seem to split between a low average rating of 2.0 for the constrained DatPP and the recast SubC on the one hand and the rest. On the bar plots, the two low groups receive the least most positive replies.

For the third question *"Would you like to use such a system more often to practice German or other foreign languages?"*, the recast SubC group receives the lowest average rating, while the metalinguistic DatPP group receives the highest rating.

Note however that none of the differences is significant, according to a Kruskal-Wallis one-way analysis of variance test.

Regarding the latter three questions which are only applicable for the free input conditions, (Figure 8.7, 8.8, 8.9), we observe the following: The learners of the recast SubC group gave the highest average rating for the question *"How natural did you perceive the dialog with the system?"*, while learners of the recast DatPP gave the worst rating. For the question *"Do you think the utterances of the system were coherent and appropriate in the given dialog context?"* both the recast SubC and metalinguistic DatPP group give an

average of 2.0. The recast DatPP group found the system slightly less coherent, with an average rating of 2.25 and none giving the highest rating. Consistent with that result, the recast DatPP group also gave the lowest average rating regarding the question *"Do you think that the system understood you?"* – 2.06. The recast SubC group gives an average rating of 1.64 (based on ratings no worse than 2) and the metalinguistic DatPP group gives an average rating of 1.71. Again, none of the differences are significant according to according to a Kruskal-Wallis one-way analysis of variance test.

If we summarize the data and compare the free input conditions with the constrained input condition across the two target structures, there is no significant difference on any of the three questions. We do not make a three-way comparison between constrained, recast and metalinguistic feedback conditions, because the latter was only applied to one of the target structures. In general the ratings of the system is more positive than negative but clearly with room for improvement.

8.4 Summary

This chapter presented background details about the implementation of the dialog system that we used for conducting the study. In particular, it described the strategies to handle learner errors. It then provided an analysis of the system performance and identified the most important failure points. In the end, we gave a summary of the user ratings which showed an acceptable, but not outstanding impression. In the following chapter we will present in detail the results of the language skill assessment before and after the treatment.

9 Findings

This chapter describes the results of the experiments that we conducted to answer the research questions put forward in Section 6.6. As we have described there, we wish to investigate (a) if there is a difference between the effects of computer-based FOCUS-ON-FORM and FOCUS-ON-FORMS instruction and (b) if there is a difference between recasts and metalinguistic feedback. We will present our findings along the dimensions of linguistic development that we described in Section 7.3. In the first part, we will present the development on grammatical accuracy, as measured by a grammaticality judgment test and a sentence construction test (Section 9.1). In the second part, we will then present the development of spoken language skills, firstly, as rated by human raters, and secondly, in terms of temporal measures of speech (Section 9.2).

It is important to note that the circumstances of our study, in particular the small amount of potential participants and their high drop-out rate, can have a negative impact on the statistical power of our tests. The power of a statistical test describes how likely it is that the test will detect a correlation if correlation exists in reality. The power is determined by three variables - the significance level, the effect size, and the sample size. The common significance level in psychological and educational studies is 0.05 - it indicates that the probability a result has occurred by chance is 5%, with a probability of 95% the result has not arisen by chance. The effect size indicates how strong a correlation is, for instance, how much more learners learn with a given method compared to learners that used a different method. In general, a test is more powerful, i.e., more likely to detect an existing correlation, if the size of effect is large and/or the sample size is large and/or the significance level is high. Since researchers cannot manipulate the effect size, and since in the circumstances of second language research the number of potential participants (i.e., the sample size) is usually restricted, it has been proposed to also report results that fail the strict, yet somewhat arbitrary level of 0.05 (Gass et al., 1999; Mackey and Gass, 2005). Gass et al. argue that reporting and discussing these results as ”meaningful trends” should be encouraged, because

they could be as important as more strictly significant results. Following this argument we will report results that are significant at the conventional level of $\alpha = 0.05$ and mark differences that were significant at $\alpha = 0.10$ ("marginally significant") to indicate interesting tendencies.

9.1 Development of grammatical accuracy

This section presents the development of learners as assessed by tests that focus on the target structures in a relatively isolated fashion. As we have described above in Section 7.3, we use a grammaticality judgment test that is directed at implicit knowledge (7.3.1) and a sentence construction test that is directed at more explicit knowledge (7.3.2). For the grammaticality judgment test we will look in more detail at the performance for grammatical and ungrammatical items separately, because it has previously been shown that learners perform differently on grammatical and ungrammatical items (Hedgcock, 1993; Loewen, 2009).

For each of these two tests and for each of the two target structures that we examined, we will first discuss the data under a descriptive perspective and in a second step apply statistical tests in order to infer if any of the observed variance is significant. The key objective of the statistical inference is to examine if the performance of the participants is different between different test times and if the different experimental groups show a different development.

In order to determine the appropriate statistical tests, we first checked if the collected data are normally distributed, since this is the standard criterion to decide between parametric and non-parametric tests. We performed the Shapiro-Wilk test (Shapiro and Wilk, 1965) on our data and found that the data was not normally distributed on some of the within-subject and/or between-subject variables. Similarly, the Levene test (Levene, 1960) revealed that the assumption of homogeneity of variance was violated; this assumption is another important criterion to check if two or more independent groups can be compared. Based on these test results, we used non-parametric versions of all tests. In order to compare within-subject differences between the different test times, we performed Friedman's Test (Friedman, 1937) with pairwise post-hoc comparisons using the Wilcoxon-Nemenyi-McDonald-Thompson test as described by Hollander and Wolfe (1999) based on the implementation in the package coin in R[1] and the implementation by described in Galili (2010).

For comparing the differences between groups (between-subject) we used the Mann-Whitney U test (Mann and Whitney, 1947) – for those situations where we only had two groups to compare. For the experiments with three different groups to compare, we used the Kruskal-Wallis one-way analysis of variance test, with multiple comparison tests as a post-hoc analysis based on the test described in Siegel and Castellan (1988) implemented in the pgirmess R package.[2]

Because the data are not normally distributed, we describe them by their median and interquartile range and use box plots to illustrate further characteristics. In particular, the box plots indicate the dispersion and outliers. The upper and lower edges

[1] http://cran.r-project.org/web/packages/coin/index.html

[2] http://cran.r-project.org/web/packages/pgirmess/index.html

of the boxes indicate the upper and lower quartile respectively, i.e., 25% of the data points are above the upper quartile and 25% of the data points are below the lower quartile. The area within the box, i.e., the distance between the upper and lower quartile is the interquartile range. The whiskers of the plots indicate the most extreme point which is still within 1.5 times the interquartile range from the upper or lower quartile respectively.

In general, our sample sizes for participants who have provided data over the complete course of the experiment are relatively small due to the relatively small number of available participants and the considerable drop-out rate as described above (Section 7.4). As we have noted above, the *power* of a statistical test, i.e., the probability of detecting an effect in the data when there is one in reality, is dependent on the sample size.

Therefore, we tried to compensate the sparseness of data of participants who provided data for the complete span of the experiments, by also taking into account the data from those participants who dropped out at later stages of the study. This means that we considered the data coming from all participants who took part only in the first two or three tests respectively, in addition to those who provided data for all four tests. As a matter of course, the value of considering these additional data is limited to analyzing the more immediate effect of the instruction only. We only present the additional analyses for the dative prepositional phrases, since the drop-out rate for that structure was more severe; for the subordinate clauses, only three participants did not provide data for the last test, to consider their data did not add new results.

We will first describe the results of the target structure word order in subordinate clauses (SubC) (Section 9.1.1) and then the results for the dative case in prepositional cases (DatPP) (Section 9.1.2).

9.1.1 Word order in subordinate clauses

This section presents the learner development for the target structure word order in subordinate clauses. Recall from Section 7.4, that for this structure, we only compared two experiment conditions - the FOCUS-ON-FORM approach with free input and recast feedback and the FOCUS-ON-FORMS approach with constrained input, which we will further refer to as **Free-recast** and **Constrained**.

Table 9.1 contains information about the test result data for both experimental groups on both tests: sentence construction (SC) and timed grammaticality judgment test (TGJT), as described by medians (md), and interquartile ranges (iqr), for each of the four test times. For the judgment test, we present the total of all items as well as the scores for grammatical and ungrammatical items separately. The table further indicates the number of participants whose data was accessible in each group. As we indicated in Table 7.9 in Section 7.9, we have data from 11 participants for each group who took part in all three sessions. Of these, we excluded some participants' data from the analysis because they started with a perfect score at the pretest T1 – since we are interested in the learning gain that the instruction yields, we only consider learners who have the possibility to improve. We excluded the perfect performers separately for each test and each subset of test items. As Table 9.1 shows, this left us with 10 learners in each group for the grammaticality judgment test (on the complete item set

and the grammatical items, excluding one perfect scorer from each group), and with 6 learners in each group for the sentence construction test and the ungrammatical items of the judgment test, excluding five perfect scorers from each group.

The numbers contained in Table 9.1 are presented graphically by the box plots depicted in Figure 9.1 to Figure 9.4 which we will discuss further below.

		T1		T2		T3		T4	
Group/test	n	md	iqr	md	iqr	md	iqr	md	iqr
Free-recast									
SC all	6	67	25	75	16	75	42	100	13
TGJT all	10	79	32	83	15	79	25	96	8
TGJT gram.	10	83	12	92	17	92	33	92	17
TGJT ungr.	6	58	42	75	29	75	54	100	0
Constrained									
SC all	6	58	17	83	25	75	29	67	38
TGJT all	10	79	36	83	34	84	17	83	17
TGJT gram.	10	83	12	92	33	83	17	83	13
TGJT ungr.	6	50	58	83	50	84	58	75	54

Table 9.1 – Test results for SubC: medians (md) and interquartile ranges (iqr) for percentage scores, sentence construction test (SC) and timed grammaticality judgment test (TGJT).

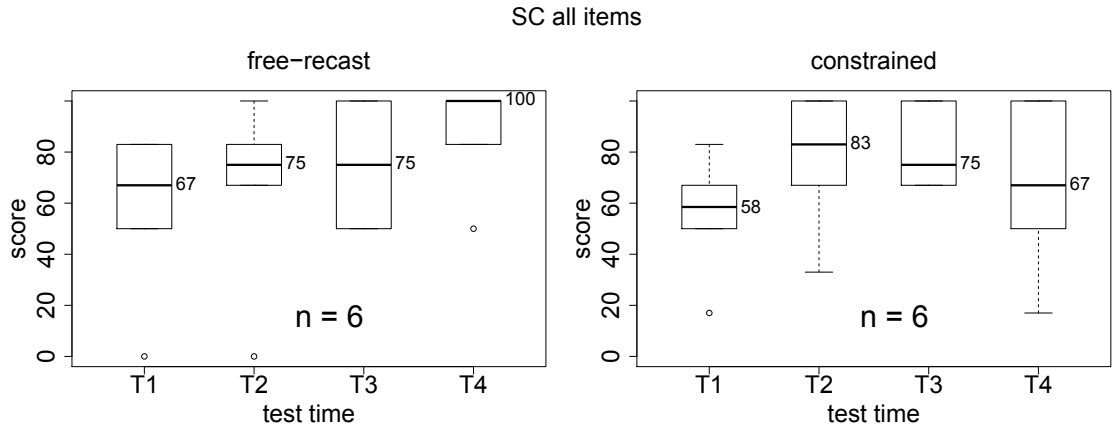


Figure 9.1 – Box plot representation of results for the sentence construction test for SubC, all items.

Sentence construction test

The results of the sentence construction test, which comprised six items, are illustrated in Figure 9.1. For both groups, the median increases between the pretest T1 and the first posttest T2, but for the recast group the dispersion decreases, while for the constrained group it increases. After that, the development is different – while the free-recast group maintains the same median at the second posttest T3 (albeit with a larger

dispersion) and finally reaches a perfect median of 100% at the delayed posttest T4, the constrained group steadily decreases after T2. According to the Friedman test, however, none of the differences between the test times is significant. Similarly, none of the differences between the two groups at any test time is significant either, according to the Mann-Whitney U test.

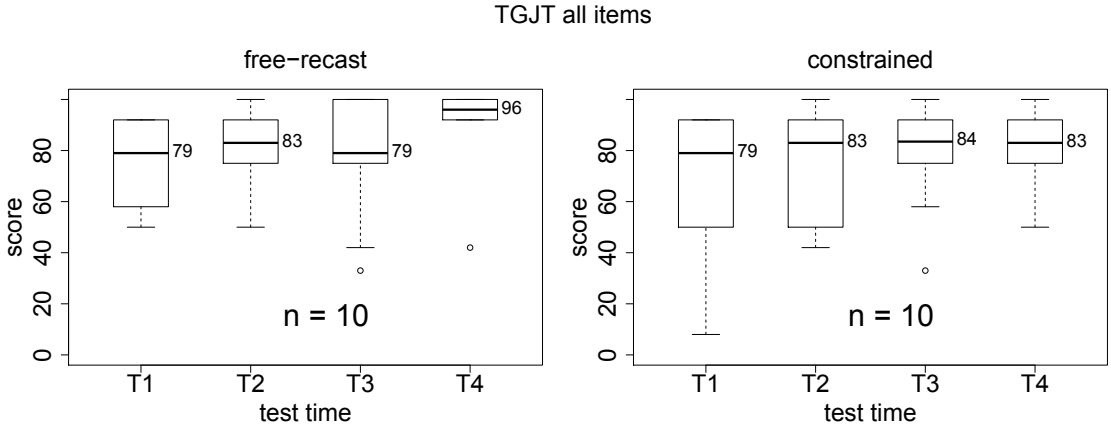


Figure 9.2 – Box plot representation of results for the judgment test for SubC, all items.

Timed grammaticality judgment test

The results of the judgment test are illustrated in Figure 9.2. This test comprised 12 items. Both groups have a very similar development – they start off from a median of 79% at T1 and very slightly increase to 83% at T2, with the recast group displaying a smaller dispersion. The free-recast group then gets back to 79% at T3 and finishes with a median score of 96% and a very small dispersion, while the constrained group shows increases only minimally to 84% at T3 and gets back to 83% at T4. Except for the performance of the free-recast group at T4, the scores seem all very similar and do not seem to hold any significant differences. And indeed, testing for differences between test times and between groups only reveals differences related to that result. There is a marginally significant difference between T1 and T4 for the free-recast group. Regarding between-group differences, the performance of the free-recast group at T4 is significantly better than the score of the constrained group.

Grammatical and ungrammatical items For the grammatical items, as illustrated in Figure 9.3, again both groups have a very similar development – they start from the same median of 83% at T1 and slightly increase to 92% at T2. The free-recast group maintains its median for T3 and T4, albeit with a lower lower quartile at T3, which indicates a slightly lower performance. The constrained group gets back to a median of 83% at T3 and T4. None of the between-test and between-group differences are significant.

For the ungrammatical items (see Figure 9.4), both groups increase their median score between T1 and T2 and then keep it at T3, with the constrained group starting slightly lower but getting slightly higher than the free-recast group. At T4, the free-recast group improves up to the perfect median score of 100%, while the constrained

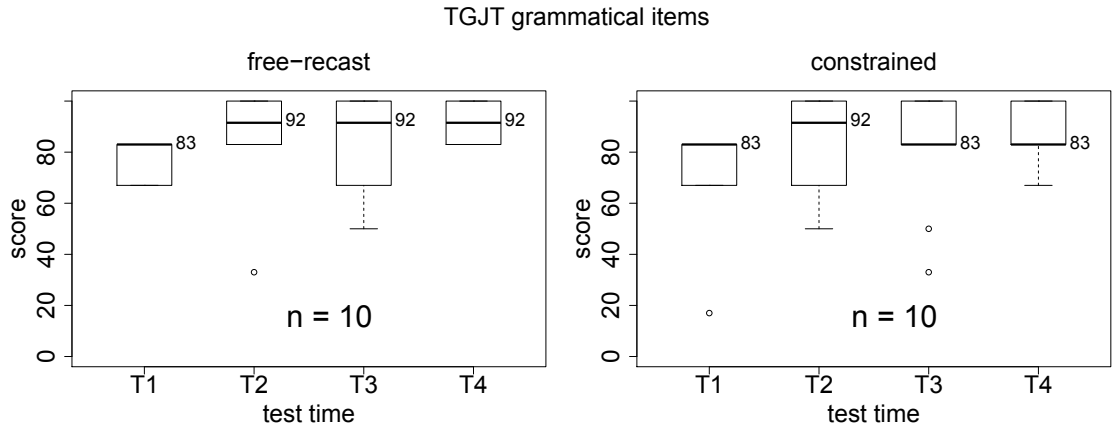


Figure 9.3 – Box plot representation of results for the judgment test for SubC, grammatical items only.

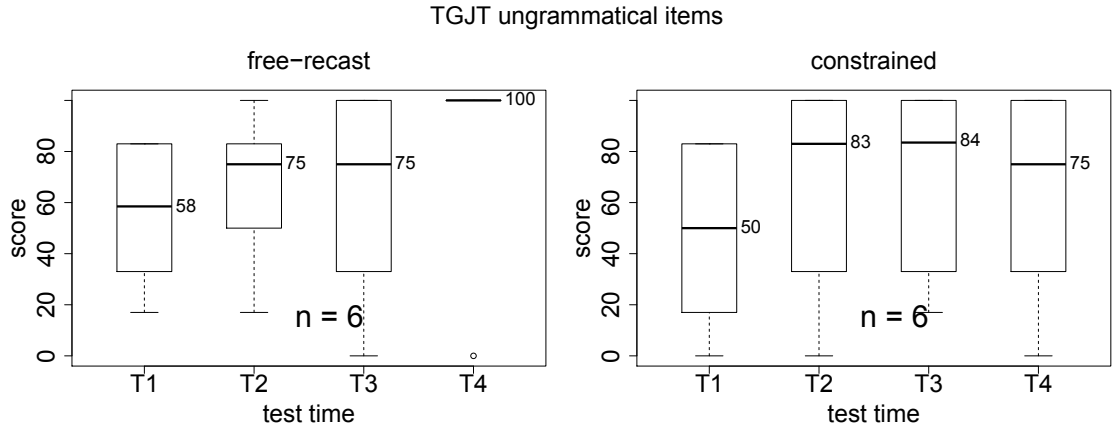


Figure 9.4 – Box plot representation of results for the judgment test for SubC, ungrammatical items only.

group decreases a bit to 75%. None of the differences between the groups at each test time is significant. However, the increase of the constrained group between T1 and T2, as well as their increase between T1 and T3 is significant.

Previous studies have shown that the performance can differ between ungrammatical and grammatical test items, and often, ungrammatical items are harder to judge correctly (Hedgcock, 1993). In our data, such a difference is not very distinct. In fact, according to the Wilcoxon signed rank test, we find such significant difference between the performance for grammatical and ungrammatical items at only one point – on the first posttest (T2). There, the participants (aggregated from both groups) are more accurate at judging grammatical items than ungrammatical items (V = 80, p-value = 0.089). Note that for this comparison we used the complete data set, including the data of the participants who scored 100% at T1, which meant 22 participants in total.

Summary: word order in subordinate clauses

For the instruction on the target structure *word order in subordinate clauses*, there was no broad and notable effect on the grammatical knowledge of either group. None of the two experimental groups showed a significant improvement on the sentence construction test – a test we assume to tap more into explicit knowledge. For the timed grammaticality judgment test, which we assume to tap more into implicit knowledge, there were three interesting developments. First, the free-recast group showed a marginally significant improvement between the pretest and the delayed posttest. Second, the free-recast group outperformed the constrained group on the delayed posttest, while their performance was on the same level for all other tests. The third interesting development is that for the ungrammatical items of the judgment test, the constrained instruction showed an immediate effect - the group who received it performed significantly better at the first and second posttest than on the pretest. There was no comparable effect for the free-recast instruction.

In summary, we can state that both types of instruction yield some small effects – the free-recast is more delayed and potentially indirect, while the constrained is more immediate and only concerns knowledge regarding the ungrammatical items.

9.1.2 Dative case in prepositional phrases

This section presents the learner development for the dative case in prepositional phrases (DatPP). As above, we will present the results of the sentence construction (SC) test and the timed grammaticality judgment test (TGJT) in separate sections, each we will start by providing descriptive statistics and then discuss the inferential statistics. We consider data from three experimental conditions: Two instruction conditions implement the FOCUS-ON-FORM approach by allowing free input from the learner – in one the corrective feedback is given in form of recasts, in the other, it is given in form of metalinguistic feedback. The third type of instruction implements the FOCUS-ON-FORMS approach and allows only constrained input from the learner. We refer to them as **Free-recast**, **Free-metaling** and **Constrained**.

As we have discussed in the introduction to 9.1, we tried to compensate the sparseness of data points by including additional data from participants who dropped out during the course of the experiment. This means that we separately analyzed all the data we had of participants that took part in the first session that comprised the pretest and the first posttest (T1-2 data set– ••○○) – which included 67 participants. Similarly, we also looked separately at the data of participants that took part in the first two sessions and provided data for the pretest and the first and second posttest (T1-3 data set– •••○) – these included data from 53 participants. There were 33 participants who took part in the complete experiment (T1-4 data set– ••••).

However, the data set was further reduced slightly because we excluded the data of those participants who achieved a perfect score of 100% at the pretest T1. We reckon that these participants cannot achieve any further learning gain and would therefore skew the results. As a result, the number of participants whose data is usable for analysis of the complete experiment was further reduced to a number between 27 and 30 depending on the particular test and item subset. For the T1-3 data set, we could use

between 43 to 49 participants, for the T1-2 data set, there were 59 and 63 participants. We indicate the exact numbers when we discuss the individual results.

We will start the presentation with an overview of the data coming from those learners who contributed data along the complete experiment. Table 9.2 shows the results for the T1-4 data set by indicating the median (md) percentage scores and the interquartile ranges (iqr) for each of the three experimental groups (free-recast, free-metalinguistic feedback, and constrained) on both tests: sentence construction test (SC) and timed grammaticality judgment test (TGJT). For the latter test, the table also shows the scores for grammatical and ungrammatical items separately. The table further indicates the number of participants in each group - for the sentence construction test we had 30 participants in total, 10 for each group. For the grammaticality judgment test we also have data of 30 participants in total, but 11 for the free-recast group, nine for the free-metaling group, and 10 for the constrained group. The numbers for the subsets of grammatical and ungrammatical items differ slightly since there were in total three more learners who scored 100% on the grammatical items than learners who scored perfect for the ungrammatical items. We will now discuss the results in more detail, by starting with the sentence construction test in Section 9.1.2, and the judgment test in Section 9.1.2.

		T1		T2		T3		T4	
Group/test	n	md	iqr	md	iqr	md	iqr	md	iqr
Free-recast									
SC all	10	38	45	67	56	78	53	78	53
TGJT all	11	76	32	82	23	88	23	76	20
TGJT gram.	9	78	33	89	11	89	22	89	11
TGJT ungr.	11	62	44	75	32	75	32	62	50
Free-metaling									
SC all	10	28	56	33	19	38	31	56	50
TGJT all	9	53	24	65	24	53	24	65	18
TGJT gram.	8	56	8	72	36	72	14	89	14
TGJT ungr.	9	50	37	38	37	50	37	38	12
Constrained									
SC all	10	44	62	100	11	100	0	67	28
TGJT all	10	74	22	94	9	94	22	79	28
TGJT gram.	10	89	11	94	11	89	19	94	19
TGJT ungr.	10	56	25	100	19	94	35	68	44

Table 9.2 – Test results for DatPP: medians (md) and interquartile range (iqr) for percentage scores, sentence construction test (SC) and timed grammaticality judgment test (TGJT)

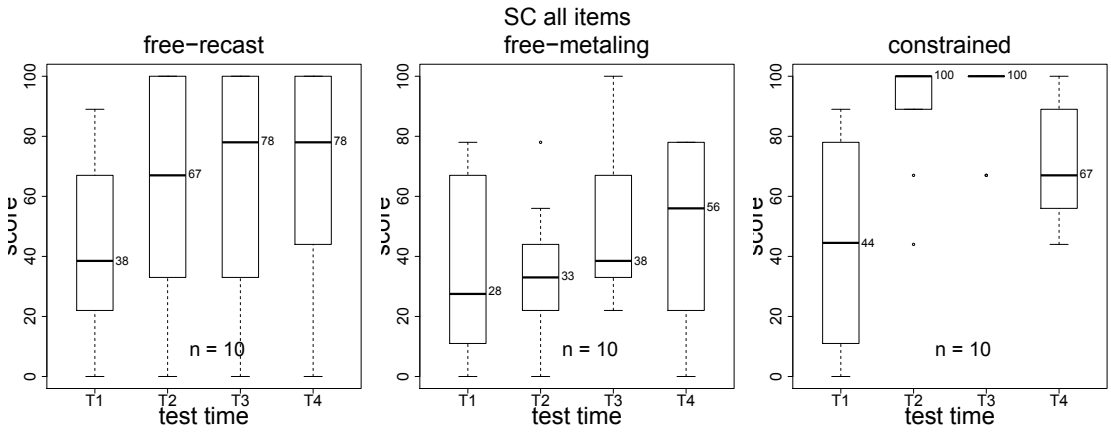


Figure 9.5 – Results for DatPP on the sentence construction test, represented as box plots.

Sentence construction test

Figure 9.5 shows box plots and means for the sentence construction test, thereby graphically representing the data contained in Table 9.2, but in addition giving more information about the dispersion of the results. At a first glance, the development of the three groups seems quite different. The performance of the free-recast group increases steadily over the course of the the four tests, but with quite a large dispersion. At the delayed posttest T4, even though the median does not increase any further, the lower quartile increases slightly. The free-metaling group starts as the lowest of all groups, but, similar to the recast group, increases steadily in terms of the median values at every posttest. The constrained group starts at the highest level of all groups, reaches the maximum median of 100% already at T2 and further increases at T3, as indicated by the increase of the lower quartile, but decreases again at the delayed posttest T4.

T1	T2	T3	T4
no diff.	$ML < C$ all	$ML < C$ all	no diff.
no diff.	$R < C$ ●●●○	no diff.	no diff.

Table 9.3 – Differences between groups for the sentence construction test, ML: metalinguistic group, C: constrained group, R: recast group

Differences between groups Table 9.3 summarizes the significant differences between groups. Based on the Kruskal-Wallis one-way analysis of variance test, we found no significant difference between the groups at T1. However, there are a few differences at T2 and T3. The post-hoc analysis indicated that the constrained group has a significantly higher score than the free-metalinguistic group at T2 and T3 for all applicable data sets. The constrained group further outperforms the free-recast group at T2, but only for the T1-3 data set.

There are no differences between the two free input groups at T2 or T3. Further, there is also no difference between any pairing of groups at the delayed posttest T4.

subset	n	T1 md	T1 iqr	T2 md	T2 iqr	T3 md	T3 iqr	T4 md	T4 iqr	T1 T2	T1 T3	T2 T3	T1 T4	T2 T4	T3 T4
Sentence construction test															
Free-Recast															
●●○○	17	33	45	56	45					■□					
●●●○	14	38	42	56	34	78	50			□□	■□	□□			
●●●●	10	38	45	67	56	78	53	78	53	□□	■□	□□	■■	□□	□□
Free-Metalinguistic Feedback															
●●○○	23	44	50	44	34					□□					
●●●○	19	44	50	44	45	56	50			□□	■■	■□			
●●●●	10	28	56	33	19	38	31	56	50	□□	□□	□□	□□	□□	□□
Constrained															
●●○○	19	56	56	89	28					■■					
●●●○	14	62	56	100	19	100	22			■■	■■	□□			
●●●●	10	44	62	100	11	100	0	67	28	■■	■■	□□	□□	□□	□□

Table 9.4 – Test results for sentence construction test, medians (md) and interquartile range (iqr) for percentage scores; differences between test times: ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

Differences between tests Table 9.4 indicates for each of the three subsets of test times which of the differences between test times are significant, based on the results of the Friedman test with pairwise post-hoc comparisons. The table further contains medians and interquartile ranges.

For the free-recast group, there is a significant increase between T1 and T4. Further, there is a marginally significant gain between T1 and T3. Finally, for the largest data set (T1-2 data set) there is a marginally significant increase between T1 and T2.

The free-metaling group improves significantly between T1 and T3, and marginally between T2 and T3, but both these differences apply only to T1-3 data set.

Finally, the constrained group shows significant increase between T1 and T2 and between T1 and T3, across all subsets of test times.

In summary, these results indicate that all experimental groups show some improvement over time on the sentence construction test. The constrained group displays the most pervasive increase, which is consistent with its superiority indicated by the between-group comparisons discussed above. In particular, the immediate effect of the instruction, as indicated by the development between T1 and T2, is absent for the metaling group and only very spotty for the recast group.

Timed grammaticality judgment test

Figure 9.6 illustrates the development over the four test times for the timed grammaticality judgment test. The box plots illustrate that all groups increase to a certain extent between T1 and T2. Between T2 and T3, the free-recast group increases further, while the free-metaling group deteriorates and the constrained group maintains its median,

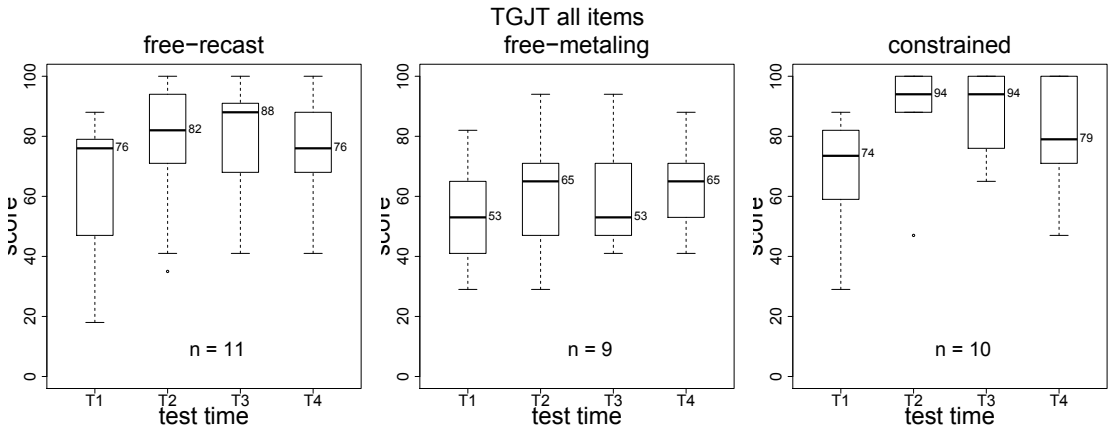


Figure 9.6 – Results for DatPP on the timed grammaticality judgment test, represented as box plots.

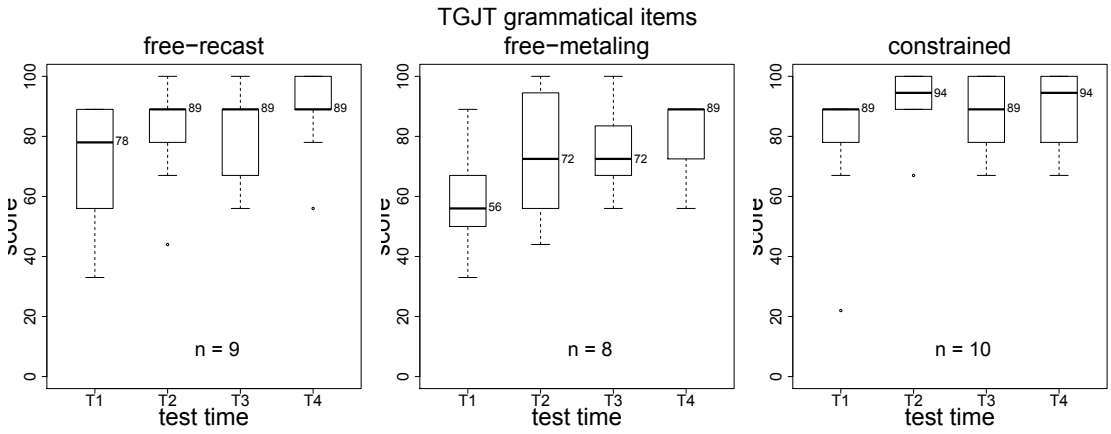


Figure 9.7 – Results for DatPP on the timed grammaticality judgment test, grammatical items only.

but decreases its lower quartile, which indicates a slight deterioration. At the delayed posttest T4, the free-metaling group increase their median again to the level of T2, while the free-recast and the constrained group decrease below their T2 level.

The development of the free-recast and the constrained group are similar in that they reach a maximum at T2 or T3 and decrease again at T4. In contrast to that, the free-metaling group, who starts from the lowest score of all groups, fluctuates between two median scores, with the same lower score at T1 and T3 and the same higher score at T2 and T4.

Grammatical and ungrammatical items When we look at the grammatical and the ungrammatical items separately, as illustrated in Figure 9.7 and Figure 9.8, we can see that the performance for the grammatical items seems consistently better than the performance for the ungrammatical items, across all test times and experimental groups. We tested for significance of these differences using the Mann-Whitney U test, however, in order to not distort the differences, we also included the data of participants who achieved a perfect score of 100% for either the grammatical or ungrammatical

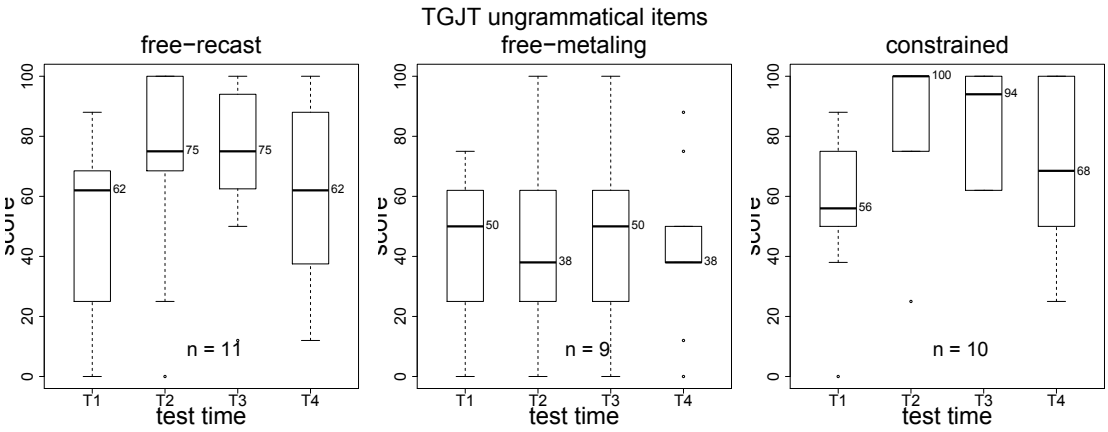


Figure 9.8 – Results for DatPP on the timed grammaticality judgment test, ungrammatical items only.

items at the pretest. The tests confirmed that the scores for the grammatical items are significantly higher than the scores for the ungrammatical items for most of the test times for all groups. For the constrained group this difference holds for all test times; for the recast and the metalinguistic feedback group it holds for T1, T2, and T3, but not for T4. This is in accordance with observations made by Hedgcock (1993); Loewen (2009), but unlike the majority of the data for the subordinate clauses discussed above.

When we compare the development of the grammatical (Figure 9.7) and ungrammatical items (Figure 9.8) with the complete item set, we notice the following: For the free-recast group, the development on the ungrammatical items is very similar to the development on the total item set, with the highest scores at T2 or T3 and a decrease at T4. The development of the grammatical items, however, shows a slightly different pattern, as there is no decrease at T4, but rather a further small increase. This increase is indicated by the increase of quartiles and of the sample minimum, even though the median is the same as at T2 and T3.

For the free-metalinguistic group, the performance on the grammatical items shows a steady increase similar to the recast group. The performance on the ungrammatical items, however, has an unusual pattern. It fluctuates in a complementary way to the score of the complete item set – there is a decrease between T1 and T2, but at T3, the median of T1 is reached again and at T4, the median increases up to the level of T2 again, albeit with a much smaller dispersion. As we will see later, however, this development, in particular the decrease between T1 and T2, seems to be a peculiarity of the T1-4 data set. For the larger data sets that contain data from the participants who did not take part at the later tests (T1-2 data set and T1-3 data set), there is an increase between T1 and T2 and a subsequent decrease between T2 and T3.

The performance of the constrained group on the grammatical items fluctuates on a high level between two rather close median values (89% and 94%). The median increases from T1 to T2, goes down again at T3 to the level of T1, and increases again at T4 to the level of T2. Although the median is the same at T2 and T4, the fact that the lower quartile is higher at T4 than at T2 indicates that the average performance is slightly better at T4. The performance on the ungrammatical items follows the same

pattern as the performance on the complete item set, with a considerable increase between T1 and T2, followed by a steady decrease after that.

Across all three groups, we can summarize that, for the grammatical items, both free groups increase over time, while the constrained group fluctuates on a high level. On the ungrammatical items, the recast and constrained group reach their maximum at T2 and then decrease until T4, but not below the median level of T1, which is similar to the development on the complete item set. The metalinguistic group shows an unusual decrease between T1 and T2 and does not increase above the level of T1 later, but as we have just noted, this development is not reflected in the larger data sets (T1-2 data set and T1-3 data set) in which the maximum median is reached at T2.

After we have described the development across test times based on the descriptive indicators, we are now going to discuss the differences between groups and between test times as confirmed by the statistical tests.

	T1	T2	T3	T4
TGJT all	no diff.	$ML < C$ all	$ML < C$ all	no diff.
TGJT gram.	$ML < C$●●●○	$ML < C$●●○○ ●●●○	no diff.	no diff.
TGJT ungr.	no diff.	$ML < C$ all	$ML < C$●●●●	no diff.

Table 9.5 – Differences between groups for the timed grammaticality judgment test, ML: metalinguistic group, C: constrained group

Differences between groups Table 9.5 gives an overview about the differences between groups. All differences regard the superiority of the constrained group over the metalinguistic group. There is no difference between the two free input groups and also no difference between the constrained group and the recast group.

At the pretest T1, the constrained group starts off from a significantly higher level than the metalinguistic group on the grammatical items, for the T1-3 data set. There is no other between-group difference at T1. At T2, the constrained group outperforms the metalinguistic group on the complete item set, as well as on the grammatical and ungrammatical items separately. For the grammatical items, however, the difference is only significant for T1-2 data set and T1-4 data set but not for T1-4 data set. At T3, the constrained group again outperforms the metalinguistic group, on the complete test item set and, for the T1-4 data set, on the ungrammatical items. At the delayed posttest, the groups do not differ.

Similar to the observations for the sentence construction test, the constrained group seems to benefit more than the metalinguistic group from the instruction in terms of immediate learning gains, even if we take into account that it starts off with a confined advantage at T1. The benefit however, does not last until the delayed posttest.

Differences between tests – all items Table 9.6 contains more detailed information about the complete item set of the judgment test – it presents medians and interquartile

subset	n	T1 md	T1 iqr	T2 md	T2 iqr	T3 md	T3 iqr	T4 md	T4 iqr	T1 T2	T1 T3	T2 T3	T1 T4	T2 T4	T3 T4
Grammaticality Judgment Test, all items															
Free-Recast															
●●○○	19	65	35	82	26					■■					
●●●○	16	74	32	85	23	74	34			■□	□□	□□			
●●●●	11	76	32	82	23	88	23	76	20	□□	■□	□□	□□	□□	□□
Free-Metalinguistic Feedback															
●●○○	24	59	24	65	30					■■					
●●●○	19	59	21	65	29	59	41			■■	□□	□□			
●●●●	9	53	24	65	24	53	24	65	18	□□	□□	□□	□□	□□	□□
Constrained															
●●○○	20	62	26	88	8					■■					
●●●○	14	65	22	94	10	88	18			■■	■■	□□			
●●●●	10	74	22	94	9	94	22	79	28	■■	■■	□□	□□	□□	□□

Table 9.6 – Test results for timed grammaticality judgment test, medians (md) and interquartile range (iqr) for percentage scores; differences between test times: ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

ranges and indicates which differences between tests are relevant for each of the three subsets of test times.

First of all, all three groups show some significant increase between T1 and T2. The recast group shows this difference only for the largest subset (T1-2 data set), and a marginally significant increase for T1-3 data set, but no difference for T1-4 data set. The metalinguistic group has a significant increase on T1-2 data set and T1-3 data set, but, like the recast group, no change at T1-4 data set. The constrained group increases significantly across all data sets.

Apart from the immediate increase between T1 an T2, there are two other changes. Between T1 and T3, the constrained group increases significantly on both applicable subsets (T1-3 data set and T1-4 data set). The recast group shows a marginally significant increase on T1-4 data set, while the metaling group does not change at all.

All three groups show some amount of immediate learning gain, but for the constrained group this gain is most comprehensive as it covers all subsets of the data. This is consistent with our previous observation that the constrained group benefits the most from the treatment. None of the groups maintain any learning gain until the delayed posttest.

Differences between tests - Grammatical and ungrammatical items Table 9.7 shows the significant differences for the grammatical items of the judgment test along with the median and interquartile range values; Table 9.8 contains the same information for the ungrammatical items.

When we compare the significant changes for the subsets of grammatical and un-

subset	n	T1 md	T1 iqr	T2 md	T2 iqr	T3 md	T3 iqr	T4 md	T4 iqr	T1 T2	T1 T3	T2 T3	T1 T4	T2 T4	T3 T4
Grammaticality Judgment Test, grammatical items															
Free-Recast															
●●○○	16	78	25	89	22					■■					
●●●○	13	78	22	89	22	78	22			□□	□□	□□			
●●●●	9	78	33	89	11	89	22	89	11	□□	□□	□□	■■	□□	□□
Free-Metalinguistic Feedback															
●●○○	20	56	26	78	25					■■					
●●●○	16	56	17	78	22	72	22			■□	■□	□□			
●●●●	8	56	8	72	36	72	14	89	14	□□	□□	□□	■□	□□	□□
Constrained															
●●○○	20	78	33	89	11					■■					
●●●○	14	78	11	94	11	89	19			■■	□□	□□			
●●●●	10	89	11	94	11	89	19	94	19	□□	□□	□□	□□	□□	□□

Table 9.7 – Test results for the grammatical items of the grammaticality judgment test, medians (md) and interquartile range (iqr) for percentage scores; differences between test times: ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

grammatical items with each other and with the complete item set, we notice the following: In general, the significant differences for the ungrammatical items are more similar to the differences for the complete item set, whereas the differences for the grammatical items are less similar to the complete set. In fact, for the metalinguistic group, the pattern of differences is exactly the same between total and ungrammatical item set. For the recast group, there is a variation on the difference between T1 and T2 - for the ungrammatical items, there is a marginally significant difference for T1-4 data set which does not exist for the complete item set, and the T1-T2 difference on T1-3 data set is significant, whereas it was only marginally significant for the complete item set. The constrained group shows a marginally significant increase between T1 and T4 for the ungrammatical items, which was not shown for the complete item set; apart from that, the significant increases are the same for the ungrammatical and the complete item set.

The grammatical items show a different set of significant changes. For the recast group, there is no increase between T1 and T3, as opposed to the ungrammatical and the complete item set, but instead a significant increase between T1 and T4. For the metalinguistic group, there are two marginally significant differences that are not present in either the complete or the ungrammatical item set – between T1 and T3 (for T1-3 data set only) and between T1 and T4. Finally, the constrained group only increases significantly between T1 and T2, but, as opposed to the ungrammatical and complete item set, this increase is not present for T1-4 data set. Furthermore, there is no increase between T1 and T3 and no increase between T1 and T4 for the constrained group.

	n	T1		T2		T3		T4		T1	T1	T2	T1	T2	T3
subset		md	iqr	md	iqr	md	iqr	md	iqr	T2	T3	T3	T4	T4	T4
Grammaticality Judgment Test, ungrammatical items															
Free-Recast															
●●○○	19	50	44	75	50					■■					
●●●○	16	56	40	75	41	75	50			■■	□□	□□			
●●●●	11	62	44	75	32	75	32	62	50	■□	■□	□□	□□	□□	□□
Free-Metalinguistic Feedback															
●●○○	24	50	40	56	50					■■					
●●●○	19	50	37	62	50	50	70			■■	□□	□□			
●●●●	9	50	37	38	37	50	37	38	12	□□	□□	□□	□□	□□	□□
Constrained															
●●○○	20	50	27	88	25					■■					
●●●○	14	50	22	100	25	82	38			■■	■■	□□			
●●●●	10	56	25	100	19	94	35	68	44	■■	■■	□□	■□	□□	□□

Table 9.8 – Test results for ungrammatical items of the grammaticality judgment test, medians (md) and interquartile range (iqr) for percentage scores; differences between test times: ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

In summary, we notice for the grammatical items that the constrained group has more learning gains than the other two groups, and, notably, is also the only group who shows some sign of longterm learning, as indicated by the marginally significant increase between T1 and T4. Opposed to that, for the ungrammatical items, both free input groups show some sign of longterm learning gain, while the constrained group does not. All groups show some immediate learning gains for the ungrammatical items as indicated by the increase between T1 and T2. Only the metalinguistic group shows a marginally significant increase between T1 and T3.

Summary: dative case in prepositional phrases

In order to summarize the development for the three different groups on the two different tests, we compiled Table 9.9, which puts the significant differences between test times next to each other. It shows that the constrained group seems to benefit most from the instruction in terms of immediate learning gains – as indicated by the significant differences between T1 and T2, as well as between T1 and T3 for both the sentence construction test and the timed grammaticality judgment test. However, the two free input groups (recast and metalinguistic feedback) also show some immediate improvement across the different tests and different subsets of test items, as well as subsets of considered test times, but not as consistently as the constrained group.

The more distinct gains of the constrained group compared to the free input groups also show in the direct comparison between group performances – the constrained group outperforms the free-recast group at T2 in the sentence construction test and the free-metalinguistic group at T2 and T3 on both the sentence construction as well as the

Free-recast				Free-metaling				Constrained			
T1-2	T1-3	T2-3	T1-4	T1-2	T1-3	T2-3	T1-4	T1-2	T1-3	T2-3	T1-4
Sentence construction test											
■□				□□				■■			
□□	■□	□□		□□	■■	■□		■■	■■	□□	
□□	■□	□□	■■	□□	□□	□□	□□	■■	■■	□□	□□
TGJT, all items											
■■				■■				■■			
■□	□□	□□		■■	□□	□□		■■	■■	□□	
□□	■□	□□	□□	□□	□□	□□	□□	■■	■■	□□	□□
TGJT, grammatical items											
■■				■■				■■			
□□	□□	□□		■□	■□	□□		■■	□□	□□	
□□	□□	□□	■■	□□	□□	□□	■□	□□	□□	□□	□□
TGJT, ungrammatical items											
■■				■■				■■			
■■	□□	□□		■■	□□	□□		■■	■■	□□	
■□	■□	□□	□□	□□	□□	□□	□□	■■	■■	□□	■□

Table 9.9 – Significant differences between test results for all groups (DatPP): ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not signficant

timed grammaticality judgment test (see Table 9.3 and 9.5).

On the other hand, it is apparent that the recast group shows the most long-term improvement compared to the other two groups, as indicated by their significant increase between T1 and T4 on the sentence construction test and the grammatical items of the judgment test. Compared to that, regarding the difference between T1 and T4, the metalinguistic group shows a marginally significant improvement on the grammatical items, while the constrained group shows such a marginally significant improvement on the ungrammatical items.

9.2 Development of oral communicative skills

In this section we present the development of the spoken language skills elicited in communicative tasks. We focus on the fluency of the learners, which we assess by two different measures. The first, described in Section 9.2.1, uses human ratings of perceived fluency, the second, described in Section 9.2.2, uses temporal measures of the learners' speech. As we have illustrated above in Figure 7.8 (Section 7.4), the constraints of the experimental setup only allowed for three times of elicitation the oral samples – as opposed to the four test times for the grammatical accuracy tests – since we could only record these samples in the beginning of each of the three sessions.

Due to the required effort and costs to rate and transcribe oral data, this analysis is

restricted to a subset of participant data - only the data collected in the first experiment (Dec 2009 / Jan 2010) was used, that means, we compare only the free-recast condition with the constrained condition. The number of participants whose data we could use is given in Table 9.10.

	Free/Recast		Constrained	
	T12	T123	T12	T123
Subordinate Clauses	10	7	9	8
Dative Prep. Phrases	8	6	10	7

Table 9.10 – Number of participants whose data was analyzed for oral communicative skills, T12 - data available on the first two tests, T123 - data available on all three tests

9.2.1 Holistic rating of perceived fluency

For the holistic rating of the speech samples, we employed three raters who rated all samples at least once. After the first round of rating, in which all three raters rated all participants once, we calculated the inter-rater agreement using Kendall's coefficient of concordance (W). The average inter-rater agreement is $W = 0.39$ across all samples. We then selected a subset of samples that contained (a) all those participants whose samples were rated with a low consistency ($W < 0.5$) and (b) a small subset of the remaining samples rated with higher consistency. These samples were rated again by the same raters in order to assess the intra-rater agreement (rater consistency). The most consistent rater achieved a Kendall's $W_3 = 0.88$, the other less consistent raters achieved $W_1 = 0.68$ and $W_2 = 0.66$ respectively. In particular the latter two values indicate that the rating task was hard, which was also confirmed by comments of the raters themselves.

For further analysis, we averaged across all existing ratings from the three raters. The following graphs depicted in Figure 9.9 to Figure 9.12 illustrate the ratings. They show for each rated participant the average of all ratings together (a bigger circle with grey filling) along with the one or two ratings of each of the three raters separately (indicated by the symbols ▫, ▵, ○ respectively). In case of repeated ratings, the first rating is to the left, the second to the right). The x-axis indicates the time at which the sample was recorded, the y-axis indicates the rating which ranges between 1 and 3, where 1 is the lowest performance and 3 the best. In case of ties - when the rater found two samples equally good, the ranks are fractions. The id of the participant is indicated at the top of each diagram.

Appointments Task/Subordinate Clauses

Figure 9.9 and Figure 9.10 illustrate the rankings for the appointment task with subordinate clauses as the target structure. Figure 9.9 shows the results for the free-recast group, Figure 9.10 shows the results for the constrained group.

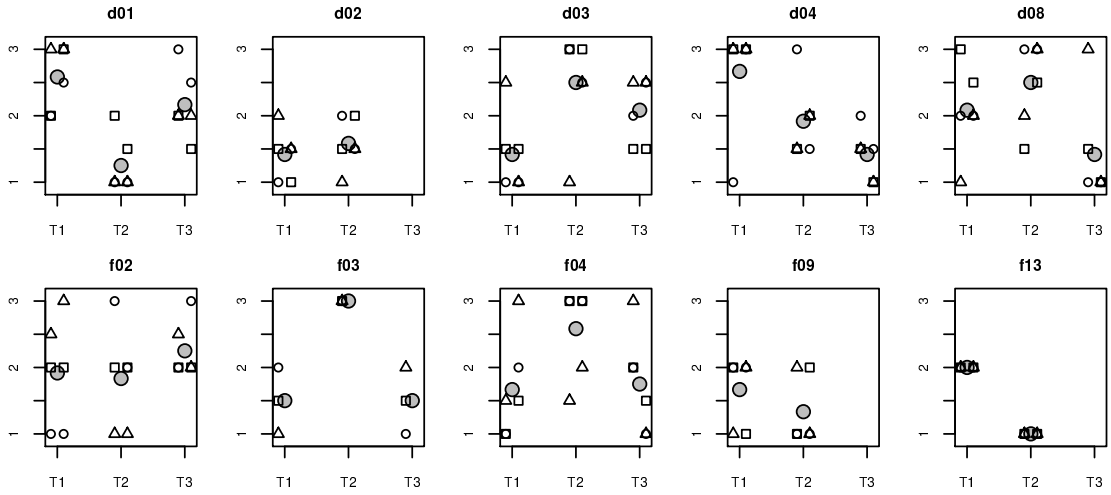


Figure 9.9 – Ratings across test times for appointment task scenario (SubC), Free-Recast group

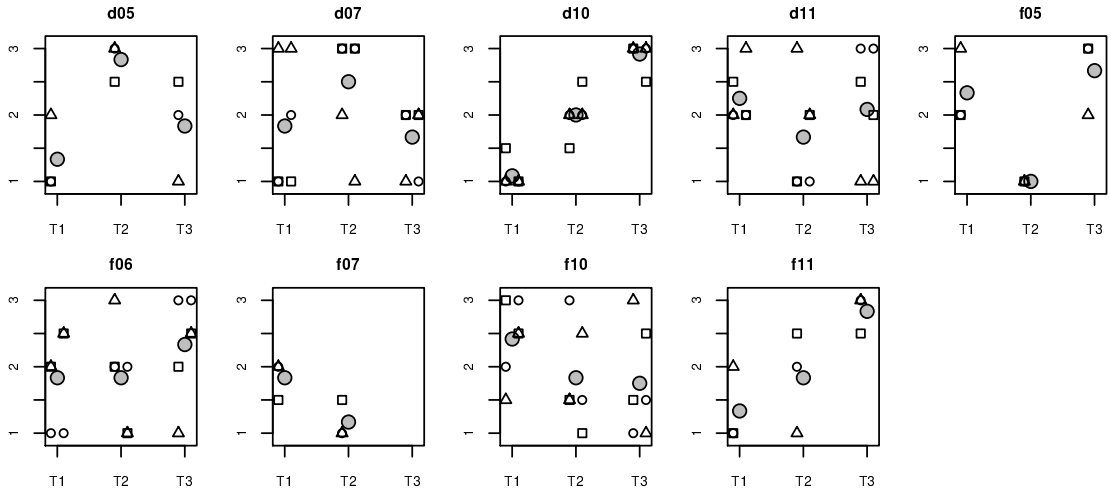


Figure 9.10 – Ratings across test times for appointment task scenario (SubC), Constrained group

For the free-recast group, five of ten participants (d02, d03, d08, f03, f04) show an improvement between T1 and T2, based on the averaged rating, but all of them show a decrease at T3 again (except for d02, for whom we did not have a sample for T3). The other five participants (d01, d04, f02, f09, f13) decreased their rating between T1 and T2, two of which had no T3 sample (f09, f13), and for the rest, one T3 rating was between T1 and T2 (d01), for one it was the lower than the previous two (d04), and for one the T3 ranking was higher than the first two rankings (f02).

Compared to that, the average rankings for the constrained group show four of nine participants increasing between T1 and T2 (d05, d07, d10), four participants decreasing (d11, f05, f07, f10) and one participant having the same rank for T1 and T2. Of those four who increase between T1 and T2, two show a further decrease at T3 (d10, f11), one decreases to a rank that is between T1 and T2 (d05), and one decreases to the lowest rank at T3 (d07). Of those four who decrease between T1 and T2, one has no T3 ranking (f07), one's T3 ranking is slightly lower than the two previous ones (f10),

another one's ranking is in the middle between T1 and T2 (d11) and the third one's ranking is higher than the previous two (f05). The one with equal rankings for T1 and T2, has a slightly higher rating at T3.

It seems that no clear tendencies emerge from the descriptive analysis so far. In both groups, half of the participants improve between T1 and T2 while the other half decrease.

We then tested whether any of the differences between the ratings at each test time are significant using the Friedman's Test with a posthoc analysis using the Wilcoxon-Nemenyi-McDonald-Thompson test as described in Hollander and Wolfe (1999) (as above in Section 9.1). We further tested the differences between the two groups using the Mann-Whitney U test.

For the appointment task scenario, we found no significant difference between the ratings of the three different test times. However, comparing the ratings of the groups showed a marginally significant difference at delayed posttest (T3) – the constrained group reached higher ratings on average than the free-recast group (W = 13, p-value = 0.0925).

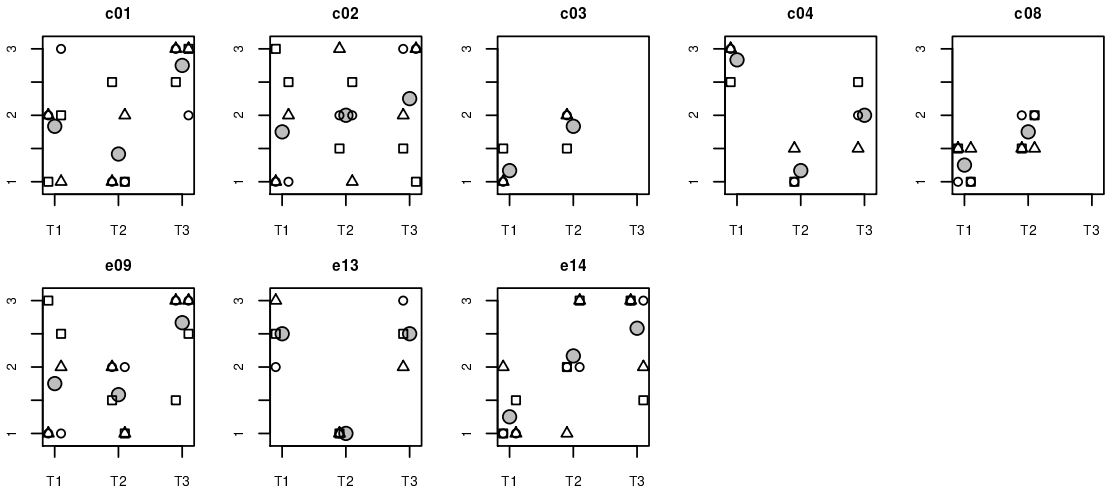


Figure 9.11 – Ratings across test times for directions giving task scenario (DatPP), Free-Recast group

Directions giving task/dative prepositional phrases

Figure 9.11 and Figure 9.12 show the rankings for each rated participant across the three different test times for the directions giving task with dative prepositional phrases as the target structure.

For the free-recast group, four of eight participants improve between T1 and T2 (c02, c03, c08, e14), the other four decrease (c01, c04, e09, e13). Of the four improvers, two improve further at T3 (c02, e14), the other two have no rating at T3 (c03, c08). Of the four participants who decrease in the beginning, the T3 ranking of two is higher than the previous two (c01, e09). For one of them (e13), the T3 ranking the same as the T1 ranking, for the other one (c04), it is between T1 and T2. Notably, for all participants who were tested at T3, their T3 rating is higher than it is at T2.

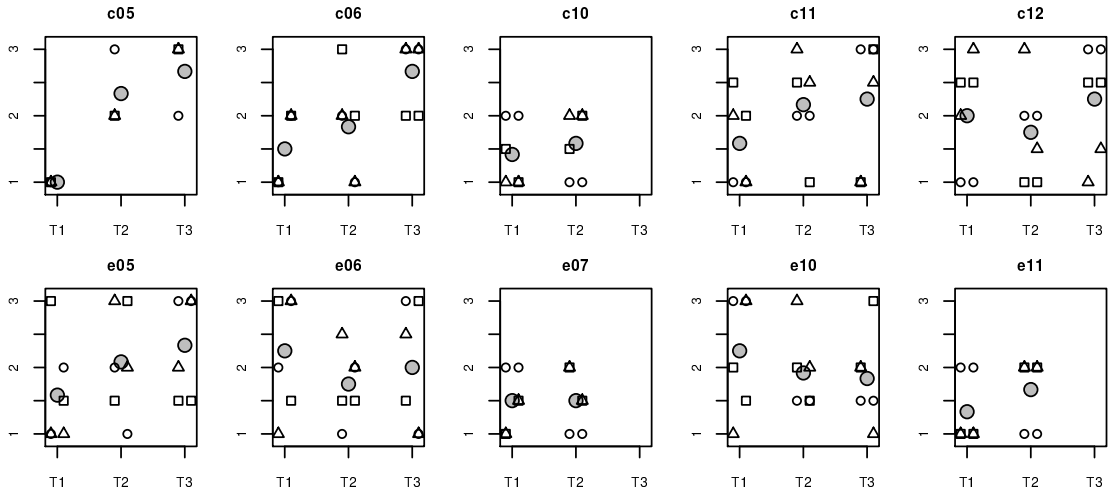


Figure 9.12 – Ratings across test times for directions giving task scenario (DatPP), Constrained group

Compared to that, the constrained group has six of ten participants who improve between T1 and T2 (c05, c06, c10, c11, e05, e11) and three who decrease (c12, e06, e10). There is one (e07) whose rankings at T1 and T2 are equal and there is no data at T3 for them. Of the six who increase between T1 and T2, four improve further at T3 (c05, c06, c11, e05) and for the other two there is no data for T3 (c10, e11). Of the three who decrease, one's T3 ranking is higher then the first two rankings (c12), one's is lower (e10), and the third one's is in the middle (e06).

From looking at the descriptive data, it seems that for the directions giving scenario, the constrained group has noticeably more participants who show a pattern of steady increase over the course of the experiment then the free-recast group and than any group on the appointments task scenario. However, none of these differences between tests turn out to be significant. Opposed to that, there is a significant increase for the free-recast group between T2 and T3 (p-value: 0.0326). There are no significant differences between the groups at any test time.

9.2.2 Temporal measures

As we have discussed above in Section 7.3.3, by transcribing and annotating the speech samples, we extracted a set of temporal measures that relate to the fluency of the sample. These measures capture the length of pauses and runs, the speech rate and the phonation-time ratio:

- mean lengths
  - of pauses in seconds
  - of runs in number of syllables (including filled pauses)
  - of runs in number of syllables (excluding filled pauses)
- speech rate

  - syllables per second (including filled pauses)
  - syllables per second (excluding filled pauses)
  - words per second
- phonation-time ratio
  - disregarding filled pauses
  - counting filled pauses as phonation
  - counting filled pauses as silence

Regarding the mean length of runs, the speech rate, and the phonation-time ratio it has been shown in previous work (cf. Section 7.3.3) that they correlate positively with impression of fluency. The mean length of pauses correlates negatively with fluency.

In the remainder of this section we will present the development of fluency over the three test times by reference to these measures. We will start in Section 9.2.2 with the data for the appointment task scenario involving subordinate clauses and in Section 9.2.2 we describe the data for the directions giving scenario with dative prepositional phrases. The majority of the data was normally distributed according to the Shapiro-Wilk test and of equal variance as asserted by the Levene test. For these data we used a paired t-test to compare the performance between each pair of tests: T1-T2, T1-T3, T2-3, in order to find out if there were any significant changes. For the instances of measures in which the data was not normally distributed, we used the non-parametric counterpart of that test, the Wilcoxon signed-rank test. Although a common approach to detect a change over time is a repeated measures anova test (or a Friedman test for the non-parametric data), this was not the most appropriate approach for us, because we lost a few data points due to technical failures in recording. In a repeated measures test, these subjects with missing data points would have to be removed and their data could not be used at all, while a pairwise comparison can make better use of the existing data. Furthermore, a repeated measures test over all three test times would be followed by a posthoc pairwise comparison anyway if it showed a difference, in order to identify between which test times the difference appears.

We do not apply an adjustment for the significance level which is usually required for multiple comparisons because we are interested in each of the pairwise differences individually. Therefore, our question is not so much whether there is a change across time but rather where exactly the change is if it is there. Furthermore, we are aware that some of the measures are highly dependent, but we do not analyze their interdependence in a multivariate approach as this is beyond the scope of this thesis.

Finally, we compare the means between the two experimental groups at each test time using a t-test and a Welch's t test (Welch, 1947) for those samples whose variance was not equal.

In each of the following two sections we will start to give an overview about the development and then present the significant changes.

Appointments Task/Subordinate Clauses

The plots depicted in Figure 9.13 illustrate the development of all measures. For each measure, there is a pair of plots – the left hand plot shows the mean values for each

of the two groups, the plot to the right indicates the values of each of the individual participants and thereby illustrates the spread of the data.

Table 9.11 summarizes the data for each measure in numbers, indicating mean values and standard deviation. Furthermore, the table also includes information about the significance of differences between test times.

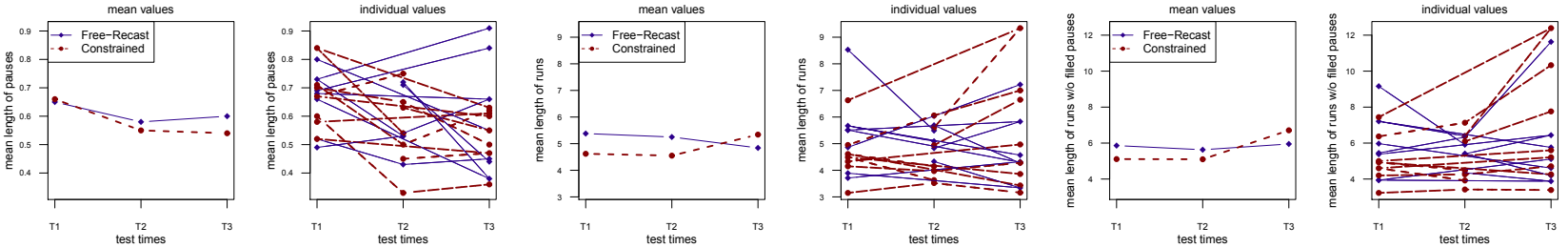
Before looking at the plots, recall that except for mean length of pauses, for all other measures it holds that higher values are related to a higher degree of fluency. In general, the plots indicate that the speech rate measures and the phonation-time ratio measures both increase over time for both groups, while the mean length of pauses decreases. This indicates that participants of both groups get more fluent over time. However, there is an exception for the measure "mean length of runs" - while the constrained group increases on that measure, the free-recast group does not. For the runs that include filled pauses,the group slightly decreases over time (Figure 9.13(2)); it stays at the same level for the runs that exclude filled pauses (Figure 9.13(3)). This means that there is one indicator that does not support an increase in fluency for the free-recast group.

Differences between groups From the plots, we see that the free-recast group has slightly higher values for most measures for most tests. If we test for the differences between groups, the only difference we find at a significance level of 0.05 is for the phonation time ration measures that counts filled pauses as silence at T1 (Figure 9.13(9)), where the free-recast groups starts of significantly higher. The difference at T2 is significant only at a level of 0.10 and at T3, the groups do not show a significant difference anymore. Furthermore, for the phonation time ratio that disregards filled pauses, the free-recast group has a higher ratio at T1 but again, only at a level of 0.1 (Figure 9.13(7)). Apart from these, there are no other significant differences between groups. Given that the existing differences are either limited to pretest T1 or follow from a difference at T1, they must be considered independent from the treatment and cannot be used to draw conclusions about any potential difference in effectiveness of the treatment.

Differences between test times Table 9.11 indicates the significant differences between test times for both groups. It shows that only the constrained group showed any significant changes and that most changes appeared between T1 and T3. The only immediate change (between T1 and T2) appeared for the mean length of pauses (at α = 0.1). At this measure, there also was a significant increase between T1 and T3. For the mean length of runs excluding filled pauses, there was a marginally significant increase between T2 and T3.

Regarding the three speech rate measures, only words per second showed an increase between T1 and T3. All three phonation time ratios showed an increase involving T3.

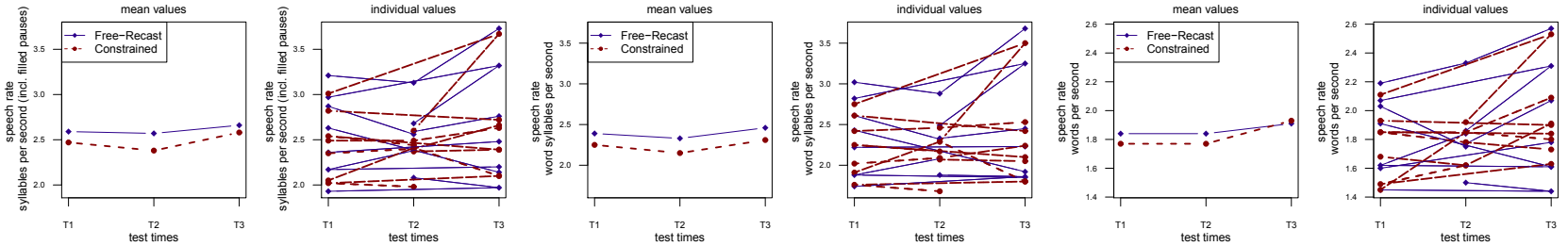
In conclusion, we can see that for the appointment task scenario, the constrained group shows a clearer increase in fluency, while the increase of the free-recast group is not significant.



((1)) mean length of pauses in seconds

((2)) mean length of runs in number of syllables (including filled pauses)

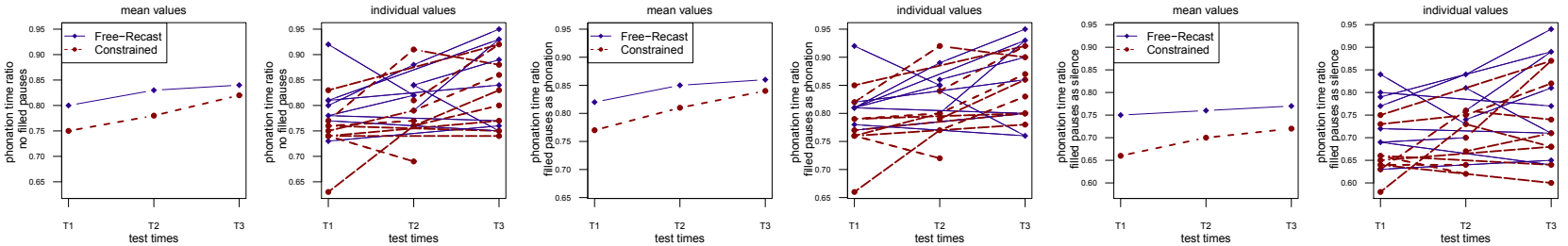
((3)) mean length of runs in number of syllables (excluding filled pauses)



((4)) speech rate
syllables per second (incl. filled pauses)

((5)) speech rate
syllables per second (excluding filled pauses)

((6)) speech rate
words per second



((7)) phonation time ratio
disregarding filled pauses

((8)) phonation time ratio
counting filled pauses as phonation

((9)) phonation time ratio
counting filled pauses as silence

Figure 9.13 – Appointment task scenario (SubC), means and individual values for different temporal measures

Measure	T1 m	T1 sd	T2 m	T2 sd	T3 m	T3 sd	Differences T1-2	Differences T1-3	Differences T2-3
Free-Recast	n=7		n=6		n=7				
mean length									
of pauses	0.65	0.11	0.58	0.11	0.6	0.21	□□	□□	□□
of runs	5.38	1.6	5.25	0.62	4.84	1.28	□□	□□	□□
of runs w/o filled pauses	5.86	1.85	5.63	0.76	5.95	2.66	□□	□□	□□
speech rate									
syllables per second (incl. filled pauses)	2.59	0.46	2.57	0.35	2.66	0.66	□□	□□	□□
syllables per second (excl. filled pauses)	2.39	0.47	2.33	0.34	2.46	0.73	□□	□□	□□
words per second	1.84	0.28	1.84	0.27	1.91	0.42	□□	□□	□□
phonation time ratio									
no filled pauses	0.8	0.06	0.83	0.03	0.84	0.08	□□	□□	□□
filled pauses as phonation	0.82	0.05	0.85	0.03	0.86	0.07	□□	□□	□□
filled pauses as silence	0.75	0.07	0.76	0.05	0.77	0.12	□□	□□	□□
Constrained	n=7		n=7		n=8				
mean length									
of pauses	0.66	0.11	0.55	0.14	0.54	0.09	■□	■■	□□
of runs	4.62	1.05	4.55	0.99	5.34	2.15	□□	□□	□□
of runs w/o filled pauses	5.11	1.4	5.1	1.42	6.71	3.18	□□	□□	■□
speech rate									
syllables per second (incl. filled pauses)	2.47	0.37	2.38	0.19	2.58	0.5	□□	□□	□□
syllables per second (excl. filled pauses)	2.25	0.37	2.15	0.25	2.31	0.55	□□	□□	□□
words per second	1.77	0.24	1.77	0.12	1.93	0.28	□□	■■	□□
phonation time ratio									
no filled pauses	0.75	0.06	0.78	0.07	0.82	0.06	□□	■■	■□
filled pauses as phonation	0.77	0.06	0.81	0.06	0.84	0.05	□□	■■	■■
filled pauses as silence	0.66	0.06	0.7	0.05	0.72	0.09	□□	■□	□□

Table 9.11 – Summary of temporal measures for appointment task scenario (SubC), indicating means (m) and standard deviation (sd), as well as pairwise significant differences between test times, ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

Directions giving task/dative prepositional phrases

The plots in Figure 9.14 indicate the development of the two groups on the different measures across test times. As with the previous scenario, the plot to the left of each pair indicates the mean values, the plot to the right provides individual values of each participant. When we compare the patterns of development with the previous scenario above, we notice first that the mean values for both groups are much more similar at T1 for most measures, except for mean length of pauses. Further, for the speech rate and phonation time ratios, we notice that the values for both groups are also very similar at T3, but that at T2, the constrained group scores markedly higher than the free-recast group. While the constrained group improved between T1 and T2, the free-recast group declined.

The development at the mean length of pauses is different – the free-recast group starts off higher than the constrained group, but at T3, the difference has inverted. For the two measures regarding the mean length of runs, the groups start off the same, both decrease slightly at T2 and the constrained group increases at T3 compared to T1, while the free-recast group stays about the same. A look at the individual development suggests that the increase of means for the constrained group at T3 can be attributed to the exceptionally high value of only one participant.

Differences between groups According to the t-test, the only significant difference between the groups is at the mean length of pauses at T1, where the free-recast group shows longer pauses than the constraint group ($\alpha = 0.05$) (Figure 9.14(1)). For all other measures and test times, the performance does not differ significantly. Since the difference regards a test before the actual treatment, it cannot be used to evaluate the effect of the treatment.

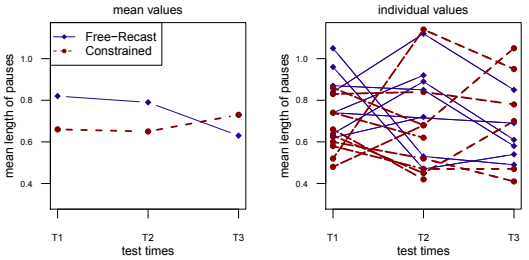
Differences between test times Table 9.12 indicates the changes between test times for each group. It strikes the eye that there are more significant changes for the free-recast group and that these changes predominantly involve T3. In terms of immediate changes (between T1 and T2), only the mean length of runs is marginally significantly changing, but it actually decreases for both groups – for the free-recast group, the length of runs including filled pauses decreases, while for the constrained group, the length of runs without filled pauses decreases.

Apart from the difference on length of runs, the only other difference between test times for the constrained group is a marginally significant increase between T2 and T3 for the phonation time ratio (filled pauses counted as phonation).

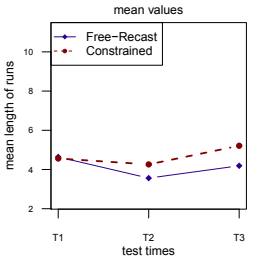
The free-recast group shows a decrease in length of pauses, most clearly between T2 and T3 and to a smaller degree between T1 and T3. For the mean length of runs, measured in number of syllables not including filled pauses, the recast group has marginally significant higher values at T1 compared to T2 and T3. The speech rate measured in syllables per seconds increases significantly between T2 and T3, whether or not filled pauses are included. The speech rate in terms of words per second did not differ between the tests. For all versions of the phonation time ratio measure, there was a marginally significant increase between T2 and T3. For the phonation time ratio

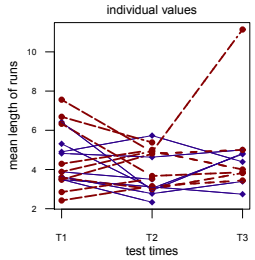
that disregards filled pauses and for the ratio that considers filled pauses as silence and not as phonation, there was also a marginally significant increase between T1 and T3.

In conclusion, we see that, for the directions giving task, the free-recast group shows distinctly more indicators of an increase in fluency, than the constrained group does.

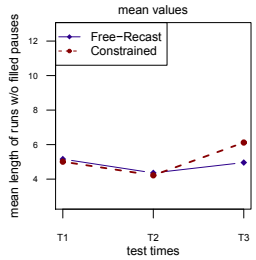


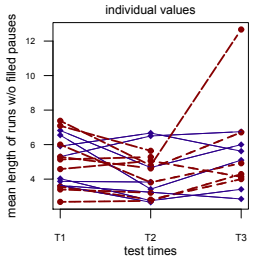
((1)) mean length of pauses in seconds



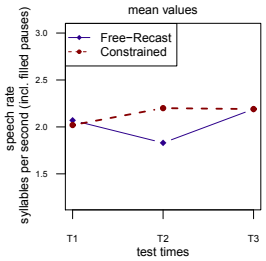


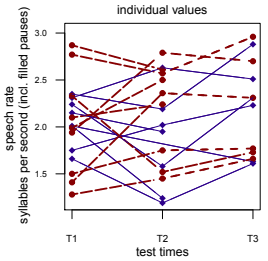
((2)) mean length of runs in number of syllables (including filled pauses)



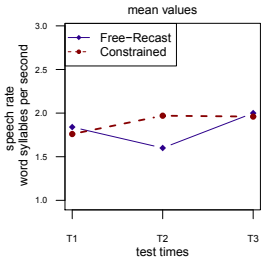


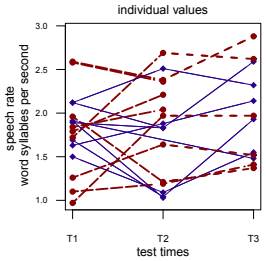
((3)) mean length of runs in number of syllables (excluding filled pauses)



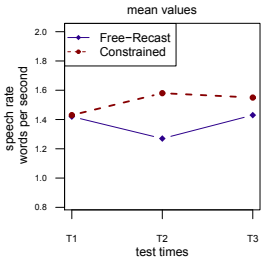


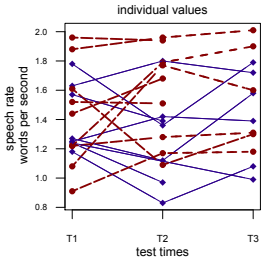
((4)) speech rate
syllables per second (incl. filled pauses)



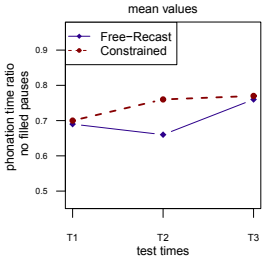


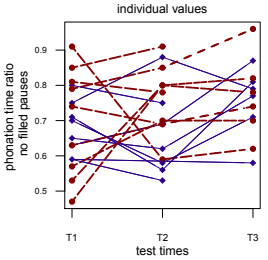
((5)) speech rate
syllables per second (excluding filled pauses)



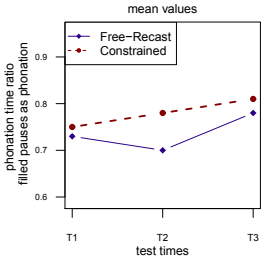


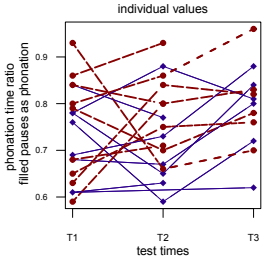
((6)) speech rate
words per second



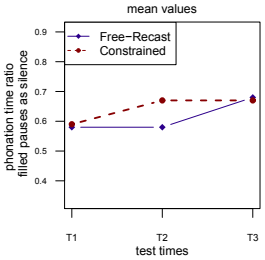


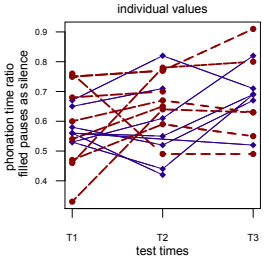
((7)) phonation time ratio
disregarding filled pauses





((8)) phonation time ratio
counting filled pauses as phonation





((9)) phonation time ratio
counting filled pauses as silence

Figure 9.14 – Directions giving task scenario (DatPP), means and individual values for different temporal measures

Measure	T1 m	T1 sd	T2 m	T2 sd	T3 m	T3 sd	Differences T1-2	Differences T1-3	Differences T2-3
Free-Recast	n=7		n=7		n=6				
mean length									
of pauses	0.82	0.16	0.79	0.23	0.63	0.13	□□	■□	■■
of runs	4.64	1.05	3.56	1.2	4.19	0.91	■□	■□	□□
of runs w/o filled pauses	5.16	1.33	4.36	1.67	4.96	1.52	□□	□□	□□
speech rate									
syllables per second (incl. filled pauses)	2.07	0.27	1.83	0.52	2.19	0.5	□□	□□	■■
syllables per second (excl. filled pauses)	1.84	0.24	1.6	0.56	2	0.44	□□	□□	■■
words per second	1.42	0.24	1.27	0.32	1.43	0.33	□□	□□	□□
phonation time ratio									
no filled pauses	0.69	0.07	0.66	0.12	0.76	0.1	□□	■□	■□
filled pauses as phonation	0.73	0.08	0.7	0.1	0.78	0.09	□□	□□	■□
filled pauses as silence	0.58	0.06	0.58	0.14	0.68	0.1	□□	■□	■□
Constrained	n=9		n=9		n=6				
mean length									
of pauses	0.66	0.13	0.65	0.23	0.73	0.25	□□	□□	□□
of runs	4.56	1.83	4.26	0.91	5.21	2.95	□□	□□	□□
of runs w/o filled pauses	5.01	1.63	4.22	1.1	6.12	3.36	■□	□□	□□
speech rate									
syllables per second (incl. filled pauses)	2.02	0.57	2.2	0.5	2.19	0.55	□□	□□	□□
syllables per second (excl. filled pauses)	1.76	0.58	1.97	0.52	1.96	0.65	□□	□□	□□
words per second	1.43	0.35	1.58	0.33	1.55	0.34	□□	□□	□□
phonation time ratio									
no filled pauses	0.7	0.15	0.76	0.1	0.77	0.12	□□	□□	□□
filled pauses as phonation	0.75	0.12	0.78	0.09	0.81	0.09	□□	□□	■□
filled pauses as silence	0.59	0.15	0.67	0.1	0.67	0.16	□□	□□	□□

Table 9.12 – Summary of temporal measures for directions giving task scenario (DatPP), indicating means (m) and standard deviation (sd), as well as pairwise significant differences between test times, ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

9.2.3 Summary of oral skills development

Appointments/SubC		
	between tests	between groups
holistic:	no differences	T3: C outperforms R
temporal:	C: some improvements (mostly between T1,T3)	T1/(T2): phonation-time-ratio: R outperforms C

Directions/DatPP		
	between tests	between groups
holistic:	R: T2 < T3	no differences
temporal:	R: some improvements at T3 C: minor improvements	T1: length of pauses: C outperforms R

Table 9.13 – Summary of oral skill development, C: constrained group, R: recast group

The development of oral skills as measured by a holistic rating by human raters and temporal measures is summarized in Table 9.13 and can be described like this:

For the appointments task, the holistic ratings show no difference between test times for any of the two groups, but the constraint group receives a higher rating than the recast group at the delayed posttest T3. In accordance with that, the constraint group shows an increase in fluency on some temporal measures, mostly between T1 and T3. There are no significant differences between test times for the recast group on any temporal measure. The recast group has higher phonation-time-ratio measures than the constrained group at T1, for one of these measures this difference is still marginally significant at T2, but since these differences existed before any treatment, no conclusions about the treatment can be drawn.

For the directions giving task, the holistic rating of the recast group shows an improvement between the posttest T2 and the delayed posttest T3. There is no difference of the holistic rating between test times for the constraint group and no differences between the two groups at any test time. The recast group shows some improvements at most of the temporal measures, most of them between T2 and T3 and between T1 and T3, while the constraint group shows only a marginal improvement at two measures. The constrained group shows shorter average length of pauses at T1 than the recast group.

It is interesting to notice that the two experimental groups showed a somewhat complementary effect for the two target scenarios. While the appointment making scenario led to more increase in fluency-related measures for the constrained group, the directions giving scenario induced more recognizable gains for the free-recast group. For both scenarios it is remarkable that the immediate changes between T1 and T2 were much rarer than changes involving the delayed posttest T3.

This chapter presented the results of the experiment we conducted in detail. We will summarize these findings in the next chapter and then discuss them.

10 Discussion

In this chapter we discuss the original findings of this thesis in the light of the questions that motivated this study. Additionally, we will point out shortcomings in the design and implementation of the study and suggest options for further work.

The questions addressed by this study are based on previous research in the field of second language acquisition which examined different types of instruction that differed with regard to (a) the weight they put on meaning or form and (b) how implicitly and explicitly they draw attention to formal aspects of the language. One important area that modifies explicitness is the feedback given in response to erroneous learner productions. In this study, we realize the instruction through an intelligent computer interface, because we are also interested in examining how language acquisition research based on human-only interaction can be implemented within a human-computer interaction setting, in which the computer provides instruction.

The questions that we addressed with this study were the following:

1. Is there a difference in effectiveness between the effects of computer-based FOCUS-ON-FORM and FOCUS-ON-FORMS instruction?

2. Is there a difference in effectiveness between computer-delivered recasts and metalinguistic feedback?

Before discussing the findings in more detail, we give a short summary of the previous chapter.

10.1 Summary of findings

In order to discuss the findings, we start off with a summary of the results as presented in Chapter 9. We then go on to discuss different aspects in more detail.

Figure 10.1 summarizes the significant changes between test times for the grammatical knowledge tests, Table 10.1 summarizes the significant differences between

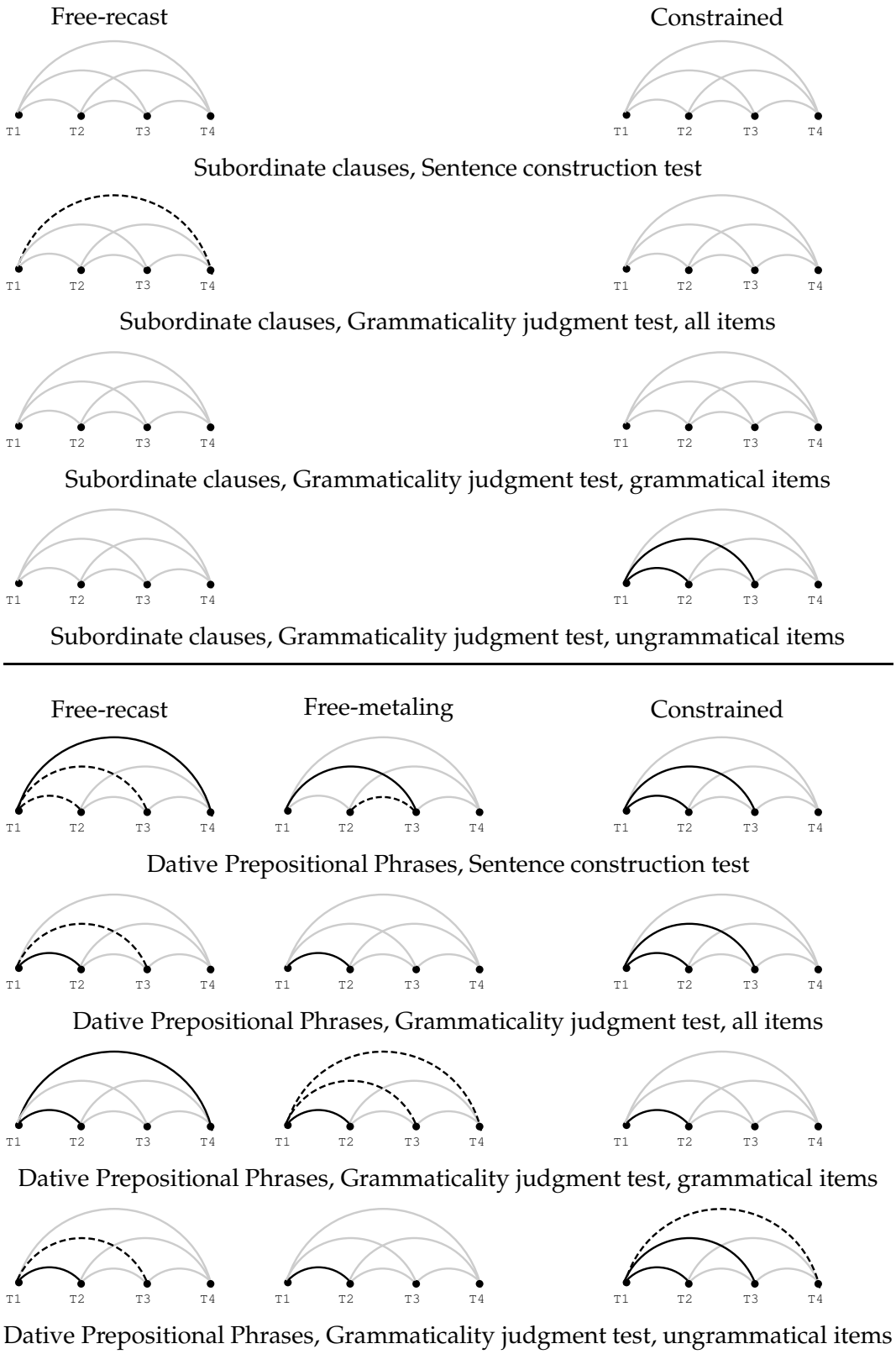


Figure 10.1 – Changes between test times for grammatical items. Solid arcs indicate a significant difference between test times at $\alpha = 0.05$, dashed arcs indicate a difference at significance level of $\alpha = 0.10$, grey arcs indicate no significant difference.

Measure	Appointments T1-2	Appointments T1-3	Appointments T2-3	Directions T1-2	Directions T1-3	Directions T2-3
Free-Recast						
mean length						
of pauses	□□	□□	□□	□□	■□	■■
of runs	□□	□□	□□	■□	■□	□□
of runs w/o filled pauses	□□	□□	□□	□□	□□	□□
speech rate						
syl. per second (w/ filled pauses)	□□	□□	□□	□□	□□	■■
syl. per second (w/o filled pauses)	□□	□□	□□	□□	□□	■■
words per second	□□	□□	□□	□□	□□	□□
phonation time ratio						
no filled pauses	□□	□□	□□	□□	■□	■□
filled pauses as phonation	□□	□□	□□	□□	□□	■□
filled pauses as silence	□□	□□	□□	□□	■□	■□
Constrained						
mean length						
of pauses	■□	■■	□□	□□	□□	□□
of runs	□□	□□	□□	□□	□□	□□
of runs w/o filled pauses	□□	□□	■□	■□	□□	□□
speech rate						
syl. per second (w/ filled pauses)	□□	□□	□□	□□	□□	□□
syl. per second (w/o filled pauses)	□□	□□	□□	□□	□□	□□
words per second	□□	■■	□□	□□	□□	□□
phonation time ratio						
no filled pauses	□□	■■	■□	□□	□□	□□
filled pauses as phonation	□□	■■	■■	□□	□□	■□
filled pauses as silence	□□	■□	□□	□□	□□	□□

Table 10.1 – Summary of between-test differences for temporal measures of communicative skills test; ■■– $p<0.05$, ■□– $p<0.10$, □□– $p\geq0.10$/not significant

		T1	T2	T3	T4
SubC	SC	no diff.	no diff.	no diff.	no diff.
	TGJT all	no diff.	no diff.	no diff.	$C < R$
	TGJT gram.	no diff.	no diff.	no diff.	no diff.
	TGJT ungr.	no diff.	no diff.	no diff.	no diff.
	Oral rating	no diff.	no diff.		$R < C$
	Temp. Meas.	$C < R$[1]	$C < R$[2]		no diff.
DatPP	SC	no diff.	$ML < C$ all $R < C$ ●●●○	$ML < C$ all	no diff.
	TGJT all	no diff.	$ML < C$ all	$ML < C$ all	no diff.
	TGJT gram.	$ML < C$●●●○	$ML < C$ ●●○ ●●●○	no diff.	no diff.
	TGJT ungr.	no diff.	$ML < C$ all	$ML < C$●●●●	no diff.
	Oral rating	no diff.	no diff.		no diff.
	Temp. Meas.	$R < C$[3]	no diff.		no diff.

[1] at two phonation-time ratio measures, for one marginally different
[2] at one phonation-time ratio measure, marginally different
[3] at mean length of pauses

Table 10.2 – Differences between groups for both target structures/task scenarios and all tests, ML: metalinguistic group, C: constrained group, R: recast group

test times on temporal measures, and Table 10.2 summarizes the differences between the experimental groups at the different test times for all measures. Recall that there were four test times for the grammatical knowledge tests, but only three test times for the communicative skills, because we could not conduct a test immediately after the first treatment. Therefore we map T2 of communicative skills to T3 of the grammatical tests, and T3 to T4 respectively in Table 10.2.

For illustrative purposes, the graphs depicted in Figure 10.1 merge the separate analyses that we made for each of the three different data sets (including T1-2 data set, ●●○○, T1-3 data set, ●●●○, and T1-4 data set ●●●●) based on Table 9.9 (page 207).

The results can be summarized as follows:

1. **General Observations**

   More effects for dative prepositional phrases Figure 10.1 illustrates rather clearly that, overall, the instruction for dative prepositional phrases showed more effects for the development of grammatical accuracy than the instruction for subordinate clauses.

   Complementary development for communicative skills For communicative skills, as measured by temporal measures related to fluency, (Table 10.1), there is a complementary development – the recast group shows some increase in fluency in the directions scenario, but not in the appointments scenario, while the constrained group shows some increase in fluency in the appointments scenario but only very limited development in the directions scenario.

   Grammatical items are judged more accurately than ungrammatical items In the timed grammaticality judgment test, the performance on well-formed items is better than on ill-formed items across most tests and groups.

2. **Development of grammatical accuracy for subordinate clauses**

   Delayed effects for recast group The recast group shows a marginally significant increase between the pretest T1 and the delayed posttest T4 for the timed grammaticality judgment test. The good performance of the recast group at T4 is also expressed by their significant superiority to the constrained group (see Table 10.2).

   Some immediate effects for the constrained group In contrast to the delayed effects shown by the recast group, the constrained group shows some more immediate effects between T1 and T2 and between T1 and T3, but only for the ungrammatical items of the judgment test.

3. **Development of communicative skills for appointment scenario**

   Constrained group shows some long-term effect The constrained group increases on some of the temporal measures related to fluency, mostly between the pretest and the delayed posttest. Related to that, for the holistic fluency ratings, the constrained group receives a higher rating than the recast group at the delayed posttest. Apart from that, however, in relation to the holistic ratings, there are no significant differences between the test times for the constrained group.

No development for recast group The holistic fluency ratings and temporal measures show no evidence for development of communicative skills for the recast group.

4. **Development of grammatical accuracy for dative prepositional phrases**

   Between group comparison: Constrained group outperforms metalinguistic group at immediate posttests The constrained group outperforms the metalinguistic feedback group at the sentence construction test and the timed grammaticality judgment test at the two posttests T2 and T3, which indicates that they seem to benefit more from the instruction. For the sentence construction test, the constrained group also outperforms the recast group at T2. However, there are no differences between groups at the delayed posttest T4.

   Constrained group shows most immediate learning gains The constrained group shows more immediate learning gains on both grammar tests than the two free production groups – as indicated by the significant improvements between T1 and T2 and between T1 and T3. The free production groups (recast and metalinguistic feedback) also show some immediate improvement between different test times, but not as pervasively as the constrained group.

   Recast group shows most long-term development The recast group shows the most long-term development compared to the other two groups (as indicated by the significant increase between T1 and T4 for the sentence construction test and the grammatical items of the judgment test). Compared to that, regarding the T1-T4 development, the metalinguistic group shows a marginally significant improvement on grammatical items, while the constrained group shows such a marginally significant improvement on ungrammatical items.

5. **Development of communicative skills for directions giving scenario**

   Distinct long-term effect for recast group. The recast group receives a significantly higher rating at the delayed posttest compared to the immediate posttest. Consistent with this, they also improve at some of the temporal measures at the delayed posttest – mostly in comparison to the pretest, but some also in comparison to the posttest.

   Only marginal development for constrained group The constrained group shows no differences between the ratings at each of the three test times and only a marginally significant change at two temporal measures.

10.2 Discussion of results

We are now going to discuss in more detail the trends that emerged. In Section 10.2.1 and Section 10.2.2 we discuss the findings in terms of the research questions. In Section 10.2.3 we look at the difference between the grammatical and ungrammatical items of the judgment test. We then compare the development on the two different target structures in Section 10.2.4. We conclude by discussing the development of the communicative skills in Section 10.2.5.

10.2.1 Constrained instruction versus free input instruction

The first question this study targeted was whether there is a difference in developmental effects between computer-based FOCUS-ON-FORMS (constrained) and FOCUS-ON-FORM (free) instruction? The findings indicate that there are indeed differences, and they mostly are most evident in the **differences between immediate and delayed effects**.

Constrained shows more immediate results

The constrained group shows overall more immediate effects than the two free production groups. For the dative prepositional phrases, the constrained group shows a clear increase between the pretest and the first two posttests on both the sentence construction test as well as on the grammaticality judgment test. The two free production groups also show some short-term development, but it is not as comprehensive and distinct. This supports the assumption expressed, among others, by Ellis (2009a) and DeKeyser (2008), that implicit learning takes longer than explicit learning, which we have discussed in Section 4.3.3.

For the subordinate clauses, which entailed very little development in general, the constrained group showed significant improvement on the ungrammatical items of the judgment test between the pretest and the first and second posttest. This may be ascribed to the fact that learners in the free-recast condition, which entailed no immediate effect, were not forced to produce *weil*-clauses unlike the learners in the constrained condition. As we have shown in our analysis of the interaction between learners and the system (Section 8.2.4), learners only used *weil*-clauses in about a quarter of the opportunities they could have used them, and only one fifth of the produced *weil*-clauses were incorrect and required a corrective recast. This implies that a considerable proportion of learners avoided using *weil*-clauses and, even though the system provided many examples of correct *weil*-clauses, the mere perception of examples did not seem to be effective. Learners in the constrained condition, on the other hand, were explicitly corrected if they produced incorrect *weil*-clauses and arguably their errors were more evident.

Recast group shows more delayed effects

In contrast to the more convincing immediate gains of the constrained group, it is noticeable that the recast group shows more delayed effects - for the subordinate clauses at the judgment test, where they show a marginally significant increase between T1 and T4 and outperform the constrained group at the delayed posttest; and also for dative prepositional phrases, where they have a significant increase between T1 and T4 for the sentence construction test and the grammatical items of the judgment test. While the constrained group, who receives explicit instruction, is faster, the recast group, who receives implicit instruction, seems to take longer to learn but their learning is more sustainable. The metalinguistic feedback group, who received explicit feedback, show immediate gains comparable to the recast group, and delayed gains roughly on par with the constrained group.

Apart from general differences in the pace of learning, another possible reason for the more long-term effects of implicit instruction may lie in the indirect effects of the instruction – it might have increased the propensity of learners to exploit the input they received outside of the actual instruction or it might have increased their motivation to learn consciously. Since it is beyond the scope of this study to look deeper into the subsequent effects of instruction, such reasoning remains speculative. In any case, overall, our results suggest that the recast instruction seems to be more durable, while the constrained, explicit FOCUS-ON-FORMS instruction seems to entail short term effects that do not last. In the following, we will briefly review the existing research on delayed learning.

Delayed learning and consolidation of new knowledge

Previous research in language learning has shown some evidence for the effects of learning taking effect after a certain delay. Relevant studies vary with regard to the content of knowledge which ranges from novel words/vocabulary over morphological phenomena to syntax. Another difference is the nature of the target language that was used, with natural languages on the one hand and artificial and semi-artificial languages which combined word stems of an existing language with artificial morphemes on the other hand. Related to that is another crucial difference between the studies which concerns the control over exposure to the participants during the time span between the initial instruction and the delayed testing – usually only when artificial languages were involved, could any intermediate exposure to them be ruled out.

Clay et al. (2007) and Davis et al. (2009) tested the learning of novel words that do not exist in any natural language and found that knowledge of these words was better after an intermediate period without any exposure. In the study by Clay et al. (2007)the delay was 6-10 days and knowledge was measured indirectly through a picture-word inference test. In the experiment by Davis et al. (2009) the delay was only 24 hours and the knowledge was tested through lexical competition and recognition tasks as well as neurocognitive processes as measured by an fMRI device. The short delay of 24 hours included a night of sleep, which points to the effect of sleep in knowledge consolidation (Diekelmann et al., 2009; Walker, 2005).

Merkx et al. (2011) and Tamminen et al. (2012) compared immediate and delayed learning of morphosyntactical structures in a semi-artificial language that combined English lexis with artificial affixes. They showed that consolidation times of two days and two months without additional exposure or practice led to more generalized knowledge than immediate learning.

Grey et al. (2014) examined delayed effects for a morphosyntactical phenomenon (case-marking) and a syntactical phenomenon (word order) in a semi-artificial language with English lexis. With no additional exposure after initial learning, they found that knowledge related to case-marking further improved two weeks after the immediate test, and to a lesser degree word order did too.

Similarly, Morgan-Short et al. (2012) found delayed effects for learning the word order of an artificial language, but in this study the delayed test was administered after 3-6 months (mean about 5 months). While the test results of a grammaticality judg-

ment test were maintained, neuro-cognitive processes as measured by event-related potentials (ERPs), appeared to be more native-like after the delay.

While studies that examine the effect of instruction for natural languages have an overall trend that shows a decrease in performance at the delayed tests compared to immediate tests (Norris and Ortega, 2000), there are a few notable exceptions (Spada and Tomita, 2010; Ellis et al., 2006; Mackey, 1999; Morgan-Short and Bowden, 2006). For all of these, however, similar to our study, exposure and additional practice in the meantime cannot be excluded for certain. In summary, studies conducted under relatively rigorous laboratory conditions give evidence for the existence of delayed learning and consolidation of linguistic knowledge. These delayed effects rarely appear in studies that try to estimate the effect of particular types of instruction. This may be a result of the different focus and consequently of the design of these latter studies. Although our study was not focused on examining the long-term effects of the different parameters of instruction, they did appear for the instruction with recast feedback.

10.2.2 Recasts versus metalinguistic feedback

When we compare the effect of the two different feedback types in the free, FOCUS-ON-FORM condition in order to answer our second question (*"Is there a difference in effectiveness between computer-delivered recasts and metalinguistic feedback?"*), we see that the immediate effects are on par, but recasts seem to be slightly superior in terms of delayed effects. Therefore, in terms of immediate effects, our study is in accordance with previous research that found no difference between the two types of feedback (Loewen and Erlam, 2006; Sauro, 2009; Razagifard and Rahimpour, 2010); as discussed in more detail in Section 5.5.3. In terms of long-term effects, the superiority of recasts compared to metalinguistic feedback that we found in our study is contrary to the studies that found metalinguistic feedback to be superior (Rezaei and Derakhshan, 2011; Sheen, 2007; Ellis et al., 2006; Carroll and Swain, 1993).

With regard to the problem that learning effects are often measured with tests that tend to tap into explicit knowledge, and therefore might give an advantage to more explicit forms of instruction (in this case metalinguistic feedback), no clear difference is evident between the sentence construction test and the timed judgment test that measured explicit and implicit knowledge respectively.

Recall that the difference between previous research and the present study is that previous research compared recasts and metalinguistic feedback between humans only, either in face-to-face conversation or via a written chat interface, while the present study compares the two in a CALL setting with a computer system as the feedback provider. Recall further that the studies that used a written chat interface found no differences between the feedback types, while the studies examining feedback in face-to-face interaction found more benefits for metalinguistic feedback (Section 5.5.3). Our findings regarding the short-term development, together with previous results suggest that metalinguistic feedback has less advantage over recast feedback in type-written interaction compared to oral interaction. One reason for this may lie in the fact that the problem of recasts – that they are harder to notice because of their relative lack of salience – might be compensated by their accessibility onscreen. Learners with a

lower degree of phonological sensitivity or working memory capacity, who have been shown to benefit less from recast feedback (Robinson, 2001; Mackey et al., 2002), might be able to profit more from recast feedback in type-written modes.

Noticing

Our study did not include any specific methods of assessing the extent to which learners noticed the recasts they were given. However, in the survey we conducted after the second session, we asked the learners if they noticed that the system corrected some of their errors. Unsurprisingly, almost all of the learners in the metalinguistic feedback condition reported that they noticed the system's corrections.

In the recast condition, across both target structures, 10 percent of the learners indicated that they noticed the feedback that was given, about 40 percent did not notice any corrective feedback, about 30 percent noticed some feedback but did not notice or did not remember what exactly it was targeted at, and finally, about 20 percent did not respond to that question. The rates of noticing differed to some degree between the two target structures – in fact, for the subordinate clauses, none of the participants replied that they noticed corrective feedback on that structure. If you remember that learners did produce many fewer incorrect subordinate clauses than incorrect dative prepositional phrases, which resulted in considerably less corrective recasts of subordinate clauses (Section 8.2.4), this difference can be expected. In light of this imbalance, it is impossible to properly compare the noticeability of recast feedback for both structures with the setup of the current study. A comparison would require us to start from a similar number of erroneous utterances, which might be hard in a task-driven, near-natural context. Furthermore, the questionnaire that we used in our study is a relatively blunt tool for assessing the actual noticing. It was administered only after the second session, which may have been too late a stage to get reliable observations regarding the first treatment session. Furthermore, self-reporting of learners may not be the most reliable measure of actual noticing. Therefore, a possible extension to the current study could be to employ more sophisticated methods to assess the amount of noticing in order to get a more accurate picture about the extent to which learners noticed feedback.

The survey results suggest that there is indeed a lack of noticing in the recast condition. In the general comments about the system some of the recast learners also expressed that they would have wished for more explicit feedback or explanations. This is consistent with the observations made by Heift (2004) and Yang and Akahori (1999), that learners preferred the more explicit feedback if they were asked about their preference directly (as discussed in Section 5.5.2). The relatively low rate of noticing is also in line with concerns and evidence expressed by VanPatten (1990) about the attentional limits that may prevent a perfectly simultaneous attention to meaning and form. All these observations suggest that it might be worthwhile to modify the recasts such that they become more salient and noticeable, as we have discussed in Section 5.5.1. However, we have to keep in mind that increasing the explicitness of recasts may jeopardize the ideal of synchronous attention to form and meaning, as it was proposed in the first conceptualization of the FOCUS-ON-FORM approach (Long, 1991; Long and Robinson, 1998).

Furthermore, it would be an interesting extension of this work to examine the effects of further types of feedback that are suited to a type-written interface, for instance, clarification requests or explicit corrections without metalinguistic explanations and without pushing the learner for a correction.

As we have discussed in Section 5.3.2, there is some evidence that feedback that pushes learners to modify their erroneous utterances yields greater learning gains than feedback which does not (Lyster, 2004; Ammar and Spada, 2006; Izumi, 2002). Our results are in contrast to these findings, the reasons for that may lie in other parameters of the feedback used in these studies, or the fact that the feedback was given in human-human interaction context. The lack of differences that we found for the short-term effects are, however, in accordance with the findings reported in Lyster and Izquierdo (2009).

10.2.3 Differences between grammatical and ungrammatical items

If we look at the differences between grammatical and ungrammatical items in the grammaticality judgment test, we notice the following: The test scores for grammatical and ungrammatical items only differ at T2 for the subordinate clauses – learners of both experimental groups judge grammatical items more accurately. For the dative prepositional phrases, the grammatical items score higher than the ungrammatical items for all test times except for the recast and metalinguistic group at the delayed posttest T4, where there is no significant difference. These results are largely consistent with previous work that has shown that grammatical items are usually more likely to be judged correctly than ungrammatical items (Hedgcock, 1993; Loewen, 2009). Loewen cites a counter-example and recounts hypotheses and speculations about the processes that may work when learners judge grammatical and ungrammatical items. These seem to be mostly speculative. Juffs (2001), for instances, surmises that ungrammatical sentences take longer to judge because learners try to match the test item with their internal grammar, and are quick to find a match for grammatical sentences, while they try a number of different hypotheses for an ungrammatical sentence before they give up. Evidence that explicit knowledge is invoked for ungrammatical items has been provided by Ellis (1991) who found out from think-aloud protocols that learners often used their explicit knowledge for sentences they judged as ungrammatical or were not sure about. Another possible reason for the higher performance on grammatical sentences might be that learners, when in doubt, may be more likely to accept a sentence than to judge it as incorrect, and therefore the positive judgment have a higher frequency, which leads to more incorrect sentences being falsely judged as correct.

Regarding differences between the development for test times, Figure 10.1 illustrates the following differences: For the subordinate clauses, there is an improvement for the constrained group only for the ungrammatical items but not for the grammatical items. For the dative prepositional phrases, the immediate development (between T1 and T2) is the same for both grammatical and ungrammatical items, for all three experimental groups. Less immediate developments, however, show a different pattern between the grammatical and ungrammatical items for each of the group: The constrained group shows long-term improvement only for the ungrammatical items but

not for the grammatical items. In contrast, the metalinguistic feedback group shows some long-term improvement for the grammatical items but not for the ungrammatical items. Finally, the recast group is in the middle ground, since it shows a significant increase between T1 and T4 for the grammatical items and a marginally significant increase between T1 and T3 for the ungrammatical items.

We can conclude from these results that explicit instruction (constrained group) seems to be more beneficial for improving performance on ungrammatical items – the feedback for errors was much more evident in the constraint condition.

In general, it may be less likely that the performance for the grammatical items shows significant improvement as it already starts from a higher level in general. However, the fact that both recast and metalinguistic group show some long-term improvement on the grammatical items suggests that the free instruction might have provided more positive evidence that helped learners to recognize a greater number of correct items as correct.

10.2.4 Differences between development for the two target structures

The results reveal an obvious difference in development between the two target structures and task scenarios. Progress on accuracy for the dative prepositional phrases was more pervasive across the different types of instruction. This difference may be grounded in general differences between the two target structures that influence their teachability and their suitability to certain kinds of instruction and feedback. In Section 4.4, we discussed frequency and regularity, salience, and the functional value of structures, as well as the developmental readiness of the learners as factors that have an impact on how effectively different structures can be learned and taught.

Both subordinate clauses and dative prepositional phrases are relatively **frequent** structures in German. As we have discussed above in Section 7.1.2, subordinate clauses are potentially problematic, because there is a growing tendency in spoken German to use coordinating structures instead of subordinating structures. In particular the trend to use *weil* as a coordinating conjunction might have had a negative influence on the learnability of the word order of *weil*-clauses.

Regarding the **developmental readiness** of learners, which manifests in orders of acquisition, the two structures are difficult to compare because developmental orders are usually observed for comparable structures (e.g., word order of different clause types, marking of different cases) rather than between unrelated structures. Even though there is some evidence given by Diehl et al. (2002) that case marking is learned later than word order, the conflicting evidence for the developmental sequence of word order alone, as discussed in Section 4.4.4, cast some doubt on the reliability of such evidence.

Although **salience and functional value** are relevant factors for the general teachability of structures, they have been discussed in particular as being important for the effectiveness of implicit feedback like recasts. For instance, Long et al. (1998) argues that recasts are more effective for for more salient and meaning-bearing structures (as we have discussed in Section 5.5.1). To our knowledge, there is no research that has directly compared the salience of case marking in German determiners and subordinate word order, nor any work that would inform such a comparison. Case marking is not

very salient as we have argued above in Section 7.1.1. The salience of word order in subordinate clauses is unclear.

The fact that the word order in subordinate clauses does not carry any meaning in itself and is a purely formal requirement might be one explanation for why the recast instruction only had a limited effect. In comparison, the dative case marking in prepositional phrases does carry some meaning, and is, in certain cases, necessary to distinguish between the local and the directional meaning of a phrase. However, as we have discussed in Section 7.1.1, in the context of the task scenario we used in our study, phrases in which the case marking is critical for conveying the intended meaning are rare.

Another factor for the differences between the two target structures could be the **prior knowledge** of the participants about the target structures. Since we had no influence on the prior exposure to the structures, and the participants had a diverse range of previous input and instruction, the only crude estimate and way of controlling for previous knowledge of the structure were the pretest results. It was evident that tests for the subordinate clauses showed more participants with perfect scoring on the pretest. Because we excluded the results of these participants, the amount of considered data was smaller. As a result, the remaining sample might have differed in certain aspects from the dative prepositional phrases sample, for instance, they might have been on a lower level in general, and therefore less responsive to instruction.

Tasks

Finally, another important influence in our setup is the suitability of target structures to be used in communicative tasks and the actual tasks. As we have discussed above in Section 4.5.2, the effectiveness of a focused task depends on how natural and necessary the target structure is for the completion of the tasks. The interaction showed that dative prepositional phrases were used much more frequently in the directions giving scenario than subordinate clauses were used in the appointments scenario (see Section 8.2). We argue that it was harder to elicit the use of subordinate clauses, and that their use was not as important for completing the task as dative prepositional phrases were for completing the directions giving task. This confirms the concerns expressed by Pica (1994) that some forms are hard to make relevant in a communicative task. Therefore, communicative tasks as the only means may not suffice for teaching certain forms.

The fact that the participants in the free production condition for the subordinate clause scenario did not use the target structure to the same degree as the participants in the dative prepositional phrase scenario did, may have influenced the development and may be one reason why the learners did not improve to the same degree.

Alongside that point, it is also to be noted that learning gains are only one option to measure the quality of tasks. As we have discussed above in Section 4.5.3, the effect of a single task, even if it is completed several times, may be too subtle to be measured. Ellis (2003) discussed two alternative ways to evaluate tasks – the student-based evaluation for which students are asked if they enjoyed the task and the response-based evaluation which examines if the learners completed the task as expected. Regarding student-based evaluation, the survey showed that learners generally had a mostly positive opinion about the tasks (Section 8.3). In terms of the response-based evaluation,

the majority of the learners completed the tasks with the intended result.

10.2.5 Communicative skill development

According to the arguments put forward for the superiority of FOCUS-ON-FORM opposed to FOCUS-ON-FORMS instruction (Section 4.2), we had reason to expect that the recast group who received instruction that required them to make use of their language in a communicative, meaning-oriented situation showed more development in terms of communicative skills. This expectation was fulfilled for the directions giving task but not for the appointments task. In fact, the two task scenarios entailed complementary developments in fluency. While the appointment making scenario led to greater increase in fluency-related measures for the constrained group, the directions giving scenario induced more recognizable gains for the free-recast group.

It has to be noted that the two tasks differed considerably in the nature of interaction they involved. While in the task for giving directions it was possible to produce relatively long utterances, which were relatively seldom interrupted, unless a clarification question arose, the appointment arranging task required much more back and forth in order to negotiate. Therefore, the nature of the speech samples in each of the two task scenarios differed. Recall that during the editing process the contributions of each partner in a task dyad were separated. For the directions giving task this was relatively easy as there were only few interruptions and overlaps. For the appointment task scenario, there were much more cuts necessary as the dialog was more interactive and turns switched more often.

The higher interactivity and the higher symmetry of roles in the appointments task may have led to different patterns of contribution to the dialog. Some participants may have taken more initiative in the dialog, which might have led to a larger contribution compared to their more passive partner. Since the pairings were not necessarily the same across all test times, some additional variation might have been added through different pairings and different dynamics between them.

For both scenarios it is remarkable that the immediate changes between T1 and T2 were much rarer than changes involving the delayed posttest T3. As with the delayed effects for the grammatical accuracy, it is not clear whether the effect of the treatment was delayed or whether the learners maturated independently in the meantime.

In terms of immediate changes of the temporal measures (between T1 and T2), only the mean length of runs is marginally significantly changing, but it actually decreases for both groups – for the free-recast group, the length of runs including filled pauses decreases, while for the constrained group, the length of runs without filled pauses decreases. This development does not accord with the expectations coming from previous work that showed that longer average length of runs correlate with fluency (Kormos and Dénes, 2004). Kormos and Dénes, however used narrative tasks, whereas our tasks are more oriented to accomplish a goal. Shorter runs in a goal-oriented task may be interpreted as sign for greater efficiency.

An important caveat regarding the holistic ratings of fluency is the difficulty of the rating task, as indicated by the relatively low consistency between the raters and the low internal consistency of two of the raters. Further, the raters gave explicit feedback

saying that they found some samples hard to judge. The difficulty may stem from the relatively small and subtle differences between the tests.

10.3 Limitations

In this section, we discuss the limitations of the current study in more detail and summarize those that were already mentioned in the previous discussion. We will further suggest options for remedying these limitations in future work.

Small sample size

The most important limitation of this study is the small number of participants that was caused by the limited access to participants. The problem was further exacerbated by the considerable drop out of learners, which was facilitated by the fact that attendance in their courses was not compulsory. Small sample sizes lead to a reduced statistical power, as we have discussed above in the introduction of Chapter 9. Mackey and Gass (2005) have argued that the common problem of small samples in second language research may warrant a reconsideration of the common significance level of 0.05 and they have proposed to report findings on a significance level of 0.10 in order to report on important trends that may lead to replication of studies. Following this suggestion, we have included reports on differences of a significance level of 0.10. In future work, with more resources, it would be desirable to recruit a larger pool of participants.

Lack of control group

Another disadvantage of our study is the lack of a true control group with null instruction. As we have argued above in Section 6.6.2, we wanted to make best use of the small number of participants we had access to. Therefore we cannot exclude with certainty that the development would have happened without instruction, solely by maturation, unspecific exposure, or side-effects of the tests. In future extensions or replications of this study it would be desirable to include a true control group.

Self-selection bias

One way that we used to compensate the loss of participants from the language courses and the restricted access we had to the courses was to recruit participants individually. They were comparable to the participants who took part during their course time because they were recruited from the same type of courses which took place a semester later. However, since they were volunteers, they might have been motivated to a higher degree than the average student. This self-selection bias (Heckman, 1979) was not problematic for the additional participants that were recruited for the subordinate clause target structure, since they were equally distributed for across both conditions. For the dative prepositional phrases, however, the additional participants served primarily to add to the metalinguistic feedback group, and made up between

20 to 40 percent, depending on how many sessions and test times are included (see Table 7.2 in Section 7.2). If the voluntary participants of the metalinguistic condition were more highly motivated to learn, their influence is not clearly apparent in the results, as they did not display superior learning results.

Learning environment and control

Since our study was conducted in a second language learning (SLL) environment which, in addition to the lessons, provided learners with considerable exposure to German, it is possible that the learners received relevant input during the course of the experiment. This additional input cannot be controlled or measured and it introduces an additional source of variance (see also Section 6.6.2). Additional exposure also makes it difficult to analyze the processes that figure for long-term effects because it is not clear how exactly instruction and maturation processes interact with the input coming from the environment in developing or maintaining knowledge and skills. Furthermore, the additional input that learners in SLL contexts get outside of class may equalize the differences of the instruction they get, whereas learners in foreign language learning (FLL) contexts are more directly impacted by the instruction and thus differences are more likely to show in the assessment of their learning gains. With regard to that, Li (2010) showed in a meta-analysis that studies conducted in a FLL context yielded larger effect sizes than studies conducted in a SLL context. However, to our knowledge, there is no study that compares these two contexts directly. One possible explanation for this difference is that learners in a FLL context tend to value formal correctness more, whereas learners in a SLL context are more keen on communicative skills, as shown in a survey conducted by Loewen et al. (2009) at a university in the USA. This difference might make FLL learners more willing to integrate corrective feedback.

In future, the study could be replicated in a FLL context, where additional input of the target language is minimal. A FLL context may also provide more control over another source of unwanted variation – the native language of the learner. It is possible that transfer processes may have influenced the results. However, given the variety of first languages, it is hard to further analyze our results under that perspective.

To establish even more control, one option might be to create an artificial language to study. However, this alternative seems unfeasible, since it would require a considerable amount of input to get to a stage at which a communication task could be carried out. Furthermore, the use of artificial languages as such is controversial since the differences between natural and artificial languages cast some doubt on the ecological validity of this paradigm.

Prior exposure

In relation to the problem of uncontrolled input during the experiment, we also have to consider the variation that comes from prior exposure to and learning of the target structures. Since we assumed that some amount of knowledge about the target structures existed and we did not attempt to focus on learners with zero knowledge, learners may have had very different types of instruction for the target structures which can

not be deduced from the pretest scores, but which might influence the effectiveness of the different types of instruction. In Ellis (2010)'s terms, we are considering acquisition as the increase in accuracy of partially acquired features (compare the discussion in Section 5.2.3). Alternatively, we could have studied the acquisition of entirely unknown grammatical structures or characterized the acquisition process as a progress along a sequence of stages. However, there are practical difficulties involved in finding target structures that have not been taught before, and there are only rare examples of studies which attempt this (Long et al., 1998). Furthermore, a sequence of acquisition stages has only be established for a certain subset of structures (Section 4.4.4), which do not necessarily lend themselves well for the focused communicative tasks that we used in the current study. The fact that structures have been acquired to differing degrees before the experiment also makes it difficult to compare the effect of instruction for different target structures directly. Finding two or more structures that were not taught before or that have comparable stages of acquisition is even less promising.

Classroom versus lab

One disadvantage of the decision to conduct the experiment within a computer-equipped classroom is that we were not able to attend to each participant individually. As a consequence, we could not always respond to questions or problems promptly and our control over the execution of the tests and tasks was limited. Occasionally, participants needed support with an exercise because they did not understand it clearly. Sometimes, this resulted in data loss because the learners were unable to complete the tasks as expected. At other times, technical problems with the computer could only be resolved after a delay. While a laboratory setting would have circumvented these problems, it would have been disproportionally more expensive to implement it. However, lab settings are associated with larger effect sizes compared to classroom settings for feedback in general (Li, 2010) and recasts in particular (Nicholas et al., 2001). Similarly to the arguments against the artificial languages paradigm though, one might object that studies in lab settings may impose constraints that make a transfer to more common conditions of language learning, e.g., classrooms or engaging with native speakers in natural contexts questionable.

Measures for development of skills

A critical point for assessing the effects of instructional parameters is the choice of measurements. It is clear that any test measure can only provide an approximation of the learner knowledge as evidenced in the performance on the test.

Grammatical judgment tests are a common and popular means for assessing language knowledge, but they are not perfect. One problem, for instance, is that it is not clear if learners judge the specific structure that was targeted or something else (Ellis, 2004). One solution to this problem is to ask learners to indicate and/or correct the items they judged as incorrect. However, this second step cannot be conducted under time pressure and it is likely to make learners draw on their explicit knowledge. Therefore, it is not really a solution which sheds more light on the implicit knowledge

that worked on the timed test. In our context, asking learners to correct the erroneous items also has the potential to constitute additional practice and make the learners more aware of the target structures.

As we have already addressed in the discussion above, the **holistic rating of the communicative skills** was a difficult task as indicated by the rater feedback as well as the low internal consistency of some raters. Furthermore, additional variation might have been introduced through the uncontrolled pairing of the learners, which may have differed from one test to the next. In a laboratory setting, the dialog partner of the learner could be a neutral examiner who adheres to a fixed protocol when engaging in the dialog to decrease the variation. Kormos and Dénes (2004) and Gass et al. (1999) used narrative tasks as a basis for rating oral skills. Since a narration is a monologic task, there is no communicative partner to introduce additional variation. However, since dialogic interaction is at the core of our setting, it is difficult to find a narrative variant of our tasks that is close enough to the original task. Maybe it is possible in the future to find other communicative tasks that are easier to translate into a narrative task.

In relation to the assessment of communicative skills, it is unfortunate that the experimental conditions did not allow the conducting of an immediate oral posttest. The solution to this would have been to conduct the experiments in a laboratory with one or two learners individually, so that more tests could be conducted without the constraints of the classroom.

Finally, another point to discuss is the **timing of the delayed posttest**. We set it at five weeks after the second treatment session for practical reasons, as we have discussed in Section 7.4. Although this time span lies well in the range of the most common reported time spans, it might be too short. Harley (1989), for instance, found that effects disappeared after three months. To our knowledge, there are no published attempts to compare different spans for delayed posttest and draw conclusions or suggest appropriate intervals for delayed posttests in the context of second language acquisition research. Disregarding that problem, some suggest that the concern about long-term effects may be a little overstated. Long (2007), for instance, argues that initial impacts are the most important for assessing the effectiveness of a treatment. However, considering that we found interesting differences between short- and long-term effects, it would be a worthwhile endeavor to add one or more later delayed test times in future work.

Tasks and elicitability of target structures

A crucial conditions for our approach to work is that communicative tasks can be designed in which certain target forms are essential. A well-designed task makes the use of the target forms likely. As we have discussed above, some learners avoided the use of *weil*-clauses in the appointments scenario, which may have harmed the effect of the task. It is hard to say if it would have been possible to design a task that is more successful in eliciting this particular target structure.

In general, to the best of our knowledge, there is no straight-forward recipe for the design of focused tasks. Instead, the process seems to be build on trial and error and experience. We would argue that some target structures are more difficult than others to elicit in a communicative task. This clearly imposes a limit for the task-based approach, in so far as it is only applicable to a limited set of target structures. Other structures may need to be taught with alternative approaches. *Weil*-clauses, however, come with the additional drawback that there is a pervasive tendency in oral communication to use *weil* as a coordinating conjunction that does not entail subordinate clause word order.

Noticing

We have argued previously that noticing is crucial for learning (Section 4.2.3). We have further noted above in Section 10.2.2 that according to the learner survey, the recast feedback given in response to missing or erroneous *weil*-clauses was not noticed by the learners. Given that implicit forms of instruction and in particular recast feedback are known for being hard to notice, it would be desirable to get a more detailed insight about the noticing processes. This would comprise more sophisticated measures of noticing and a further examination of the factors that support or hamper noticing. This should then be related to the development of knowledge and skills.

System performance

The analysis of the system performance discussed in Section 8.2 shows that the system failed to give appropriate recast feedback in about eight percent of the opportunities and it failed in about 24 percent of the opportunities to give correct metalinguistic feedback. This performance, in particular for metalinguistic feedback, leaves room for improvement. It is possible that more reliable system feedback would have resulted in higher learning gains for the free input conditions.

This chapter discussed the findings of our study and pointed out the limitations arising from the conditions of the context in which we conducted it. In the next chapter we will draw some final conclusions.

11 Concluding Remarks

11.1 Summary of contributions

The goal of this thesis was to explore how language learning can be facilitated through the use of NLP-based ICALL technology. ICALL was realized in the form of a task-based dialog system that provides corrective feedback. We investigated how different parameters of the interaction affect the learning progress. Based on a review of underlying methods and existing comparable ICALL applications, we selected parameters linked to the sophistication and effort required to implement a particular form of interaction, and related them to parameters that are based on open issues from the field of SLA. In this way, we narrowed the focus of the exploration, with the goal of providing a deeper, more precise assessment of the learning gains.

By establishing a tight connection between SLA and ICALL this work contributes to the as yet small field of existing research and development which integrates ICALL and SLA perspectives. In this way, we transfer to a human-computer interaction setting pedagogical concepts that have until now been examined mostly in more traditional human-human settings.

The findings of this thesis indicate that there are small differences in the language skill development afforded by different types of computer-provided instruction. We found that constrained, explicit FOCUS-ON-FORMS instruction in general yields greater immediate learning gains, while free, largely meaning-oriented FOCUS-ON-FORM instruction yields more delayed effects. Similarly, comparing implicit recast feedback with explicit metalinguistic feedback we find that the immediate effects are on par but recast feedback leads to greater delayed effects.

These differences interact considerably with other parameters of the experimental setting, in particular with the selected target structures. Grammatical forms are different in respect to how easy it is to elicit them in a meaning-driven task. This suggests that the effectiveness of certain types of instruction is highly dependent on the

particular goal of the instruction. It also confirms that the use of focused tasks is limited by the propensity of grammatical structures to be natural, useful or essential for a meaning-driven task context. Furthermore, the design of focused tasks even for essential structures is by no means a trivial, straight-forward process but relies heavily on the skills and experience of the task designer. Thus, it seems clear that the task-based approach may have to be combined with other forms of instruction.

Our findings are largely consistent with research results from human-human interaction settings, both with respect to the difference between explicit and implicit instruction in general as well as with respect to the comparison of recast and metalinguistic feedback in particular. This is consistent with the findings presented by Petersen (2010), who found that recasts provided in a type-written ICALL interaction were as effective as recasts provided in oral teacher-learner interaction. Both findings suggest that the differences between human-computer interaction and human-only interaction do not bring about vastly different conditions for language learning, at least not in particular contexts. This means that we may assume that other, sufficiently similar SLA research results that originate from human-human interaction may lead to comparable results if they were reproduced in a human-computer setting.

However, considering the fact that the communicative skills of an artificial system are in many ways still not comparable to human performance, this transfer is limited to the range of instructional settings that do not depend on the high level of human performance. To identify the particular instructional conditions, which allow for learning through limited, not quite human-like, but still complex and entertaining performance is a worthwhile goal.

The superior long term effects of meaning-oriented, more implicit instruction with free input can be used to justify the more expensive development of systems that afford such instruction compared to simpler, more explicit accuracy-focused drill-like activities. While our results and previous work show that drills enable faster learning, they also show that the learning gains are not as sustainable.

However, our results do not warrant the abolition of the use of relatively simple interactive drill activities in general. Embedding them into a meaningful context instead of providing them as decontextualized items can further help to make such activities more engaging. In fact, according to the usability ratings of our system, which included enjoyment, perceived usefulness, and likelihood of future usage, the drill-like nature of the constrained conditions was not perceived more negatively than the free input system. Possible reasons for this similarity in ratings are our efforts to keep the rest of the context similar, or the existing flaws of the free input system, which may have caused some dissatisfaction.

Our positive results for all three types of ICALL instruction are consistent with the findings by Grgurović et al. (2013), who showed in a meta-analysis that CALL applications (comprising simple as well as intelligent CALL), were at least as effective as instruction without technology and superior in studies using strictly controlled designs.

Thus, we conclude that both simple and more advanced approaches to CALL are justified and effective means of supporting language learning. While more advanced sophisticated approaches that draw closer to aspects of human performance may be

more entertaining and more beneficial for sustained learning gains, their effect hinges on largely flawless performance which require extensive development efforts. Thus, from the perspective of cost-benefit analysis, cheaper approaches continue to have their place.

Also within the area of the more sophisticated approaches which provide feedback, there are different grades of sophistication that need to be carefully deployed. In our example, the provision of recasts does not require near-perfect error recognition because recasts are not harmful when produced in response to correct input, as they could be interpreted as regular acknowledging grounding moves. Basically, they do not claim that the learner's utterance was erroneous. Metalinguistic feedback, or other more explicit corrective feedback types, on the other hand can be more confusing and harmful if they are produced in response to a correct learner utterance. Thus, the parameters of ICALL interaction should be adapted according to the confidence on error recognition in order to avoid harm for the learner.

11.2 Outlook

The results of this thesis can be used as a basis for further research. The potential future directions of our work fall into two different strands. One regards the exploration and comparison of further pedagogical parameters, the other is related to implementational issues.

For the first strand, additional types and variants of feedback can be examined. First, it would be interesting to add other prevalent types of feedback to the investigation. Second, recasts could be enhanced in different ways in order to increase their noticeability. The effect of such enhanced recasts could then be compared with regular recasts both in terms of learning gains but also in terms of how they are perceived using more fine-grained assessments of noticing. Third, more versions of metalinguistic feedback could be realized and compared. Possible variants could provide the correct form or a more detailed linguistic explanation of the structure.

From the implementational perspective then, it may be worthwhile to try to reason about the misconceptions or gaps in knowledge that caused the error and adapt the feedback accordingly. However, this is a complex problem and probably only feasible for very well-defined narrow error types. In a similar vein, it would be useful to model confidence measures for error diagnosis, which could be assigned to any interpretation of learner input, basically coding how sure the system is that a particular utterance is accurate, erroneous, or possibly not covered by the interpretation grammar. Confidence levels could then be used to select the optimal feedback, balancing the potential harm of unwarranted corrections with the harm of missed opportunities for corrections.

For our study, we chose depth over breadth and examined three relatively narrow instances of instruction. An alternative approach would be to examine a much wider range of possible ICALL parameters but to evaluate them in less depth regarding their pedagogic effect. Parameters could cover a broader scale of feedback types and variants, including different degrees of informativity or explicitness. Parameters could also be expanded such that they create more levels of constraint on learner in-

put. Alternative, shallower types of evaluations could include learner questionnaires on usability and user experience, or an analysis of interactions patterns.

An expansion of parameters may also require more advanced and elaborate approaches to providing dialog interaction and feedback. For the purpose of our study, a comparatively simple implementation was sufficient. However, a more general and more flexible approach may require to come closer to or even surpass the limits of the current state of the art.

Parallel to the expansion of interaction parameters, one might also pursue to extend the existing work to include other target structures and tasks. The extension could also cover other levels of linguistic knowledge, e.g., pronunciation or pragmatics. This would serve the practical purpose of providing a more comprehensive collection of instruction material for a wider population of learners. However, at the same time, it opens the opportunity to gain theoretical insights into the constraints and prerequisites for applying our approach to wider areas.

In conclusion, we recommend that efforts in evaluating ICALL applications should always take into consideration existing research results and open issues in the field of SLA. By turning a blind eye to the achievements and issues of a discipline that is so clearly relevant, any efforts in ICALL run the risk of becoming a mere boast of engineering accomplishments irrelevant to actual pedagogical requirements. As we have illustrated in the review of existing systems, while several ICALL developments are based on SLA concepts, rigorous evaluations of learning progress along the lines of SLA experiments are still relatively rare.

Beyond the concrete contributions to the specific SLA issues that we examined, the more general contribution of this work lies in connecting the three disciplines SLA, NLP, and ICALL in a principled way. The approach and methodology of our study can thus serve as a framework and paradigm for further examinations of SLA within an ICALL context.

Bibliography

Abrams, Zsuzsanna Ittzes (2003). The Effect of Synchronous and Asynchronous CMC on Oral Performance in German. *The Modern Language Journal*, 87(2): 157–167.

Allen, James and Mark Core (1997). Draft of DAMSL: Dialog Act Markup in Several Layers. Unpublished manuscript.

Allen, James, George Ferguson, Bradford W. Miller, Eric K. Ringger and Teresa Sikorski Zollo (2000). Dialogue Systems: From Theory to Practice in TRAINS-96. Robert Dale, Hermann Moisl and Harold Somers (Editors), *Handbook of Natural Language Processing*, Marcel Dekker, 347–376.

Allen, James F., Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent (2001). Toward Conversational Human-Computer Interaction. *AI Magazine*, 22(4): 27–37.

Allen, James F. and C. Raymond Perrault (1980). Analyzing Intention in Utterances. *Artificial Intelligence*, 15(3): 143–178.

Allwood, Jens, Joakim Nivre and Elisabeth Ahlsén (1992). On the Semantics and Pragmatics of Linguistic Feedback. *Journal of Semantics*, 9(1): 1–26.

Amaral, Luiz (2007). *Designing Intelligent Language Tutoring Systems for Integration into Foreign Language Instruction*. Ph.D. thesis, Ohio State University, Columbus, OH, USA.

Amaral, Luiz and Detmar Meurers (2009). Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL. *CALICO Journal*, 26(3): 580–591.

Amaral, Luiz and Detmar Meurers (2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL*, 23(01): 4–24.

Amaral, Luiz, Detmar Meurers and Ramon Ziai (2011). Analyzing Learner Language: Towards a Flexible NLP Architecture for Intelligent Language Tutors. *Computer Assisted Language Learning*, 24(1): 1–16.

Ammar, Ahlem and Nina Spada (2006). One Size Fits it All?: Recasts, Prompts, and L2 Learning. *Studies in Second Language Acquisition*, 28(04): 543–574.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson and Regina Weinert (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4): 351–366.

Anderson, James N., Nancie Davidson, Hazel Morton and Mervyn A. Jack (2008). Language Learning with Interactive Virtual Agent Scenarios and Speech Recognition: Lessons Learned. *Computer Animation and Virtual Worlds*, 19(5): 605–619.

Anderson, John R. (1983). *The Architecture of Cognition*. Harvard University Press.

Andringa, Sible (2005). *The Effect of Form-Focused Instruction on Second Language Proficiency*. Ph.D. thesis, Rijksuniversiteit Groningen.

Austin, John L. (1962). *How to Do Things with Words*. Harvard University Press.

Barnett, Jim, Rahul Akolkar, RJ Auburn, Michael Bodell, Daniel C. Burnett, Jerry Carter, Scott McGlashan, Torbjörn Lager, Mark Helbing, Rafah Hosn, T.V. Raman, Klaus Reifenrath, No'am Rosenthal and Johan Roxendal (2012). State Chart XML (SCXML): State Machine Notation for Control Abstraction - W3C Last Call Working Draft 29 May 2014. [Online]. (URL http://www.w3.org/TR/scxml/). (Accessed 27.08.2014).

Beal, Carole R., W. Lewis Johnson, Richard Dabrowski and Shumin Wu (2005). Individualized Feedback and Simulation-Based Practice in the Tactical Language Training System: An Experimental Evaluation. *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED 2005*, 747–749.

Benson, Phil (2001). *Teaching and Researching Autonomy in Language Learning*. Pearson Education.

Beretta, Alan and Alan Davies (1985). Evaluation of the Bangalore Project. *ELT Journal*, 39(2): 121–127.

Bialystok, Ellen (1979). Explicit and Implicit Judgements of L2 Grammaticality. *Language Learning*, 29(1): 81–103.

Bialystok, Ellen (1994). Representation and Ways of Knowing: Three Issues in Second Language Acquisition. Ellis (1994a), 549—569.

Bierwisch, Manfred (1963). *Grammatik des Deutschen Verbs*. Studia Grammatica, Berlin: Akademie Verlag.

Bley-Vroman, Robert W., Sascha W. Felix and Georgette L. loup (1988). The Accessibility of Universal Grammar in Adult Language Learning. *Second Language Research*, 4(1): 1–32.

Bos, Johan, Ewan Klein, Oliver Lemon and Tetsushi Oka (2003). DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, 115–124.

de Bot, Kees (1996). The Psycholinguistics of the Output Hypothesis. *Language Learning*, 46(3): 529–555.

Boyd, Adriane Amelia (2012). *Detecting And Diagnosing Grammatical Errors for Beginning Learners of German: From Learner Corpus Annotation to Constraint Satisfaction Problems*. Ph.D. thesis, The Ohio State University.

Breen, Michael P. and Christopher N. Candlin (1980). The Essentials of a Communicative Curriculum in Language Teaching. *Applied Linguistics*, 1(2): 89–112.

Brown, John Seely and Richard R. Burton (1975). Multiple Representations of Knowledge for Tutorial Reasoning. Daniel Gureasko Bobrow and Allan Collins (Editors), *Representation and Understanding: Studies in Cognitive Science*, Academic Press, 311–350.

Brumfit, Christopher (1984). *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge University Press.

Bump, Jerome (1990). Radical Changes in Class Discussion Using Networked Computers. *Computers and the Humanities*, 24(1-2): 49–65.

Bunt, Harry (2000). Dynamic Interpretation and Dialogue Theory. Martin M. Taylor, Franççoise Néel and Don Bouwhuis (Editors), *The Structure of Multimodal Dialogue II*, John Benjamins Publishing, 139–166.

Bussmann, Hadumod (1998). *Routledge Dictionary of Language and Linguistics*. Routledge.

Buyer, Linda S. and Roger L. Dominowski (1989). Retention of Solutions: It Is Better to Give than to Receive. *The American Journal of Psychology*, 102(3): 353–363.

Caines, Andrew and Paula J. Buttery (2014). The Effect of Disfluencies and Learner Errors on the Parsing of Spoken Learner Language. *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, 74–81.

Carroll, Susanne and Merrill Swain (1993). Explicit and Implicit Negative Feedback: An Empirical Study of the Learning of Linguistic Generalizations. *Studies in Second Language Acquisition*, 15(3): 357–86.

Chafe, Wallace and Jane Danielewicz (1987). Properties of Spoken and Written Language. Rosalind Horowitz and S. Jay Samuels (Editors), *Comprehending Oral and Written Language*, Academic Press, 83–113.

Chan, Wai Meng and Dong-Ha Kim (2004). Towards Greater Individualization and Process-Oriented Learning through Electronic Self-Access: Project "e-daf". *Computer Assisted Language Learning*, 17(1): 83–108.

Chandler, Jean (2003). The Efficacy of Various Kinds of Error Feedback for Improvement in the Accuracy and Fluency of L2 Student Writing. *Journal of Second Language Writing*, 12(3): 267–296.

Chapelle, Carol (2001). *Computer Applications in Second Language Acquisition*. Cambridge University Press.

Chaudron, Craig (1988). *Second Language Classrooms: Research on Teaching and Learning*. Cambridge University Press.

Chiu, Scott Chien-Hsiung (2008). Review of ActiveChinese: Chinese Language Skills for the Business World. *Language Learning & Technology*, 12(3): 23–32.

Chiu, Tsuo-Lin, Hsien-Chin Liou and Yuli Yeh (2007). A Study of Web-Based Oral Activities Enhanced By Automatic Speech Recognition for EFL College Learning. *Computer Assisted Language Learning*, 20(3): 209–233.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press.

Christiansen, Morten H. and Nick Chater (1999). Connectionist Natural Language Processing: The State of the Art. *Cognitive Science*, 23(4): 417–437.

Chun, Dorothy M. (1994). Using Computer Networking to Facilitate the Acquisition of Interactive Competence. *System*, 22(1): 17–31.

Clahsen, Harald (1984). The Acquisition of German Word Order: A Test Case for Cognitive Approaches to Second Language Acquisition. Roger Andersen (Editor), *Second Languages: A Cross-Linguistic Perspective*, Newbury House, 219—-242.

Clahsen, Harald and Pieter Muysken (1986). The Availability of Universal Grammar to Adult and Child Learners - A Study of the Acquisition of German Word Order. *Second Language Research*, 2(2): 93–119.

Clark, Alexander and Shalom Lappin (2011). Computational Learning Theory and Language Acquisition. Ruth Kempson, Tim Fernando and Nicholas Asher (Editors), *Handbook of the Philosophy of Science. Volume 14: Philosophy of Linguistics*, Elsevier.

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

Clark, Herbert H. and Edward F. Schaefer (1989). Contributing to Discourse. *Cognitive Science*, 13(2): 259–294.

Clay, Felix, Jeffrey S. Bowers, Colin J. Davis and Derek A. Hanley (2007). Teaching Adults New Words: The Role of Practice and Consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5): 970–976.

Cohen, Philip R. and C. Raymond Perrault (1979). Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science*, 3(3): 177–212.

Colby, Kenneth Mark (1975). *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. Elsevier.

Colby, Kenneth Mark (1981). Modeling a Paranoid Mind. *Behavioral and Brain Sciences*, 4(04): 515–534.

Coniam, David (2008). Evaluating the Language Resources of Chatbots for Their Potential in English as a Second Language. *ReCALL*, 20(01): 98–116.

Cook, Vivian (2003). The Poverty-of-the-Stimulus Argument and Structure-Dependency in L2 Users of English. *IRAL - International Review of Applied Linguistics in Language Teaching*, 41(3): 201–221.

Copestake, Ann (2002). *Implementing Typed Feature Structure Grammars*, volume 110. Stanford: CSLI Publications.

Copestake, Ann, Dan Flickinger, Carl Pollard and Ivan A. Sag (2005). Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(2-3): 281–332.

Coppola, Bonaventura, Alessandro Moschitti and Giuseppe Riccardi (2009). Shallow Semantic Parsing for Spoken Language Understanding. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 85–88.

Core, Mark and James Allen (1997). Coding Dialogs with the DAMSL Annotation Scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, 28–35.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Strasbourg: Council of Europe.

Craik, Fergus I. M. and Robert S. Lockhart (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, 11(6): 671–684.

Curme, George O. (1970). *A Grammar of the German Language*. Ungar. Second Revised Edition.

Curtin, Constance, Douglas Clayton, Cheryl Finch, David Moor and Lois Woodruff (1972). Teaching the Translation of Russian by Computer. *The Modern Language Journal*, 56(6): 354–360.

Davis, Matthew H., Anna Maria Di Betta, Mark J.E. Macdonald and M. Gareth Gaskell (2009). Learning and Consolidation of Novel Spoken Words. *Journal of Cognitive Neuroscience*, 21(4): 803–820.

De Felice, Rachele (2008). *Automatic Error Detection in Non-Native English*. Ph.D. thesis, University of Oxford.

De Felice, Rachele and Stephen G. Pulman (2008). A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, United Kingdom.

De Mori, Renato, Frédéric Bechet, Dilek Hakkani-Tur, Michael McTear, Giuseppe Riccardi and Gokhan Tur (2008). Spoken Language Understanding. *Signal Processing Magazine, IEEE*, 25(3): 50–58.

DeKeyser, Robert (1998). Beyond Focus on Form: Cognitive Perspectives on Learning and Practicing Second Language Grammar. Doughty and Williams (1998a), 42–63.

DeKeyser, Robert (2007a). Skill Acquisition Theory. Jessica Williams and Bill VanPatten (Editors), *Theories in Second Language Acquisition: An Introduction*, Lawrence Erlbaum, 97–113.

DeKeyser, Robert (2008). Implicit and Explicit Learning. Catherine J. Doughty and Michael H. Long (Editors), *The Handbook of Second Language Acquisition*, Blackwell Publishing.

DeKeyser, Robert M. (1995). Learning Second Language Grammar Rules: An Experiment With a Miniature Linguistic System. *Studies in Second Language Acquisition*, 17(03): 379–410.

DeKeyser, Robert M. (1997). Beyond Explicit Rule Learning. *Studies in Second Language Acquisition*, 19(02): 195–221.

DeKeyser, Robert M. (Editor) (2007b). *Practice in a Second Language: Perspectives from Applied Linguistics and Cognitive Psychology*. Cambridge University Press.

Dennett, Daniel C. (1998). Can Machines Think? Daniel C. Dennett (Editor), *Brainchildren: Essays on Designing Minds*, MIT Press.

DeSmedt, William H. (1995). Herr Kommissar: An ICALL Conversation Simulator for Intermediate German. Holland et al. (1995).

Diehl, Erika, Hannelore Pistorius and Annie Fayolle Dietl (2002). Grammatikerwerb im Fremdsprachenunterricht - ein Widerspruch an sich? Wolfgang Börner and Klaus Vogel (Editors), *Grammatik und Fremdsprachenerwerb. Kognitive, psycholinguistische und erwerbstheoretische Perspektiven*, Tübingen: Narr, 143–163.

Diekelmann, Susanne, Ines Wilhelm and Jan Born (2009). The Whats and Whens of Sleep-Dependent Memory Consolidation. *Sleep Medicine Reviews*, 13(5): 309–321.

Doughty, Catherine (1991). Second Language Instruction Does Make a Difference - Evidence from an Empirical Study of SL Relativization. *Studies in Second Language Acquisition*, 13(04): 431–469.

Doughty, Catherine (1994). Fine-Tuning of Feedback by Competent Speakers to Language Learners. James E. Alatis (Editor), *Georgetown University Roundtable on Language and Linguistics (GURT) 1993: Strategic Interaction and Language Acquisition*, Georgetown University Press, 96–108.

Doughty, Catherine and Elizabeth Varela (1998). Communicative Focus on Form. Doughty and Williams (1998a), 114–138.

Doughty, Catherine and Jessica Williams (Editors) (1998a). *Focus on Form in Classroom Second Language Acquisition*. Cambridge University Press.

Doughty, Catherine J. (2003). Instructed SLA: Constraints, Compensation, and Enhancement. Catherine J. Doughty and Michael H. Long (Editors), *The Handbook of Second Language Acquisition*, Blackwell Publishing, 256–310.

Doughty, Cathy and Jessica Williams (1998b). Issues and Terminology. Doughty and Williams (1998a), 1–11.

Doughty, Cathy and Jessica Williams (1998c). Pedagogical Choices in Focus on Form. Doughty and Williams (1998a), 197–261.

Dulay, Heidi C. and Marina K. Burt (1973). Should We Teach Children Syntax? *Language Learning*, 23(2): 245–258.

Eckerth, Johannes, Karen Schramm and Erwin Tschirner (2009). Review of Recent Research (2002-2008) on Applied Linguistics and Language Teaching with Specific Reference to L2 German (Part 1). *Language Teaching*, 42(01): 41–66.

Ehsani, Farzad and Eva Knodt (1998). Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. *Language Learning & Technology*, 2(1): 45–60.

Eisenberg, Peter (1999). *Grundriß der deutschen Grammatik. Der Satz*. Stuttgart: Metzler.

Ellis, Nick C. (Editor) (1994a). *Implicit and Explicit Learning of Languages*. Academic Press.

Ellis, Nick C. (1994b). Introduction: Implicit and Explicit Language Learning - An Overview. Ellis (1994a), 1—31.

Ellis, Nick C. (2002). Frequency Effects in Language Processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24(02): 143–188.

Ellis, Rod (1984). Can Syntax be Taught? A Study of the Effects of Formal Instruction on the Acquisition of WH Questions by Children. *Applied Linguistics*, 5(2): 138–155.

Ellis, Rod (1986). *Understanding Second Language Acquisition*. Oxford University Press. 2nd, improved edition.

Ellis, Rod (1989). Are Classroom and Naturalistic Acquisition the Same? A Study of the Classroom Acquisition of German Word Order Rules. *Studies in Second Language Acquisition*, 11(03): 305–328.

Ellis, Rod (1991). Grammatically Judgments and Second Language Acquisition. *Studies in Second Language Acquisition*, 13(02): 161–186.

Ellis, Rod (1994c). A Theory of Instructed Second Language Acquisition. Ellis (1994a), 79–114.

Ellis, Rod (2001). Introduction: Investigating Form-Focused Instruction. *Language Learning*, 51(s1): 1–46.

Ellis, Rod (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Ellis, Rod (2004). The Definition and Measurement of L2 Explicit Knowledge. *Language Learning*, 54(2): 227—-275.

Ellis, Rod (2005). Measuring Implicit and Explicit Knowlegde of a Second Language: A Psychometric Study. *Studies in Second Language Acquisition*, 27(02): 141–172.

Ellis, Rod (2006). Modelling Learning Difficulty and Second Language Proficiency: The Differential Contributions of Implicit and Explicit Knowledge. *Applied Linguistics*, 27(3): 431–463.

Ellis, Rod (2009a). Implicit and Explicit Learning, Knowledge and Instruction. Ellis et al. (2009), 3–25.

Ellis, Rod (2009b). Measuring Implicit and Explicit Knowledge of a Second Language. Ellis et al. (2009), 31–64.

Ellis, Rod (2010). Epilogue - A Framework for Investigating Oral and Written Corrective Feedback. *Studies in Second Language Acquisition*, 32(Special Issue 02): 335–349.

Ellis, Rod, Shawn Loewen, Catherine Elder, Rosemary Erlam, Jenefer Philp and Hayo Reinders (2009). *Implicit And Explicit Knowledge In Second Language Learning, Testing And Teaching*. Multilingual Matters.

Ellis, Rod, Shawn Loewen and Rosemary Erlam (2006). Implicit and Explicit Corrective Feedback and the Acquisition of L2 Grammar. *Studies in Second Language Acquisition*, 28: 339–368.

Ellis, Rod and Younghee Sheen (2006). Reexamining the Role of Recasts in Second Language Acquisition. *Studies in Second Language Acquisition*, 28(04): 575–600.

Elmauer, Ute and Rolf Müller (1974). Belegung der Freiburger Forschungshypothese über die Beziehung zwischen Redekonstellation und Textsorte. *Sprache der Gegenwart*, 26(Gesprochene Sprache. Jahrbuch 1972.): 98–120.

Engel, Ulrich (1974). Syntaktische Besonderheiten der deutschen Alltagssprache. Hugo Moser (Editor), *Gesprochene Sprache*, Pädagogischer Verlag Schwann.

Erlam, Rosemary (2006). Elicited Imitation as a Measure of L2 Implicit Knowledge: An Empirical Validation Study. *Applied Linguistics*, 27(3): 464–491.

Ervin-Tripp, Susan (1979). Children's Verbal Turn-Taking. Bambi B. Schieffelin and Elinor Ochs (Editors), *Developmental Pragmatics*, Academic Press, 391–414.

Eskenazi, Maxine (2009). An Overview of Spoken Language Technology for Education. *Speech Communication*, 51(10): 832–844.

Fauth, Camille, Anne Bonneau, Frank Zimmerer, Jürgen Trouvain, Bistra Andreeva, Vincent Colotte, Dominique Fohr, Denis Jouvet, Jeanin Jügler, Yves Laprie, Odile Mella and Bernd Möbius (2014). Designing a Bilingual Speech Corpus for French and German Language Learners: A Two-Step Process. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*.

Feigenbaum, Edward A. (2003). Some Challenges and Grand Challenges for Computational Intelligence. *Journal of the ACM*, 50(1): 32–40.

Ferguson, George and James Allen (2005). Mixed-Initiative Dialogue Systems for Collaborative Problem-Solving. David W. Aha and Gheorghe Tecuci (Editors), *Mixed-Initiative Problem-Solving Assistants: Papers from the 2005 AAAI Fall Symposium*, 57–62.

Ferreira, Anita, Johanna D. Moore and Chris Mellish (2007). A Study of Feedback Strategies in Foreign Language Classrooms and Tutorials with Implications for Intelligent Computer-Assisted Language Learning Systems. *International Journal of Artificial Intelligence in Education*, 17(4): 389–422.

Ferris, Dana (1999). The Case for Grammar Correction in L2 Writing Classes: A Response to Truscott (1996). *Journal of Second Language Writing*, 8(1): 1–11.

Ferris, Dana R. (2004). The "Grammar Correction" Debate in L2 Writing: Where are we, and where do we go from here? (and what do we do in the meantime ...?). *Journal of Second Language Writing*, 13(1): 49–62.

Fillmore, Charles J (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1): 20–32.

Fitze, Michael (2006). Discourse and Participation in ESL Face-to-Face and Written Electronic Conferences. *Language Learning & Technology*, 10(1): 67–86.

Folsom, Marvin H. (1981). Four Approaches to the Dative/Accusative Prepositions. *Die Unterrichtspraxis / Teaching German*, 14(2): 222–231.

Folsom, Marvin H. (1984). Prepositions with the Dative or Accusative in Written and Spoken German. J. Alan Pfeffer (Editor), *Studies in Descriptive German Grammar*, Heidelberg: Groos, 19–32.

Foth, Kilian, Michael Daum and Wolfgang Menzel (2004). A Broad-coverage Parser for German Based on Defeasible Constraints. *KONVENS 2004, Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache*, 45–52.

Friedman, Milton (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200): 675–701.

Galili, Tal (2010). Post hoc analysis for Friedman's Test (R code). [Online]. (URL http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code). (Accessed 27.08.2014).

Gamper, Johann and Judith Knapp (2002). A Review of Intelligent CALL Systems. *Computer Assisted Language Learning*, 15(4): 329–342.

Gass, Susan, Alison Mackey, María José Alvarez-Torres and Marisol Fernández-García (1999). The Effects of Task Repetition on Linguistic Output. *Language Learning*, 49(4): 549–581.

Gass, Susan, Ildikó Svetics and Sarah Lemelin (2003). Differential Effects of Attention. *Language Learning*, 53(3): 497–546.

Gass, Susan M. (1997). *Input, Interaction, and the Second Language Learner*. Lawrence Erlbaum.

Gildea, Daniel and Daniel Jurafsky (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3): 245–288.

Glaboniat, Manuela, Martin Müller and Paul Rusch (2005). *Profile Deutsch*. Langenscheidt.

Gohl, Christine and Susanne Günthner (1999). Grammatikalisierung von weil als Diskursmarker in der gesprochenen Sprache. *Zeitschrift für Sprachwissenschaft*, 18(1): 39–75.

Goldschneider, Jennifer M. and Robert M. DeKeyser (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-Analysis of Multiple Determinants. *Language Learning*, 51(1): 1–50.

Goronzy, Silke (2004). Generating Non-Native Pronunciation Variants for Lexicon Adaptation. *Speech Communication*, 42(1): 109–123.

Gove, Philip Babcock et al. (Editors) (1993). *Webster's Third New International Dictionary of the English Language, Unabridged: A Merriam-Webster*. Springfield, Massachusetts: Merriam-Webster.

Granqvist, Svante (2003). The Visual Sort and Rate Method for Perceptual Evaluation in Listening Tests. *Logopedics, Phoniatrics, Vocology*, 28(3): 109–116.

Green, Peter S. and Karlheinz Hecht (1992). Implicit and Explicit Grammar:An Empirical Study. *Applied Linguistics*, 13(2): 168–184.

Grey, Sarah, John N. Williams and Patrick Rebuschat (2014). Incidental Exposure and L3 Learning of Morphosyntax. *Studies in Second Language Acquisition*, FirstView: 1–35.

Grgurović, Maja, Carol A. Chapelle and Mack C. Shelley (2013). A Meta-Analysis of Effectiveness Studies on Computer Technology-Supported Language Learning. *ReCALL*, 25(02): 165–198.

Grice, Paul H. (1975). Logic and Conversation. *Syntax and Semantics*, 3: 41–58.

Günthner, Susanne (1996). From Subordination to Coodination? Verb-Second Position in German Causal and Concessive Constructions. *Pragmatics*, 6(3): 323–371.

Günthner, Susanne (2008). 'weil – es ist zu spät'. Geht die Nebensatzstellung im Deutschen verloren? Markus Denkler, Susanne Günthner, Wolfgang Imo, Jürgen Macha, Dorothee Meer, Benjamin Stoltenburg and Elvira Topalovic (Editors), *Frischwärts und Unkaputtbar. Sprachverfall oder Sprachwandel im Deutschen?*, Münster: Aschendorff, 103–128.

Gusfield, Dan (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Haag, Winfried (1985). An Analytical Approach to the Teaching of Spoken German. Gordon und Brian Griffiths Doble (Editor), *Oral Skills in the Modern Languages Degree*, Centre for Information on Language Teaching and Research, 51–72.

Han, Youngju and Rod Ellis (1998). Implicit Knowledge, Explicit Knowledge and General Language Proficiency. *Language Teaching Research*, 2(1): 1–23.

Han, Zhaohong (2002). A Study of the Impact of Recasts on Tense Consistency in L2 Output. *TESOL Quarterly*, 36(4): 543–572.

Handley, Zöe and Marie-Josée Hamel (2005). Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (CALL). *Language Learning & Technology*, 9(3): 99–120.

Handley, Zöe (2009). Is text-to-speech synthesis ready for use in computer-assisted language learning? *Speech Communication*, 51(10): 906 – 919. Spoken Language Technology for Education Spoken Language.

Harel, David (1987). Statecharts: A Visual Formalism for Complex Systems. *Science of Computer Programming*, 8(3): 231–274.

Harless, William, Marcia Zier and Robert Duncan (1999). Virtual Dialogues with Native Speakers: The Evaluation of an Interactive Multimedia Method. *CALICO Journal*, 16(3): 313–37.

Harless, William G., Marcia A. Zier, Michael G. Harless and Robert C. Duncan (2003). Virtual Conversations: An Interface to Knowledge. *IEEE Computer Graphics and Applications*, 23(5): 46–52.

Harley, Birgit (1989). Functional Grammar in French Immersion: A Classroom Experiment. *Applied Linguistics*, 10(3): 331–360.

Harris, Randy Allen (2005). *Voice Interaction Design: Crafting the New Conversational Speech Systems*. Elsevier.

Hatch, Evelyn (1978). Acquisition of Syntax in a Second Language. Jack C. Richards (Editor), *Understanding Second and Foreign Language Learning*, Newbury House, 34–70.

He, Yulan and Steve Young (2005). Semantic Processing Using the Hidden Vector State Model. *Computer Speech & Language*, 19(1): 85–106.

He, Yulan and Steve Young (2006). Spoken Language Understanding Using the Hidden Vector State Model. *Speech Communication*, 48(3): 262–275.

Heckman, James J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1): 153–161.

Hedgcock, John (1993). Well-Formed vs. Ill-Formed Strings in L2 Metalingual Tasks: Specifying Features of Grammaticality Judgements. *Second Language Research*, 9(1): 1–21.

Heift, Trude (2001a). Error-Specific and Individualised Feedback in a Web-Based Language Tutoring System: Do they read it? *ReCALL*, 13(01): 99–109.

Heift, Trude (2001b). Intelligent Language Tutoring Systems for Grammar Practice. *Zeitschrift für Interkulturellen Fremdsprachenunterricht (Online)*, 6.

Heift, Trude (2003). Multiple Learner Errors and Meaningful Feedback: A Challenge for ICALL Systems. *CALICO Journal*, 20(3): 533–548.

Heift, Trude (2004). Corrective Feedback and Learner Uptake in CALL. *ReCALL*, 16(02): 416–431.

Heift, Trude (2010a). Developing an Intelligent Language Tutor. *CALICO Journal*, 27(3): 443–459.

Heift, Trude (2010b). Prompting in CALL: A Longitudinal Study of Learner Uptake. *The Modern Language Journal*, 94(2): 198–216.

Heift, Trude and Devlan Nicholson (2001). Web Delivery of Adaptive and Interactive Language Tutoring. *International Journal of Artificial Intelligence in Education*, 12(4): 310–325.

Heift, Trude and Mathias Schulze (2007). *Errors and Intelligence in Computer-Assisted Language Learning. Parsers and Pedagogues*. New York: Routledge.

Holland, V. Melissa, Jonathan D. Kaplan and Mark A. Sabol (1999). Preliminary Tests of Language Learning in a Speech-Interactive Graphics Microworld. *CALICO Journal*, 16(3): 339–359.

Holland, V. Melissa, Michelle R. Sams and Jonathan D. Kaplan (Editors) (1995). *Intelligent Language Tutors: Theory Shaping Technology*. Lawrence Erlbaum.

Hollander, Myles and Douglas A. Wolfe (1999). *Nonparametric Statistical Methods*. New York: John Wiley & Sons.

Housen, Alex and Folkert Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4): 461–473.

Housen, Alex and Michel Pierrard (2005a). Investigating Instructed Second Language Acquisition. Housen and Pierrard (2005b), 1–27.

Housen, Alex and Michel Pierrard (Editors) (2005b). *Investigations in Instructed Second Language Acquisition*. Berlin: Mouton de Gruyter.

Housen, Alex, Michel Pierrard and Siska Van Daele (2005). Structure Complexity and the Efficacy of Explicit Grammar Instruction. Housen and Pierrard (2005b), 235–270.

Hulstijn, Jan (2002). Towards a Unified Account of the Representation, Processing and Acquisition of Second Language Knowledge. *Second Language Research*, 18(3): 193–223.

Hulstijn, Jan H. and Rick de Graff (1994). Under What Conditions Does Explicit Knowledge of a Second Language Facilitate the Acquisition of Implicit Knowledge? A Research Proposal. *AILA Review*, 11: 97–112.

Hurtado, Lluís F., David Griol, Emilio Sanchis and Encarna Segarra (2005). A Stochastic Approach to Dialog Management. *Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding – ARSU 2005*, 226–231.

Izumi, Shinichi (2002). Output, Input Enhancement, and the Noticing Hypothesis: An Experimental Study on ESL Relativization. *Studies in Second Language Acquisition*, 24(04): 541–577.

Jacoby, Larry L. (1978). On Interpreting the Effects of Repetition: Solving a Problem Versus Remembering a Solution. *Journal of Verbal Learning and Verbal Behavior*, 17(6): 649 – 667.

Jehle, Fred (1987). A Free-Form Dialog Program in Spanish. *CALICO Journal*, 5(2): 11–22.

Johnson, Jacqueline S. and Elissa L. Newport (1989). Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language. *Cognitive Psychology*, 21(1): 60–99.

Johnson, Mark (1994). Two Ways of Formalizing Grammars. *Linguistics and Philosophy*, 17(3): 221–248.

Johnson, W. Lewis and Carole Beal (2005). Iterative Evaluation of a Large-Scale, Intelligent Game for Language Learning. *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED 2005*, 290–297.

Johnson, W. Lewis, Sunhee Choi, Stacy Marsella, Nicolaus Mote, Shrikanth Narayanan and Hannes Vilhjálmsson (2004a). Tactical Language Training System: Supporting the Rapid Acquisition of Foreign Language and Cultural Skills. *Proceedings of InSTIL/ICALL 2004 – Computer Assisted Learning, NLP and Speech Technologies in Advanced Language Learning Systems*.

Johnson, W. Lewis, Stacy Marsella and Hannes Vilhjaálmsson (2004b). The DARWARS Tactical Language Training System. *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Johnson, W. Lewis and Andre Valente (2009). Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures. *AI Magazine*, 30(2): 72–83.

Johnson, W. Lewis and Shumin Wu (2008). Assessing Aptitude for Learning with a Serious Game for Foreign Language and Culture. Beverley P. Woolf, Esma Aïmeur, Roger Nkambou and Susanne Lajoie (Editors), *Intelligent Tutoring Systems*, Springer Berlin Heidelberg, 520–529.

Juffs, Alan (2001). Psycholinguistically Oriented Second Language Research. *Annual Review of Applied Linguistics*, 21: 207–220.

Jurafsky, Dan and James H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall.

Kadia, Kayiba (1988). The Effect of Formal Instruction on Monitored and on Spontaneous Naturalistic Interlanguage Performance: A Case Study. *TESOL Quarterly*, 22(3): 509–515.

Kamp, Hans and Uwe Reyle (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Kanno, Kazue (1997). The Acquisition of Null and Overt Pronominals in Japanese by English Speakers. *Second Language Research*, 13(3): 265–287.

Kaplan, Jonathan D. and V. Melissa Holland (1995). Application of Learning Principles to the Design of a Second Language Tutor. Holland et al. (1995).

Kaplan, Jonathan D., Mark A. Sabol, Robert A. Wisher and Robert J. Seidel (1998). The Military Language Tutor (MILT) Program: An Advanced Authoring System. *Computer Assisted Language Learning*, 11(3): 265–287.

Kempe, Vera and Brian MacWhinney (1998). The Acquisition of Case-marking by Adult Learners of Russian and German. *Studies in Second Language Acquisition*, 20(04): 543–587.

Kepner, Christine Goring (1991). An Experiment in the Relationship of Types of Written Feedback to the Development of Second-Language Writing Skills. *The Modern Language Journal*, 75(3): 305–313.

Kerminen, Antti and Kristiina Jokinen (2003). Distributed Dialogue Management in a Blackboard Architecture. *Proceedings of the EACL-2003 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management*, Budapest, 55–62.

Kern, Richard G. (1995). Restructuring Classroom Interaction with Networked Computers: Effects on Quantity and Characteristics of Language Production. *The Modern Language Journal*, 79(4): 457–476.

Klein, Phil (1998). Software Evaluation Review of "El Corrector" ('un medio fácil para corregir rápidamente el español'). *CALICO Journal*, 16(1): 64–75.

Knott, Alistair, John Moorfield, Tamsin Meaney and Lee-Luan Ng (2003). A Human-Computer Dialogue System for Maori Language Learning. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003*, 3336–3343.

Knott, Alistair and Peter Vlugter (2008). Multi-Agent Human-Machine Dialogue: Issues in Dialogue Management and Referring Expression Semantics. *Artificial Intelligence*, 172(2–3): 69–102.

Komatani, Kazunori, Katsuaki Tanaka, Hiroaki Kashima and Tatsuya Kawahara (2001). Domain-Independent Spoken Dialogue Platform Using Key-phrase Spotting Based on Combined Language Model. *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, Aalborg, Denmark, 1319–1322.

Kormos, Judit and Mariann Dénes (2004). Exploring Measures and Perceptions of Fluency in the Speech of Second Language Learners. *System*, 32(2): 145–164.

Krashen, Stephen D. (1981). *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press.

Krashen, Stephen D. (1982). *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon Press.

Krashen, Stephen D. (1985). *Input Hypothesis: Issues and Implications*. London: Longman.

Kwasny, Stan C. and Norman K. Sondheimer (1981). Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems. *Computational Linguistics*, 7(2): 99–108.

Lafford, Barbara A. (2004). Review of Tell Me More Spanish. *Language Learning & Technology*, 8(3): 21–34.

Lai, Chun, Fei Fei and Robin Roots (2008). The Contingency of Recasts and Noticing. *CALICO Journal*, 26(1): 70–90.

Lai, Chun and Yong Zhao (2006). Noticing and Text-Based Chat. *Language Learning & Technology*, 10(3): 102–120.

Larsen-Freeman, Diane (2006). The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics*, 27(4): 590–619.

Larsen-Freeman, Diane E. (1975). The Acquisition of Grammatical Morphemes by Adult ESL Students. *TESOL Quarterly*, 9(4): 409–419.

Larsson, Staffan (2002). *Issue-Based Dialogue Management*. Ph.D. thesis, Goteborg University.

Larsson, Staffan and David R. Traum (2000). Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit. *Natural Language Engineering*, 6(3-4): 323–340.

Lech, Till Christopher and Koenraad De Smedt (2006). Dreistadt: A Language Enabled Moo for Language Learning. *Proceedings of the ECAI-06 Workshop on Language-Enabled Educational Technology*.

Leeman, Jennifer (2007). Feedback in L2 learning: Responding to errors during practice. DeKeyser (2007b), 111–137.

Leeman, Jennifer, Igone Arteagoitia, Boris Fridman and Catherine Doughty (1995). Integrating Attention to Form With Meaning: Focus on Form in Content-Based Spanish Instruction. Richard Schmidt (Editor), *Attention and Awareness in Foreign Language Learning and Teaching*, Hawaii: University of Hawai'i Press, 217–258.

Lemon, Oliver and Alexander Gruenstein (2004). Multithreaded Context for Robust Conversational Interfaces: Context-Sensitive Speech Recognition and Interpretation of Corrective Fragments. *ACM Transactions on Computer-Human Interaction*, 11(3): 241–267.

Lemon, Oliver, Alexander Gruenstein, Alexis Battle and Stanley Peters (2002). Multi-Tasking and Collaborative Activities in Dialogue Systems. *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, 113–124.

Lemon, Olivier and Olivier Pietquin (2007). Machine Learning for Spoken Dialogue Systems. *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, 1761–1764.

Lennon, Paul (1990). Investigating Fluency in EFL: A Quantitative Approach. *Language Learning*, 40(3): 387–417.

Levene, Howard (1960). Robust Tests for Equality of Variances. Ingram Olkin (Editor), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, 278–292.

Levenshtein, Vladimir (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, volume 10.

Levin, Esther, Roberto Pieraccini and Wieland Eckert (2000). A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1): 11–23.

Levin, Lori S. and David A. Evans (1995). ALICE-chan: A Case Study in ICALL Theory and Practice. Holland et al. (1995).

Levin, Lori S., David A. Evans and Donna M. Gates (1991). The Alice System: A Workbench for Learning and Using Language. *CALICO Journal*, 9(1): 27–56.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge University Press.

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

Levy, Mike (2009). Technologies in Use for Second Language Learning. *The Modern Language Journal*, 93(s1): 769–782.

L'Haire, Sébastien (2004). Vers un feedback plus intelligent. Les enseignements du project Freetext. *Proceedings of the Journée d'étude de l'ATALA on NLP and Language Learning*, 1—-12.

Li, Shaofeng (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning*, 60(2): 309–365.

Liamkina, Olga (2008). Making Dative a Case for Semantic Analysis: Differences in Use Between Native and Non-Native Speakers of German. Andrea Tyler, Yiyoung Kim and Mari Takada (Editors), *Language in the Context of Use: Usage-Based Approaches to Language and Language Learning*, Berlin: Mouton de Gruyter, 145–165.

Lightbown, Patsy M. (1998). The Importance of Timing in Focus on Form. Doughty and Williams (1998a), 177–196.

van der Linden, Elisabeth (1993). Does Feedback Enhance Computer-Assisted Language Learning? *Computers & Education*, 21(1-2): 61–65.

Lison, Pierre (2014). *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, University of Oslo.

Loewen, Shawn (2009). Grammaticality Judgement Tests and the Measurement of Implicit and Explicit L2 Knowledge. Ellis et al. (2009), 94–112.

Loewen, Shawn and Rosemary Erlam (2006). Corrective Feedback in the Chatroom: An Experimental Study. *Computer Assisted Language Learning CALL*, 19(1): 1–14.

Loewen, Shawn, Shaofeng Li, Fei Fei, Amy Thompson, Kimi Nakatsukasa, Seongmee Ahn and Xiaoqing Chen (2009). Second Language Learners' Beliefs About Grammar Instruction and Error Correction. *The Modern Language Journal*, 93(1): 91–104.

Loewen, Shawn and Toshiyo Nabei (2007). The Effect of Oral Corrective Feedback on Implicit and Explicit L2 Knowledge. Mackey (2007), 361–378.

Loewen, Shawn and Jenefer Philp (2006). Recasts in the Adult English L2 Classroom: Characteristics, Explicitness, and Effectiveness. *The Modern Language Journal*, 90(4): 536–556.

Long, Michael H. (1981). Input, Interaction and Second Language Acquisition. *Annals of the New York Academy of Sciences*, 379: 259–278.

Long, Michael H. (1991). Focus on Form: A Design Feature in Language Teaching Methodology. Ralph B. Ginsberg De Bot, Kees and Claire Kramsch (Editors), *Foreign Language Research in Cross-Cultural Perspective*, John Benjamins Publishing, 39–52.

Long, Michael H. (1996). The Role of the Linguistic Environment in Second Language Acquisition. William C. Ritchie and Tej K. Bhatia (Editors), *Handbook of Second Language Acquisition*, Academic Press, 413–468.

Long, Michael H. (2000). Focus on Form in Task-Based Language Teaching. Richard D. Lambert and Elana Shohamy (Editors), *Language Policy and Pedagogy: Essays in Honor of A. Ronald Walton*, John Benjamins Publishing.

Long, Michael H. (2007). *Problems in SLA*. Lawrence Erlbaum.

Long, Michael H., Shunji Inagaki and Lourdes Ortega (1998). The Role of Implicit Negative Feedback in SLA: Models and Recasts in Japanese and Spanish. *The Modern Language Journal*, 82(3): 357–371.

Long, Michael H. and Peter Robinson (1998). Focus on Form: Theory, Research, and Practice. Doughty and Williams (1998a), 15–41.

Loschky, Lester and Robert Bley-Vroman (1993). Grammar and Task-Based Methodology. Susan M. Gass and Graham Crookes (Editors), *Tasks and Language Learning: Integrating Theory and Practice*, Multilingual Matters, 123–167.

Lund, Randall J. (2004). Erwerbssequenzen im Klassenraum. *Deutsch als Fremdsprache*, 41(2): 99–103.

Lyman-Hager, Mary Ann (2000). Bridging the Language-Literature Gap: Introducing Literature Electronically to the Undergraduate Language Student. *CALICO Journal*, 17(3): 431–452.

Lyster, Roy (1998). Recasts, Repetition, and Ambiguity in L2 Classroom Discourse. *Studies in Second Language Acquisition*, 20(01): 51–81.

Lyster, Roy (2004). Differential Effects of Prompts and Recasts in Form-Focused Instruction. *Studies in Second Language Acquisition*, 26(03): 399–432.

Lyster, Roy and Jesús Izquierdo (2009). Prompts Versus Recasts in Dyadic Interaction. *Language Learning*, 59(2): 453–498.

Lyster, Roy and Hirohide Mori (2006). Interactional Feedback and Instructional Counterbalance. *Studies in Second Language Acquisition*, 28(02): 269–300.

Lyster, Roy and Leila Ranta (1997). Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms. *Studies in Second Language Acquisition*, 19(01): 37–66.

Lyster, Roy and Kazuya Saito (2010). Oral Feedback in Classroom SLA - A Meta-analysis. *Studies in Second Language Acquisition*, 32(Special Issue 02): 265–302.

Mackey, Alison (1999). Input, Interaction and Second Language Development: An Empirical Study of Question Formation in ESL. *Studies in Second Language Acquisition*, 21(04): 557–587.

Mackey, Alison (Editor) (2007). *Conversational Interaction and Second Language Acquisition: A Series of Empirical Studies*. Oxford University Press.

Mackey, Alison and Susan M. Gass (2005). *Second Language Research: Methodology and Design*. Lawrence Erlbaum.

Mackey, Alison and Susan M. Gass (2006). Pushing the Methodological Boundaries in Interaction Research: An Introduction to the Special Issue. *Studies in Second Language Acquisition*, 28(02): 169–178.

Mackey, Alison, Susan M. Gass and Kim McDonough (2000). How Do Learners Perceive Interactional Feedback? *Studies in Second Language Acquisition*, 22(04): 471–497.

Mackey, Alison and Jaemyung Goo (2007). Interaction Research in SLA: A Meta-Analysis and Research Synthesis. Mackey (2007), 407–452.

Mackey, Alison and Jenefer Philp (1998). Conversational Interaction and Second Language Development: Recasts, Responses, and Red Herrings? *The Modern Language Journal*, 82(3): 338–356.

Mackey, Alison, Jenefer Philp, Takako Egi, Akiko Fujii and Tomoaki Tatsumi (2002). Individual Differences in Working Memory, Noticing of Interactional Feedback, and L2 Development. Peter Robinson (Editor), *Individual Differences and Instructed Language Learning*, John Benjamins Publishing, 181–209.

Mandell, Paul B. (1999). On the Reliability of Grammaticality Judgement Tests in Second Language Acquisition Research. *Second Language Research*, 15(1): 73–99.

Mann, Henry B. and Donald R. Whitney (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. *Annals of Mathematical Statistics*, 18(1): 50–60.

Mc Carthy, Christopher (2008). Interactional Corrective Feedback and Context in the Swedish EFL Classroom. [Online]. (URL http://su.diva-portal.org/smash/record.jsf?pid=diva2:199455). (Accessed 28.08.2014) Student thesis at Stockholm University, Department of English.

McTear, Michael F. (2002). Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys*, 34(1): 90–169.

McTear, Michael F. (2004). *Spoken Dialogue Technology: Towards the Conversational User Interface*. Springer London Berlin Heidelberg.

Meisel, Jürgen M. (1997). The Acquisition of the Syntax of Negation in French and German: Contrasting First and Second Language Development. *Second Language Research*, 13(3): 227–263.

Meisel, Jürgen M., Harald Clahsen and Manfred Pienemann (1981). On Determining Developmental Stages in Natural Second Language Acquisition. *Studies in Second Language Acquisition*, 3(02): 109–135.

Menzel, Barbara (2004). *Genuszuweisung im DaF-Erwerb: Psycholinguistische Prozesse und didaktische Implikationen*. Berlin: Weißensee-Verlag.

Menzel, Wolfgang and Ingo Schröder (1999). Error Diagnosis for Language Learning Systems. Mathias Schulze, Marie-Josée Hamel and June Thompson (Editors), *Language Processing in CALL ReCALL Special Publication*, 20–30.

Merkx, Marjolein, Kathleen Rastle and Matthew H. Davis (2011). The Acquisition of Morphological Knowledge Investigated through Artificial Language Learning. *The Quarterly Journal of Experimental Psychology*, 64(6): 1200–1220.

Meurers, Detmar (2012). Natural Language Processing and Language Learning. Carol A. Chapelle (Editor), *Encyclopedia of Applied Linguistics*, Blackwell Publishing.

Meurers, Detmar, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf and Niels Ott (2010). Enhancing Authentic Web Pages for Language Learners. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, Stroudsburg, PA, USA: Association for Computational Linguistics, 10–18.

Mika, Egmont (2005). *Formeln und Routinen : Zum Genuserwerb italienischer, portugiesischer und spanischer Gastarbeiter mit Deutsch als Zweitsprache*. Ph.D. thesis, Uppsala University, Department of Modern Languages.

Möllering, Martina (2004). *The Acquisition of German Modal Particles. A corpus-based approach*. Bern: Lang.

Morgan-Short, Kara and Harriet Wood Bowden (2006). Processing Instruction and Meaningful Output-based Instruction: Effects on Second Language Development. *Studies in Second Language Acquisition*, 28(01): 31–65.

Morgan-Short, Kara, Ingrid Finger, Sarah Grey and Michael T. Ullman (2012). Second Language Processing Shows Increased Native-like Neural Responses after Months of No Exposure. *PloS ONE*, 7(3): e32974.

Morton, Hazel, Nancie Davidson and Mervyn Jack (2008). Evaluation of a Speech Interactive CALL System. Felicia Zhang and Beth Barber (Editors), *Handbook of Research on Computer-Enhanced Language Acquisition and Learning*, Idea Group Publishing, 219–239.

Morton, Hazel and Mervyn Jack (2005). Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning*, 18(3): 171 – 191.

Muranoi, Hitoshi (2007). Output Practice in the L2 Classroom. DeKeyser (2007b).

Naber, Daniel (2003). *A Rule-Based Style and Grammar Checker*. Master's thesis, Universität Bielefeld.

Nagata, Noriko (1993). Intelligent Computer Feedback for Second Language Instruction. *The Modern Language Journal*, 77(3): 330–339.

Nagata, Noriko (1997). The Effectiveness of Computer-Assisted Metalinguistic Instruction: A Case Study in Japanese. *Foreign Language Annals*, 30(2): 187–200.

Nagata, Noriko (2002). BANZAI: An Application of Natural Language Processing to Web based Language Learning. *CALICO Journal*, 19(3): 583–599.

Nagata, Noriko (2009). Robo-Sensei's NLP-Based Error Detection and Feedback Generation. *CALICO Journal*, 26(3): 562–579.

Nagata, Noriko and M. Virginia Swisher (1995). A Study of Consciousness-Raising by Computer: The Effect of Metalinguistic Feedback on Second Language Learning. *Foreign Language Annals*, 28(3): 337–347.

Nerbonne, John (2003). Natural Language Processing in Computer-Assisted Language Learning. Ruslan Mitkov (Editor), *Handbook of Computational Linguistics*, Oxford University Press, 670–698.

Nerbonne, John and Duco Dokter (1999). An Intelligent Word-Based Language Learning Assistant. *Traitement Automatique des Languages*, 40(1): 125–142.

Nerbonne, John, Duco Dokter and Petra Smit (1998). Morphological Processing and Computer-Assisted Language Learning. *Computer-Assisted Language Learning*, 11(5): 543–559.

Newport, Elissa L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, 14: 11–28.

Nicholas, Howard, Patsy M. Lightbown and Nina Spada (2001). Recasts as Feedback to Language Learners. *Language Learning*, 51(4): 719–758.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov and Erwin Marsi (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13(02): 95–135.

Norris, John M. and Lourdes Ortega (2000). Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis. *Language Learning*, 50(03): 417–528.

Norris, John M. and Lourdes Ortega (2001). Does Type of Instruction Make a Difference? Substantive Findings from a Meta-Analytic Review. Rod Ellis (Editor), *Form-Focused Instruction and Second Language Learning*, Blackwell Publishing, 157–213.

Nunan, David (1989). *Designing Tasks for the Communicative Classroom*. Cambridge University Press.

Oliver, Rhonda (1995). Negative Feedback in Child NS-NNS Conversation. *Studies in Second Language Acquisition*, 17(04): 459–481.

Ott, Niels and Ramon Ziai (2010). Evaluating Dependency Parsing Performance on German Learner Language. *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, volume 9, 175–186.

Panova, Iliana and Roy Lyster (2002). Patterns of Corrective Feedback and Uptake in an Adult ESL Classroom. *TESOL Quarterly*, 36(4): 573–595.

Paradis, Michel (1994). Neurolinguistic Aspects of Implicit and Explicit Memory: Implications for Bilingualism. Ellis (1994a), 393–419.

Payne, J. Scott and Paul J. Whitney (2002). Developing L2 Oral Proficiency through Synchronous CMC: Output, Working Memory, and Interlanguage Development. *CALICO Journal*, 20(1): 7–32.

Pérez-Leroux, Ana T. and William R. Glass (1999). Null Anaphora in Spanish Second Language Acquisition: Probabilistic Versus Generative Approaches. *Second Language Research*, 15(2): 220–249.

Perrault, C. Raymond and James F. Allen (1980). A Plan-Based Analysis of Indirect Speech Acts. *Computational Linguistics*, 6: 167–182.

Petersen, Ken (2010). *Implicit Corrective Feedback in Computer-Guided Interaction: Does Mode Matter?* Ph.D. thesis, Georgetown University.

Philp, Jenefer (2003). Constraints on "Noticing the Gap": Nonnative Speakers' Noticing of Recasts in NS-NNS Interaction. *Studies in Second Language Acquisition*, 25(01): 99–126.

Pica, Teresa (1983). Adult Acquisition of English as a Second Language Under Different Conditions of Exposure. *Language Learning*, 33(4): 465–497.

Pica, Teresa (1994). Research on Negotiation: What Does It Reveal About Second-Language Learning Conditions, Processes, and Outcomes? *Language Learning*, 44(3): 493–527.

Pienemann, Manfred (1984). Psychological Constraints on the Teachability of Languages. *Studies in Second Language Acquisition*, 6(02): 186–214.

Pienemann, Manfred (1988). Determining the Influence of Instruction on L2 Speech Processing. *AILA Review*, 5: 40–72.

Pienemann, Manfred (1989). Is Language Teachable? Psycholinguistic Experiments and Hypotheses. *Applied Linguistics*, 10(1): 52–79.

Pienemann, Manfred, Malcolm Johnston and Geoff Brindley (1988). Constructing an Acquisition-Based Procedure for Second Language Assessment. *Studies in Second Language Acquisition*, 10(02): 217–243.

Pietquin, Olivier and Thierry Dutoit (2006). A Probabilistic Framework for Dialog Simulation and Optimal Strategy Learning. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2): 589–599.

Pittner, Karin and Judith Berman (2008). *Deutsche Syntax. Ein Arbeitsbuch*. Tübingen: Narr.

Poesio, Massimo (2000). Semantic Analysis. Robert Dale, Hermann Moisl and Harold Somers (Editors), *Handbook of Natural Language Processing*, Marcel Dekker Inc., 93–122.

Pollard, Carl and Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pollard, David and Masoud Yazdani (1993). A Multilingual Multimedia Restaurant Scenario. Masoud Yazdani (Editor), *Multilingual Multimedia: Bridging the Language Barrier with Intelligent Systems*, Intellect Books, 1–13.

Price, Charlotte, Andrea Bunt and Gordon McCalla (1999). L2tutor: A Mixed-Initiative Dialogue System for Improving Fluency. *Computer Assisted Language Learning*, 12(2): 83–112.

Purpura, James Enos (2004). *Assessing Grammar*. Cambridge University Press.

Rambow, Owen, Srinivas Bangalore and Marilyn Walker (2001). Natural Language Generation in Dialog Systems. *Proceedings of the First International Conference on Human Language Technology Research*, Stroudsburg, PA, USA: Association for Computational Linguistics, 1–4.

Raux, Antoine (2004). Automated Lexical Adaptation and Speaker Clustering Based on Pronunciation Habits for Non-Native Speech Recognition. *Proceedings of INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*.

Raux, Antoine and Maxine Eskenazi (2004a). Non-Native Users in the Let's Go!! Spoken Dialogue System: Dealing with Linguistic Mismatch. Daniel Marcu Susan Dumais and Salim Roukos (Editors), *HLT-NAACL 2004: Main Proceedings*, Boston, Massachusetts, USA: Association for Computational Linguistics, 217–224.

Raux, Antoine and Maxine Eskenazi (2004b). Using Task-Oriented Spoken Dialogue Systems for Language Learning: Potential, Practical Applications and Challenges. *Proceedings of InSTIL/ICALL 2004 – Computer Assisted Learning, NLP and Speech Technologies in Advanced Language Learning Systems*.

Razagifard, Parisa and Massoud Rahimpour (2010). The Effect of Computer-Mediated Corrective Feedback on the Development of Second Language Learners' Grammar. *International Journal of Instructional Technology and Distance Learning*, 7(5): 11–30.

Reber, Arthur S., Faye F. Walkenfeld and Ruth Hernstadt (1991). Implicit and Explicit Learning: Individual Differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5): 888–896.

Reuer, Veit (2003). Error Recognition and Feedback with Lexical Functional Grammar. *CALICO Journal*, 20(3): 497–512.

Rezaei, Saeed and Ali Derakhshan (2011). Investigating Recast and Metalinguistic Feedback in Task-based Grammar Instruction. *Journal of Language Teaching and Research*, 2(3): 655–663.

Rezaei, Saeed, Farzaneh Mozaffari and Ali Hatef (2011). Corrective Feedback in SLA: Classroom Practice and Future Directions. *International Journal of English Linguistics*, 1(1): 21–29.

Rieser, Verena (2008). *Bootstrapping Reinforcement Learning-based Dialogue Strategies from Wizard-of-Oz Data*. Ph.D. thesis, Saarland University.

Rieser, Verena and Oliver Lemon (2008). Learning Effective Multimodal Dialogue Strategies from Wizard-of-Oz Data: Bootstrapping and Evaluation. *Proceedings of ACL-08: HLT*, 638—-646.

Rieser, Verena and Oliver Lemon (2011). *Reinforcement Learning for Adaptive Dialogue Systems: A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Springer Berlin Heidelberg.

Rimrott, Anne and Trude Heift (2008). Evaluating Automatic Detection of Misspellings in German. *Language Learning & Technology*, 12(3): 73–92.

Roberts, Michael A. (1995). Awareness and the Efficacy of Error Correction. Richard W. Schmidt (Editor), *Attention and Awareness in Foreign Language Learning*, Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Robinson, Peter (1996). Learning Simple and Complex Second Language Rules Under Implicit, Incidental, Rule-Search, and Instructed Conditions. *Studies in Second Language Acquisition*, 18(01): 27–67.

Robinson, Peter (2001). Individual Differences, Cognitive Abilities, Aptitude Complexes and Learning Conditions in Second Language Acquisition. *Second Language Research*, 17(4): 368–392.

Rogers, Margaret (1982). Interlanguage Variability: Verb Placement Errors in German. Gerhard Nickel and Dietrich Nehls (Editors), *Error Analysis, Contrastive Linguistics and Second Language Learning, Selected Papers from the 6th International Congress of Applied Linguistics, Lund 1981*, Heidelberg: Groos, 43–83.

Rogers, Margaret (1984). On Major Types of Written Error in Advanced Students of German. *IRAL - International Review of Applied Linguistics in Language Teaching*, 22(1): 1–40.

Rogers, Margaret (1987). Learners Difficulties with Grammatical Gender in German as a Foreign Language. *Applied Linguistics*, 8(1): 48–74.

Rosé, Carolyn Penstein, Barbara Di Eugenio, Lori S. Levin and Carol Van Ess-Dykema (1995). Discourse Processing of Dialogues with Multiple Threads. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, ACL '95, 31–38.

Rösler, Dietmar (1982). Teaching German Modal Particles. *IRAL - International Review of Applied Linguistics in Language Teaching*, 20(1–4): 33—-38.

Russell, Jane and Nina Spada (2006). The Effectiveness of Corrective Feedback for the Acquisition of L2 Grammar: A Meta-Analysis of the Research. John M. Norris and Lourdes Ortega (Editors), *Synthesizing Research on Language Learning and Teaching*, John Benjamins Publishing, 133–164.

Sachs, Rebecca and Bo-Ram Suh (2007). Textually Enhanced Recasts, Learner Awareness, and L2 Outcomes in Synchronous Computer-Mediated Interaction. Mackey (2007), 197–227.

Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50(4): 696–735.

Salaberry, Rafael M. (1996). A Theoretical Foundation for the Development of Pedagogical Tasks in Computer Mediated Communication. *CALICO Journal*, 14(1): 5–34.

Sanders, Ruth H. and Alton F. Sanders (1995). History of an AI Spy Game: Spion. Holland et al. (1995).

Sauro, Shannon (2009). Computer-Mediated Corrective Feedback and the Development of L2 Grammar. *Language Learning & Technology*, 13(1): 96–120.

Saxton, Matthew (1997). The Contrast Theory of Negative Input. *Journal of Child Language*, 24(01): 139–161.

Saygin, Ayse Pinar, Ilyas Cicekli and Varol Akman (2000). Turing Test: 50 Years Later. *Minds and Machines*, 10(4): 463–518.

Schachter, Jacquelyn (1996). Maturation and the Issue of Universal Grammar in Second Language Acquisition. William C. Ritchie and Tej K. Bhatia (Editors), *Handbook of Second Language Acquisition*, Academic Press, 159–163.

Schegloff, Emanuel A. and Harvey Sacks (1973). Opening up Closings. *Semiotica*, 8(4): 289–327.

Schlangen, David and Gabriel Skantze (2009). A General, Abstract Model of Incremental Dialogue Processing. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, EACL '09, 710–718.

Schmidt, Richard (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, 11(2): 129–158.

Schmidt, Richard (1993). Consciousness, Learning and Interlanguage Pragmatics. Gabriele Kasper and Shoshana Blum-Kulka (Editors), *Interlanguage Pragmatics*, Oxford University Press, 21–42.

Schmidt, Richard (1994). Deconstructing Consciousness In Search Of Useful Definitions For Applied Linguistics. *AILA Review*, 11: 11–26.

Schmidt, Richard (2001). Attention. Peter Robinson (Editor), *Cognition and Second Language Instruction*, Cambridge University Press.

Schröder, Jochen (1978). Zum Zusammenhang von Lokativität und Direktionalität bei einigen wichtigen deutschen Präpositionen. *Deutsch als Fremdsprache*, 15: 9–15.

Schubert, Lenhart (2014). Computational Linguistics. Edward N. Zalta (Editor), *The Stanford Encyclopedia of Philosophy*, Spring 2014 edition. (URL http://plato.stanford.edu/archives/spr2014/entries/computational-linguistics/). (Accessed 28.08.2014).

Schulze, Mathias (2008). AI in CALL–Artificially Inflated or Almost Imminent? *CALICO Journal*, 25(3): 510–527.

Schumacher, Nicole (2005). *Tempus als Lerngegenstand: Ein Modell für Deutsch als Fremdsprache und seine Anwendung für italienische Lernende*. Tübingen: Narr.

Schwartz, Bonnie D. (1993). On Explicit and Negative Data Effecting and Affecting Competence and Linguistic Behavior. *Studies in Second Language Acquisition*, 15(02): 147–163.

Schwind, Camilla B. (1995). Error Analysis and Explanation in Knowledge Based Language Tutoring. *Computer Assisted Language Learning*, 8(4): 295–924.

Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Searle, John R. (1976). A Classification of Illocutionary Acts. *Language in Society*, 5(01): 1–23.

Selinker, Larry (1972). Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching*, 10(2): 209–232.

Semke, Harriet D. (1984). Effects of the Red Pen. *Foreign Language Annals*, 17(3): 195–202.

Seneff, Stephanie, Chao Wang and Julia Zhang (2004). Spoken Conversational Interaction for Language Learning. *Proceedings of InSTIL/ICALL 2004 – Computer Assisted Learning, NLP and Speech Technologies in Advanced Language Learning Systems*.

Shapiro, Samuel S. and Martin B. Wilk (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4): 591–611.

Sharwood Smith, Michael (1993). Input Enhancement in Instructed SLA: Theoretical Bases. *Studies in Second Language Acquisition*, 15: 165–179.

Shea, Peter (2000). Leveling the Playing Field: A Study of Captioned Interactive Video for Second Language Learning. *Journal of Educational Computing Research*, 22(3): 243–63.

Sheen, Younghee (2007). The Effects of Corrective Feedback, Language Aptitude and Learner Attitudes on The Acquisition of English Articles. Mackey (2007), 301–322.

Sheen, Younghee (2010a). Differential Effects of Oral and Written Corrective Feedback in the ESL Classroom. *Studies in Second Language Acquisition*, 32(02): 203–234.

Sheen, Younghee (2010b). Introduction. *Studies in Second Language Acquisition*, 32(Special Issue 02): 169–179.

Sheppard, Ken (1992). Two Feedback Types: Do They Make A Difference? *RELC Journal*, 23(1): 103–110.

Shieber, Stuart M. (Editor) (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. MIT Press.

Siegel, Sidney and N. John Castellan (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Skehan, Peter (1996a). A Framework for the Implementation of Task-based Instruction. *Applied Linguistics*, 17(1): 38–62.

Skehan, Peter (1996b). Second Language Acquisition Research and Task-based Instruction. Jane Willis and Dave Willis (Editors), *Challenge and Change in Language Teaching*, Oxford: Heinemann.

Skehan, Peter (1998). *A Cognitive Approach to Language Learning*. Oxford University Press.

Skehan, Peter and Pauline Foster (1999). The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning*, 49(1): 93–120.

Slobin, Dan Isaac (1973). Cognitive Prerequisites for the Development of Grammar. Charles A. Ferguson and Dan Isaac Slobin (Editors), *Studies of Child Language Development*, Holt, Rinehart, and Winston.

Smith, Bryan (2004). Computer-Mediated Negotiated Interaction and Lexical Acquisition. *Studies in Second Language Acquisition*, 26(03): 365–398.

Smith, Bryan (2005). The Relationship between Negotiated Interaction, Learner Uptake, and Lexical Acquisition in Task-Based Computer-Mediated Communication. *TESOL Quarterly*, 39(1): 33–58.

Spada, Nina and Patsy M. Lightbown (1993). Instruction and the Development of Questions in L2 Classrooms. *Studies in Second Language Acquisition*, 15(02): 205–224.

Spada, Nina and Yasuyo Tomita (2010). Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning*, 60(2): 263–308.

Spinner, Patti and Alan Juffs (2008). L2 Grammatical Gender in a Complex Morphological System: The Case of German. *IRAL - International Review of Applied Linguistics in Language Teaching*, 46(4): 315–348.

Stalnaker, Robert (2002). Common Ground. *Linguistics and Philosophy*, 25(5): 701–721.

Statan, Larry (2006). Review of Side by Side Interactive. *Language Learning & Technology*, 10(3): 36–43.

Stewart, Iain A. D. and Portia File (2007). Let's Chat: A Conversational Dialogue System for Second Language Practice. *Computer Assisted Language Learning*, 20(2): 97–116.

Stockwell, Glenn (2007). A Review of Technology Choice for Teaching Language Skills and Areas in the CALL Literature. *ReCALL*, 19(02): 105–120.

Suzuki, Mikiko (2004). Corrective Feedback and Learner Uptake in Adult ESL Classrooms. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 4(2).

Swain, Merrill (1985). Communicative Competence: Some Roles of Comprehensible Input and Comprehensible Output in Its Development. Susan M. Gass and Carolyn G. Madden (Editors), *Input in Second Language Acquisition*, Newbury House, 235–253.

Swain, Merrill (1995). Three Functions of Output in Second Language Learning. Guy Cook and Barbara Seidlhofer (Editors), *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*, Oxford University Press, 125–144.

Swain, Merrill (2005). The Output Hypothesis: Theory and Research. Eli Hinkel (Editor), *Handbook of Research in Second Language Teaching and Learning*, Lawrence Erlbaum, 471–483.

Sykes, Julie M. (2005). Synchronous CMC and Pragmatic Development: Effects of Oral and Written Chat. *CALICO Journal*, 22(3): 399–431.

Szagun, Gisela (1997). Some Aspects of Language Development in Normal-Hearing Children and Children With Cochlear Implants. *The American Journal of Otology*, 18(6): 131–134.

Tamminen, Jakke, Matthew H. Davis, Marjolein Merkx and Kathleen Rastle (2012). The Role of Memory Consolidation in Generalisation of New Linguistic Information. *Cognition*, 125(1): 107–112.

Taylor, Paul (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Timmermann, Waltraud (2005). *Tempusverwendung in chinesisch-deutscher Lernersprache. Eine Analyse auf sprachenvergleichender Basis*. Münster: Waxmann.

Tomokiyo, Laura (2001). *Recognizing Non-Native Speech: Characterizing and Adapting to Non-Native Usage in Speech Recognition*. Ph.D. thesis, Carnegie Mellon University.

Towell, Richard, Roger Hawkins and Nives Bazergui (1996). The Development of Fluency in Advanced Learners of French. *Applied Linguistics*, 17(1): 84–119.

Traum, David R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

Traum, David R. (1999). Computational Models of Grounding in Collaborative Systems. *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, 124–131.

Traum, David R. (2008). Extended Abstract: Computational Models of Non-Cooperative Dialogues. *Proceedings of LONDIAL 2008, the 12th Workshop on the Semantics and Pragmatics of Dialogue*, 11–14.

Traum, David R. and James F. Allen (1994). Discourse Obligations in Dialogue Processing. *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, ACL '94, 1–8.

Traum, David R. and Hinkelman Elizabeth (1992). Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, 8(3): 579–599.

Traum, David R. and Staffan Larsson (2003). The Information State Approach to Dialogue Management. *Current and New Directions in Discourse & Dialogue*, Kluwer Academic Publishers, 325–353.

Traum, David R. and Jeff Rickel (2002). Embodied Agents for Multi-Party Dialogue in Immersive Virtual Worlds. *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, 766–773.

Trim, John, Brian North and Daniel Coste (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Niveau A1, A2, B1, B2*. Langenscheidt.

Truscott, John (1996). The Case Against Grammar Correction in L2 Writing Classes. *Language Learning*, 46(2): 327–369.

Truscott, John (1999a). The Case for "The Case Against Grammar Correction in L2 Writing Classes": A Response to Ferris. *Journal of Second Language Writing*, 8(2): 111 – 122.

Truscott, John (1999b). What's Wrong with Oral Grammar Correction. *Canadian Modern Language Review*, 55(4): 437–56.

Truscott, John (2004). Evidence and Conjecture on the Effects of Correction: A Response to Chandler. *Journal of Second Language Writing*, 13: 337–343.

Van Noord, Gertjan, Gosse Bouma, Rob Koeling and Mark-Jan Nederhof (1999). Robust Grammatical Analysis for Spoken Dialogue Systems. *Natural Language Engineering*, 5(1): 45–93.

VanPatten, Bill (1990). Attending to Form and Content in the Input: An Experiment in Consciousness. *Studies in Second Language Acquisition*, 12(03): 287–301.

Vlugter, Peter, Alistair Knott, J. McDonald and C. Hall (2009). Dialogue-Based CALL: A Case Study on Teaching Pronouns. *Computer Assisted Language Learning*, 22(2): 115–131.

Vlugter, Peter, Edwin Van Der Ham and Alistair Knott (2006). Error Correction Using Utterance Disambiguation Techniques. *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, 123–130.

Wahlster, Wolfgang (2006). Dialogue Systems Go Multimodal: The SmartKom Experience. *Smartkom: Foundations of Multimodal Dialogue Systems*, Springer Berlin Heidelberg, Cognitive Technologies, 3–27.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm and Alicia Abella (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, EACL '97, 271–280.

Walker, Marilyn A. and Owen C. Rambow (2002). Spoken Language Generation. *Computer Speech & Language. Special Issue on Spoken Language Generation*, 16(3—4): 273–281.

Walker, Matthew P. (2005). A Refined Model of Sleep and the Time Course of Memory Formation. *Behavioral and Brain Sciences*, 28(01): 51–64.

Wang, Chao and Stephanie Seneff (2007). A Spoken Translation Game for Second Language Learning. *Proceedings of the 13th International Conference on Artificial Intelligence in Education, AIED 2007*, 315–322.

Warschauer, Mark (1996). Comparing Face-to-Face and Electronic Discussion in the Second Language Classroom. *CALICO Journal*, 13(2-3): 7–26.

Weijenberg, Jan (1980). *Authentizität gesprochener Sprache in Lehrwerken für Deutsch als Fremdsprache*. Heidelberg: Groos.

Weinert, Regina (1994). Some Effects of a Foreign Language Classroom on the Development of German Negation. *Applied Linguistics*, 15(1): 76–101.

Weizenbaum, Joseph (1966). ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1): 36–45.

Welch, Bernard L. (1947). The Generalization of 'Student's' Problem when Several Different Population Varlances Are Involved. *Biometrika*, 34(1-2): 28–35.

White, Lydia (1987). Against Comprehensible Input: The Input Hypothesis and the Development of Second-Language Competence. *Applied Linguistics*, 8(2): 95–110.

White, Lydia (1991). Adverb Placement in Second Language Acquisition: Some Effects of Positive and Negative Evidence in the Classroom. *Second Language Research*, 7(2): 133–161.

Wilks, Yorick and Roberta Catizone (2000). Human-Computer Conversation. *Encyclopedia of Microcomputers*, New York: Dekker.

Williams, John N. (2005). Learning Without Awareness. *Second Language Research*, 27(02): 269–304.

Wilske, Sabrina and Magda Wolska (2011). Meaning versus Form in Computer-assisted Task-based Language Learning: A Case Study on the German Dative. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(1): 23–37.

Wittgenstein, Ludwig (1953/2009). *Philosophical Investigations*. Wiley-Blackwell.

Witzel, Jeffrey D. and Lesley Ono (2003). Recasts, Salience, and Morpheme Acquisition. *Paper presented at the 26th Annual Second Language Research Forum (SLRF)*, University of Arizona, Tucson, AZ.

Yang, Jie Chi and Kanji Akahori (1999). An Evaluation of Japanese CALL Systems on the WWW Comparing a Freely Input Approach With Multiple Selection. *Computer Assisted Language Learning*, 12(1): 59–79.

Young, Steve, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson and Kai Yu (2010). The Hidden Information State Model: A Practical Framework for POMDP-Based Spoken Dialogue Management. *Computer Speech & Language*, 24(2): 150–174.

Zhang, Pengyuan, Qingwei Zhao and Yonghong Yan (2007). A Spoken Dialogue System Based on Keyword Spotting Technology. *Proceedings of the 12th International Conference on Human-Computer Interaction (HCI International 2007). HCI Intelligent Multimodal Interaction Environments*, Springer Berlin Heidelberg, 253–261.

Zhao, Yong (1997). The Effects of Listeners' Control of Speech Rate on Second Language Comprehension. *Applied Linguistics*, 18(1): 49–68.

Zhao, Yong (2003). Recent Developments in Technology and Language Learning: A Literature Review and Meta-Analysis. *CALICO Journal*, 21(1): 7–28.