Talent in nonnative phonetic convergence

Von der Philosophisch-Historischen Fakultät der Universität Stuttgart
zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.)
genehmigte Abhandlung

Vorgelegt von

Natalie Lewandowski

aus Helmstedt

Hauptberichter: Prof. Dr. Grzegorz Dogil
Mitberichter: Prof. Dr. Bernd Möbius

Tag der mündlichen Prüfung: 02.12.2011

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

2012

Motto

”Language (...) lies on the borderline between oneself and the other.
The word in language is half someone else’s.”

Michail M. Bakhtin [BH81, 293]

Acknowledgments

Najlepszym polskim przyjaciołom w kraju i na „obczyźnie“– za całokształt:
Natalia, Asika, Kamila, Jagoda, Mateusz, Kuba i Jarek – bucha!

Meinem Betreuer Professor Grzegorz Dogil, all meinen ehemaligen und aktuellen KollegInnen am IMS, IfLA und im Graduiertenkolleg ein großes Dankeschön für die unzähligen praktischen Tipps und ein immer offenes Ohr für alle wissenschaftlichen (und nicht nur solche) Fragen.

I would like to thank Professor Ann R. Bradlow and Midam Kim for sharing their comments and giving me new impulses for my work.

And a great thank you goes to my infinitely patient native speaker subjects - JP and TM!

Contents

List of Figures

List of Tables

List of Abbreviations

ART	Adaptive Resonance Theory
ATM	Autonomous Transmission Model
BAS	Behavior Activation System
BIS	Behavior Inhibition System
CAT	Communication Accommodation Theory
CLS	Complementary Learning Systems approach
CSM	Context Sequence Model
DICE	Dissociated interactions and conscious control model
DM	Declarative memory
DP	Declarative-procedural model
EFL	English as a Foreign Language
EPIC	Executive processing interactive control model
FIT	Feature Integration Theory
GA	General American (accent)
IAM	Interactive Alignment Model
IDs	Individual Differences
L1	First language
L2	Second/foreign language
LTM	Long-term memory
MLAT	Modern Languages Aptitude Test

NNS	Nonnative speaker
NS	Native speaker
PDH	Procedural Deficit Hypothesis
PLAB	Pimsleur’s Language Aptitude Battery
PM	Procedural memory
Polysp	Polysystemic speech perception (system)
SLA	Second Language Acquisition
SSBE	Standard Southern British English (accent)
STM	Short-term memory
VOT	Voice onset time
WM	Working memory

Zusammenfassung

Phonetische Konvergenz beschreibt das Phänomen des sich Annäherns zweier Personen in Bezug auf ihre Aussprache, das aus einer kommunikativen Interaktion heraus entsteht. Diese Tendenz für mehr Synchronität in der phonetischen Domäne der Sprache umfasst sowohl Veränderungen der segmentalen als auch der suprasegmentalen Eigenschaften. Gegenstand dieser Arbeit war die Untersuchung phonetischer Konvergenz in einem fremdprachlichen Kontext, in gemischten Dialogen zwischen Muttersprachlern des Deutschen und Englischen. Vorrangig dabei war festzustellen, in wieweit der individuelle Faktor *phonetisches Talent* die natürlich auftretende Konvergenz im Dialog beeinflusst.

Der Ursprung und Zweck von Konvergenz im Allgemeinen und speziell im Bereich der Aussprache wurde bereits seit den siebziger Jahren untersucht. Zu der Zeit wurde Konvergenz, bzw. die Anpassung zwischen Sprechern, meist vom Gesichtspunkt sozialer Faktoren analysiert, die den Verlauf des Prozesses beeinflussen. Die Communication Accommodation Theory (CAT)[1] stellte die Behauptung auf, dass alle positiven und negativen Verschiebungen im Verhalten einer Person (Phonetik miteingeschlossen), durch den Drang weniger oder mehr soziale Distanz zu schaffen, begründet sind. Neuere Ansätze, wie das Interactive Alignment Modell[2], präsentieren eine prozessorientierte Theorie zur Enstehung von Konvergenz, ohne dabei den Einfluss von sozialen Faktoren zu diskutieren. Die scheinbare Unvereinbarkeit beider Theorien, die den Kontrast zwischen kontrollierbaren und automatischen Abläufen betrifft, sorgte seither für viele Debatten um die Beeinflussbarkeit von Konvergenz.

Die Untersuchung phonetischer Anpassung in einem fremdsprachlichen Kontext, rief zuallererst die Notwendigkeit hervor allgemeine Theorien der Identität und Identitätsnegotiation, sowie der hierfür eingesetzten Mittel – den Sprachstil mit speziellem Fokus auf Phonetik – zu beschreiben. Im Weiteren wurden auch die speziellen Gegebenheiten, die durch die unterschiedlichen Statusrollen der

[1] siehe [GP75, GO06].

[2] siehe [PG04a].

Muttersprachler und nicht-Muttersprachler, sowie deren entsprechende Kompetenzen hervorgerufen wurden, miteinbezogen. Da Konvergenz sich als subjektiv sehr variables Phänomen erweist, wurden zusätzlich auch die verschiedenen Einflussmöglichkeiten individueller Faktoren erläutert.

Die phonetische Anpassung ist ein hochgradig sprecher- und kontextbezogener Prozess, daher können die ihr zugrundeliegenden Abläufe am effektivsten innerhalb eines theoretischen Ansatzes beschrieben werden, der eine Vielzahl von situationellen Variablen miteinbezieht. Ein Ansatz, der diesen Voraussetzungen gerecht wird, ist ein benutzungsorientierter Ansatz der Sprachverarbeitung – die Exemplartheorie[3]. Da exemplarbasierte Theorien die Abspeicherung von individuellen Wortformen im Gedächtnis mit einer Vielzahl von sprecher- und kontextspezifischen Indizes vorhersehen, erlauben sie zugleich eine genaue Modellierung von Variabilität in der Sprachverarbeitung. Diese Eigenschaften gewährleisten, dass exemplarbasierte Ansätze die passenden Rahmenbedingungen für die Beschreibung aller Prozesse bieten, die phonetischer Konvergenz zugrunde liegen.

Für eine Untersuchung des genauen Einflussgrades von Talent auf Konvergenz im fremdsprachlichen Kontext, wurden zwei Experimente konzipiert – das Hauptexperiment in Dialogform und ein auf gelesener Sprache basierender Kontrolltest. Zwanzig deutsche Muttersprachler unterhielten sich in zwei getrennten Dialogsituationen mit jeweils einem englischen Muttersprachler – einem amerikanischen Sprecher und einer britischen Sprecherin. Am Ende des jeweiligen Dialogs wurden die deutschen Sprecher zusätzlich gebeten den Verlauf des Dialogs und die Ergebnisse der zu lösenden Aufgabe zusammenzufassen. Vor und nach jedem Dialog lasen die Teilnehmer eine Liste mit Zielwörtern aus dem Dialog vor, die den Kontrolltest darstellte. Die Dialoge waren als quasi-spontane aufgabenorientierte Interaktionen konzipiert, in denen die zu lösende Aufgabe darin bestand ein Fehler-

[3] siehe u.a. [Joh97, BH01, Pie01].

suchspiel mit zwei zusammengehörigen Bildern zu lösen – den Diapix[4]. Die angewandte Methode ließ es zu natürliche Sprache mit ausbalancierten Sprechanteilen aufzuzeichnen, die darüber hinaus auch eine ausreichende Anzahl von Wiederholungen der Zielwörter enthielt.

Die akustischen Messungen basierten auf der Extraktion von Amplituden-Hüllen aus dem Sprachsignal. Die Analyse erfolgte auf Wortebene durch den Vergleich dreier Zeitpunkte innerhalb des Dialogs – früh, spät und während der Zusammenfassung – sowie zwischen den einzelnen Wortlisten. Die Amplituden-Hüllen der Wörter beider Dialogpartner wurden mittels einer Kreuzkorrelation miteinander verglichen, um ihren spektralen Ähnlichkeitsgrad zu bewerten. Die dadurch erhaltenen Vergleichswerte stellten die Basis für die darauffolgenden statistischen Analysen dar.

Die Hauptannahme über den Einfluss von Talent auf phonetische Konvergenz konnte im Folgenden bestätigt werden. Im Vergleich zu den weniger talentierten Sprechern, wiesen die talentierten Sprecher signifikant mehr Konvergenz zu ihrem englischen Gesprächspartner zwischen einem frühen und späten Zeitpunkt des Dialogs auf. Das Geschlecht der Probanden hatte dagegen keinen signifikanten Einfluss auf das Ergebnis des Dialogexperiments, ebenso wie auf das des Kontrolltests. Die englischen Muttersprachler wurden vor dem Beginn des Experiments über dessen Ziele in Kenntnis gesetzt und angewiesen, ihre Aussprache weitgehend zu kontrollieren, so dass keine positiven oder negativen Verschiebungen stattfinden. Trotz der expliziten Anweisung an die Sprecher, ihre Aussprache konstant zu halten, konvergierten beide englischen Sprecher in Richtung ihrer deutschen Gesprächspartner. Längerfristige Auswirkungen des während der Dialoge erzielten Konvergenzeffektes auf den Zusammenfassungsteil oder den Kontrolltest konnten nicht bestätigt werden. Dies spricht dafür, dass die aus der Dialogsituation stammende Konvergenz nicht auf andere Sprachstile, wie den Erzählstil oder gelesene Sprache, übertragen wird.

[4] siehe [vEBBB$^+$10].

Die beschriebenen Funde sprechen dafür ein Hybridmodell für Konvergenz anzunehmen, das sowohl hauptsächlich unterbewusst ablaufende, automatische Prozesse, als auch die Komponenten berücksichtigen kann, die partiell der bewussten Kontrolle unterliegen. Phonetisches Talent scheint dabei direkten Einfluss auf den zentralen Mechanismus für phonetische Konvergenz in einer Fremdsprache auszuüben. Der Wirkungsort des Talentfaktors sind möglicherweise Prozesse innerhalb des Aufmerksamkeits- und Gedächtnisnetzwerkes in der Sprachverarbeitung. Das Fehlen von Übertragungseffekten für Konvergenz vom Dialog in einen monologischen Sprachstil suggeriert, dass phonetische Anpassung in starkem Maße an die Präsenz einer natürlichen dialogischen Interaktion gebunden ist. Ferner weist es darauf hin, dass Sprecher über einen Zugang zu Gedächtnisspeichern mit vielfach indexierten Sprachexemplaren verfügen, die situationsbedingt umgehend abgerufen werden können.

Summary

Phonetic convergence describes the phenomenon in which two people interacting with each other get closer to each other's pronunciation. This tendency for more synchrony in the phonetic domain of speech covers changes in segmental as well as suprasegmental features. The purpose of this study was to investigate phonetic convergence in a second language environment, namely in native-nonnative dialogs between speakers of German and English. Crucial for the analysis was to determine to what extent the individual factor of *phonetic talent* influences the outcome of naturally occurring convergence in dialog.

The origin and purpose of convergence in general, and specifically in the area of pronunciation, has been investigated since the 1970s. Back then, convergence, or more generally speaking, accommodation was predominantly analyzed from the angle of social factors influencing the outcome of the process. The Communication Accommodation Theory (CAT)[5] proposed that all positive or negative shifts in someone's behavior (including phonetics) are conditioned by the need to, respectively, reduce or create more social distance. Newer accounts, such as the Interactive Alignment Model[6], present a mechanistic theory of how accommodative processes arise, without the discussion of the social factors involved. The apparent exclusivity of both theories, concerning the controllable vs. automatic dichotomy, has been the reason for many disputes in the field.

Research into phonetic accommodation in a second language, has, first of all, required the consideration of general theories of identity and identity negotiation and the means employed to do this, basically speech style with a special focus on phonetics. In addition, the special conditions given by the native and nonnative status of the speakers and their dictinct competences needed to be brought into focus as well. Since convergence seems to be a very variable phenomenon, the possible impact of several individual differences has additionally been discussed.

[5] see [GP75, GO06].

[6] see [PG04a].

Accommodation is a highly speaker- and context-dependent process, and its underlying mechanics are therefore best described within a theory accounting for those multiple situational variables. An account which was found to be especially suitable for such a purpose is a usage-based account of language – exemplar theory[7]. Since exemplar theory foresees a rich indexing of speech in memory, including speaker- and context-specific details, it allows for the modeling of variability in speech and speech processing. These features turn exemplar-based accounts into convenient frameworks for the description of all processes underlying phonetic convergence.

In order to investigate the degree to which talent affects convergence in a second language setting, the speakers were involved in a main dialog task and a read speech pre- and post-test. Twenty German speakers were paired with two native speakers of English, a male speaker of American English and a female speaker of Standard Southern British English, in two consecutive dialogs. At the end of each dialogic interaction, the German subjects were additionally asked to summarize the findings of the task. Before and after each dialog the German subjects were asked to read out a word list with target words from the dialogs, serving as a pre- and post-test. The dialogs were quasi-spontaneous task-oriented interactions elicited with the Diapix[8] picture-matching game. The applied method allowed for the collection of natural speech, with balanced amounts from both speakers, which also contained a sufficient number of repetitions of the relevant target words.

The acoustic measurement was based on the extraction of amplitude envelopes from the speech signal. The unit of analysis were words, compared at three different points in time within the dialogs – early, late and summary – and across the three readings of the word list. The amplitude envelopes of the words of both dialog partners were matched against each other using a cross-correlation function

[7] see, e.g., [Joh97, BH01, Pie01].

[8] see [vEBBB+10].

to estimate the degree of their spectral similarity. The match values obtained thereby were the basis for the statistical analyses.

The main hypothesis about the involvement of talent in phonetic convergence was confirmed: the talented subjects displayed significantly higher convergence toward their native speakers than the less talented subjects between an early and a late point in the dialogs. Gender, on the other hand, was not a significant factor for accommodation in neither the dialog nor the read speech task. The native speakers of English were informed about the purpose of the study and were asked to control their pronunciation in order not to display positive[9] shifts toward the nonnative speakers. Despite the request for maintenance, both native speakers showed on average significant convergence toward the German subjects. Any longer-lasting effects of the dialog convergence could not be confirmed – neither for the summary part, nor for the read speech pre- and post-test. This indicates that convergence from the dialog did not carry over to other speech styles, such as a first person narrative or read speech.

The above findings suggest that convergence mechanisms require a hybrid model to account for processes functioning largely subconsciously ands also those components that can be partially consciously controlled. Talent apparently influences the core of phonetic convergence mechanisms in an L2, and is probably connected to the joint network of attention and memory, responsible for the storage, processing and selection of exemplars. Moreover, the lack of convergence carrying-over to monologic speech styles not only suggests that phonetic accommodation is strongly tied to the presence of a natural dialogic interaction but also that speakers have access to memory pools with richly indexed speech exemplars that can be instantly retrieved if situationally required.

[9] or negative shifts.

Chapter 0

Introduction

> ”I gradually came to see that phonetics had an *important bearing on human relations* – that when people of different nations *pronounce* each other’s language really well (even if vocabulary and grammar not perfect), it has an astonishing effect on bringing them together; it *puts people on terms of equality*, and *a good understanding between them immediately springs up*.”
>
> Daniel Jones [JCM03]

Pronunciation indeed is a special part of the acquisition of a second language. It often allows us to identify a person, sometimes after only a few words are spoken. The phonetics of a language are probably the most prominent window to a person’s identity. This identity often stands wide open, although many of us would rather have it closed and double-locked. It can bring people together, as Jones said, but it can also be a reason for negative attitudes and prejudices. It can also be a source of resentment. And, while some might curse the apparently unchangeable nature of their accents, others are left wishing for more stability in their pronunciation, which often skips beyond their control as soon as they find themselves communicating with another person.

This loss of control over our own pronunciation is connected to the phenomenon of phonetic convergence. It describes a process in which the pronunciation of directly interacting partners becomes more similar to each other. Investigating convergence[1] in a second language environment adds one component to the equation not present in convergence between native speakers, and that is the differences in mastering the pronunciation of the foreign language. The special status of phonetics in the process of acquiring a second language has led to the assumption of a distinct talent component responsible for a person’s success in the L2 phonetics. This talent factor might also be involved in the mechanism controlling phonetic convergence.

[1] The term *convergence* usually stands for a style shift toward a conversational partner, while *divergence* indicates a shift in the opposite direction. In this case, however, *convergence* is used to mean *accommodation* or style shifting in general, since one could simply assume the existence of positive and negative convergence.

The evaluation of the exact nature of the relationship between talent and phonetic convergence is the purpose of this study.

The analysis of convergence or accommodation requires the inclusion of a multitude of research aspects. It is not only the *what* but also the *how* and the *why* which is behind the process needing to be clarified. Those features depend both on the situational context and the conversational addressee. The multitude of theoretical models and viewpoints concerning how to approach accommodation or style shifting (or even what terminology to use), turns the delivery of a full account of the aspects relevant in phonetic accommodation in dialog into a balancing act.

What remains clear is that, apart from the production or output component (the "what"), where we consider phonetic aspects, and through the procedural component that is to shed light on the functioning of the link between production and perception (the "how"), we still need a starting point - an answer to our "why". This starting point calls for a closer look at the socio- and psycholinguistic background of communication in general and of dialogic behavior in detail. Studying the above in a multilingual setting adds to this complexity, since SLA mechanisms need to be incorporated into the considerations as well. Consequently, all questions of identity construction, native and nonnative speaker competences, basic theories of accommodation, and individual differences between speakers will be addressed in **Chapter 1** on the **sociolinguistic motivation for accommodation**.

Modern linguistic and phonetic research has been moving more and more in the direction of usage-based approaches, while at the same time moving away from structuralist models. Exemplar theory is grounded in a usage-based account of language, grammar and language change. Within this usage-based framework it has been postulated turning to the observation of linguistic performance instead of describing directly assumed underlying linguistic competence. It is one's experience with language that is taken to be central, with the cognitive organization of this experience eventually building up grammar[2].

[2] [Byb06].

Implications following as a natural consequence from this view include the change of the description models for linguistic categories, away from traditionally used abstract rules, processes and structures, towards actual patterns of occurrence of those linguistic categories. Such *rich memory* or *exemplar models* prove to be a very good means for modelling naturally-occurring accommodation processes. A detailed account of the basic mechanisms for exemplar storage and retrieval as well as the multiple connections to convergence and its subcomponents are presented in **Chapter 2** on the **modelling of convergence in a usage-based account**. The last two sections of Chapter 2 additionally discuss the relationship between exemplar models on one side, and second language processing and the individual factor *talent*, on the other side.

After presenting the sociolinguistic background for accommodation and an exemplar-theoretic model of naturally-occurring convergence, **Chapter 3** will be concerned with the presentation of the **state-of-the art in measuring phonetic convergence**. This includes an overview of the current methodologies, experimental settings, measured parameters and the identified convergence effects. Crucial design differences between the current study and experiments that have been carried out in the past will be referred to in **Chapter 4** on the applied **methodology and data analysis technique** – the measurement of spectral similarity with amplitude envelopes.

The subsequent two chapters, **Chapter 5 and Chapter 6**, report the **results of the dialog experiment and the pre- and post-test**. A **discussion** of the presented results follows directly in **Chapter 7** – with separate sections for the dialog experiment and the read speech pre- and post-test. The **conclusion and outlook** in **Chapter 8** will summarize the findings, contribute ideas for related practical applications, and provide an outlook for future directions in studying phonetic convergence, especially in a second language context.

Chapter 1

Sociolinguistic motivation for accommodation

We are no doubt dealing with at least four interdependent layers underlying a two-person-encounter where neither shares the same mother tongue nor the same socio-cultural background. Firstly, the socio-psychological notion of identity and its negotiation within the dialog situation; secondly, the foreign language spoken in this situation, adding further complexity to the analysis; thirdly, the process of style shifting or accommodation in speech; and finally, the question of how individuals differ in their performance in a second language and possibly also in the amount of accommodation they engage in. All these layers will be thoroughly discussed in the present chapter.

1.1 Construction of Identity

Identity is no longer regarded as a static entity but as a changeable and fluid construct – it can be perceived as having undergone a change from being treated as a constant toward being seen more as a negotiable variable within sociolinguistic and sociopsychological research [Par07]. Instead of defining the identity or *face* of a person based solely on information about this person, researchers have been looking at who this person is interacting with and the nature of the interaction, as factors with probably the greatest explanatory power for what people sound like in a given moment. Blackledge and Pavlenko [BP01, 244] refer to many studies in which identity construction (in contrast to a mere expression of identity) is considered in a local environment and cannot be separated from its interactional context.

Identity is also embedded in and contingent on the broad situational setting, including geo-political aspects, social and economic changes, globalization, power relations, language ideologies, choice and attitudes and the way one's own identity and the identity of the others are posited and evaluated [PB04a, 1-2]. Pavlenko and Blackledge discuss a range of theories and approaches to identity and identity negotiation especially in multilingual contexts [PB04a, 4]. According to sociopsychological theorists, language was said to *equal* ethnic identity, or, as in later studies (as e.g. Giles and Byrne 1982 [GB82]), language was viewed as one of the *salient* *markers* of ethnic identity and group membership. The approaches, though, have attracted criticism for assuming a general state of monoculturalism and monolingualism as being a default, rather than reflecting the more complex identities and linguistic diversity of bi- and multilingual people. The authors also point to the fact that language not always only reflects ethnic identity but can also serve as a means of communication in professional situations or in the work place – in which case it forms only a part of a larger and much more complex identity [PB04a, 5].

Further criticisms have been levelled at the assumption that weak in-group relationships and open group boundaries facilitate faster assimilation of a second language and a higher proficiency level. General criticism of this pure sociopsychological account has thus been expressed [PB04a, 6]:

> Recent research in second language acquisition clearly demonstrates that the relationship between individuals' multiple identities and second language learning outcomes is infinitely more complex than portrayed in the sociopsychological paradigm and cannot be reduced to few essentialized variables(...).

The limiting factors in the sociopsychological approaches and their unidimensionality have been said to hinder the recognition that there are many social contexts that can limit or stop individuals from taking up new identities or accessing linguistic resources [PB04a, 7].

Although the interactional sociolinguistic approaches deal primarily with language choice and code-switching phenomena and do not directly concern the second language[1] learner and the second language acquisition[2] process, they nevertheless provide us with a definition of identity that is also applicable in an SLA context [Gum82]. Social identities are seen as fluid and not constant, and as it has been pointed out earlier, as something constructed[3] in linguistic and social interaction and not automatically given [PB04a, 8].

[1] Subsequently also abbreviated *L2*.

[2] Subsequently also abbreviated *SLA*.

[3] Hence the name "social constructionists" used in the literature.

Post-structuralist explorations [Hel92, Hel95] add an important aspect to our considerations of identity in a second language environment. They draw our attention to the relation between language and power relations. Language can be used to influence other people by gaining access to and exercising power. Language choice in a multilingual setting, therefore, is always an expression of existing language ideologies and legitimized identity options [PB04a, 12]. In an attempt to bring together social constructionist and post-structuralist approaches, Pavlenko and Blackledge have presented a comprehensive definition of identity and the interplay between language and identity in discourse [PB04a, 14]:

> (...)we see identity options as constructed, validated, and offered through discourses available to individuals at a particular point in time and place (...). On the one hand, languages, or rather particular discourses within them, supply the terms and other linguistic means with which identities are constructed and negotiated. On the other, ideologies of language and identity guide ways in which individuals use linguistic resources to index their identities and to evaluate the use of linguistic resources by others.

Language and identity are thus interwoven, which leads to the emergence of an extremely complex picture when dealing with more than just one language (e.g. in bi- or multilingual groups), and most certainly as well in a second language learning situation, where the emerging linguistic competence and new identity options go hand in hand.

1.1.1 Identity in second language learning

The whole process of second language learning can be seen as a struggle for participation, according to Pavlenko and Lantolf [PL00]. This participation involves the life and culture of the newly-entered society or group and calls for finding or defining a partially or totally new identity for the language learner. This view broadens the traditional understanding of the SLA process, which focused on the linguistic resources only, thereby neglecting the sociological and psychological issues.

Sfard [Sfa98] introduces two metaphors for the learning process – the learning as participation (PM) and learning as acquisition metaphor (AM). Learning in the AM understanding is connected to acquiring knowledge, ideas, notions, senses, representations, filling our minds with new concepts, just as we would acquire material goods in the real world [Sfa98, 5]. In SLA this means grasping and mastering linguistic knowledge: grammar, lexis and phonetics, which, however, does not suffice in guaranteeing successful communication in the foreign language.

The participation metaphor, in contrast, incorporates the context of learning as well, which is "rich and multifarious, and its importance is pronounced by talk about situatedness, contextuality, cultural embeddedness, and social mediation" [Sfa98, 6]. The learner is now seen as a participant in certain activities and as someone becoming a member of a certain community who is able to "communicate in the language of this community and act according to its particular norms" [Sfa98, 6]. This includes the pragmatics and cultural usage-rules of that language[4].

Block [Blo07, 113] argues in favor of an even further-reaching distinction, addressing the whole problem from a different angle – the pragmatic angle. Approaching the problem from this perspective, he introduces the distinction between pragmalinguistics and sociopragmatics. The first term covers the essential linguistic knowledge needed to carry out speech acts, whereas the latter relates to the social knowledge for a concrete sociocultural context, on the one hand for understanding what is happening, and on the other hand for acting in accordance with those rules.

Pavlenko and Lantolf [PL00] have elaborated the process of learning as participation by examining the first person narratives of bilingual writers. They have defined two phases of language learning in a migration context that describe the struggle of people to learn a new language and integrate in the culture without losing all of their "old" identities [PL00, 162p.]:

[4] Meaning a.o. the cultural routines of greeting, congratulating, expressing grief and sorrow and ways of addressing conversational partners and politeness rules.

- the phase of loss[5]
- the phase of recovery and (re)construction[6]

Despite this new identity search in a migration and cultural immersion context being much deeper-rooted and involving an almost complete re-definition of the self in contrast to a mere SLA context, there are some parallels to be found. Even if the L2 is used only during holidays or in a professional situation, the L2 learner still has to resign from his usual identity option *native speaker of language X* and cope with being seen as a *nonnative speaker of language Y* and treated accordingly.

Park [Par07, 1] points to an important relation issue by stating that both native and nonnative speaker identities "are social categories that are made procedurally relevant to the ongoing interaction and that consequently invoke an asymmetrical alignment of the participants", meaning that any conversation between an L2 learner and a native speaker of that L2 bears a status inequality, subject to negotiation. We will return to this issue in more detail in 1.1.3 (the mechanisms of identity negotiation) and in 1.2 (the underlying competence differences between NS and NNS relevant for negotiation and accommodation).

The next section will be devoted to the linguistic means (or, the speaking style) influencing the construction of identities within a dialog and also their subsequent negotiation.

1.1.2 Style

As laid out before, one's identity and one's language(s) are intrinsically intertwined. A particular linguistic repertoire is the reflection of one's available and chosen identity options, while at the same time the usage of this style already leads to a redefinition and renegotiation of this identity. Both notions, therefore, can only be seen

[5] Including a.o. the loss of one's linguistic identity, loss of the link between the signifier and the signified, and first language attrition [PL00, 162].

[6] Including a.o. the emergence of the person's new voice and translation therapy through reconstructing the past [PL00, 163].

as dynamically changing and context-dependent[7] entities (e.g. Pavlenko and Blackledge 2004 [PB04b], Coupland 2001 [Cou01]).

Style, being itself an extremely ephemeral notion, also encompasses many disciplines and approaches, just as the notion of identity does. Thus, before we attempt to define style, we need to become aware of the complexity of its nature. Coupland has argued for a multi-perspectivity where neither the theoretical understanding of style nor the individual stylistic performance are limited by any particular empirical or interpretive procedure [Cou01, 186]:

> A more broadly conceived "dialect stylistics" can explore the role of style in projecting speakers' often-complex identities and in defining social relationships and other configurations of context. [Cou01, 186]

He also demands that language be seen as a bidirectionally operating entity, being not only conditioned by social situation but at the same time defining the social encounter [Cou01, 189]. Style itself is seen "as situational *achievement*, and as the fulfillment of communicative purposes (whether consciously or non-consciously represented) in relation to those social situations" [Cou01, 189].

In order to overcome the narrow meaning of style, where *style* is equated solely with *dialect style*, Coupland [Cou01, 189] suggests differentiating between the following:

- dialect style[8]
- expressive or attitudinal style[9]

Many aspects which would not have found a place in an analysis of dialect style variation alone can be analyzed by assuming much broader boundaries for the concept of style (as e.g., forms of address, lexically-expressed formality, politeness, dominance in conversation, degree of self-disclosure) [Cou01, 189]. An even broader definition of style was introduced by Hymes [Hym74] under the term "ways of

[7] "context-dependent" here means: speaker-, situation- and context-dependent.

[8] features linked to social group/class differentiation.

[9] features not associated with social group membership.

speaking", which includes the presence of underlying patterns of ideational selection[10].

Style has also repeatedly posed more as a process than as a static entity which we assign the "quality of 'thing-ness '" [Cou07, 2]. The main focus of research therefore should be confined to understanding "how people *use* or *enact* or *perform* social styles for a range of symbolic purposes" [Cou07, 3]. Coupland [Cou07] here is comparing social styles to resource packages which can serve to express multiple personal and interpersonal meanings. Hence, his view of linguistic style remains tied to a processing (or usage) view rather than to a product (finite state) perspective [Cou07, 3].

The levels at which a user of language has (or has to *make*) a certain choice of style are extremely multifaceted. We can draw a line between *dialect* and *register* (the first implying geography), where the latter is "the semantic organisation of linguistic choices taking account of communicative purposes and circumstances" ([Hal96] in [Cou07, 13]). The choices one can make range from ideational selections[11], textual selections[12] and interpersonal selections, which are said to relate to the social distance between speakers[13] [Cou07, 13].

While the investigation of *register* has rather been neglected these days, the notion of *genre* has found an established place in sociolinguistic research. Coupland [Cou07, 15] sees genres as "culturally recognised, patterned ways of speaking, or structured cognitive frameworks for engaging in discourse". We could, for instance, think of many types of different genres on an institutionalized-personal scale, including genres such as a politician's speech, a university lecture, a sports interview or the show of a stand-up comedian. Coupland also points to the fact that the more personal the nature of a communicative act gets (e.g. small talk, an argument or story-telling), the harder (though still not impossible) it is to categorize it clearly as a disctinct genre. One important feature of a genre is that there are certain demands

[10] in Coupland's words [Cou01, 190]: "what we choose to mean, to whom, when, and where."

[11] i.e. which topics, things, facts or reports to choose from.

[12] i.e. ways of applying deixis, sequencing, or communicative mode/manner.

[13] i.e. expressing attitude and varying the communicative tone.

upon its design which make it recognizable and easy to label for its participants, who entered the communicative situation with certain expectations and probably also an accommodation to the encountered framework or context [Cou07, 15].

This recognition, however, is usually not a matter of generally applicable objectivity. What is common to all subcategorizations of style is that they are all settled in context. As has already been pointed out, context is not only limited to situational context but rather expands to personal context as well. Here we enter an area where our objective understanding of style is put to a test, since many experiments have shown objective measurements of style change not to be what individuals perceive and take as their starting point in a conversation (e.g., Bourhis et al. 1979 and Thakerar et al. 1982, [BGLT79, TGC82]). As Giles [Gil01, 214] underlines,

> our perceptions of, and labels for, speech style – and intra-individual variations of it – are subject to our social expectations and contextual knowledge (...)

which causes an immense need for a subjective operationalization of style change. This is in fact a very important finding for studying convergence in conversational speech, since what we might expect to find is a bias between objective measurements of how a person behaves linguistically and how this is interpreted by the conversational partner (compare [Gil01]). We could e.g. hypothesize that a speaker perceived his partner to speak with a typical given accent[14] based solely upon the information about the geographical origin of this person (as in our case, USA vs. Great Britain). Whether nonnative speakers of a language already have active access to all these choices and variations in speaking styles is yet another question to be considered (see 1.2 for further details).

[14] being indexed with many concrete features gathered through years of experience with speakers of the target language, see 2.1 for a detailed explanation of the exemplar based storage process of linguistic knowledge.

1.1.3 Identity negotiation

Just as the descriptions of identities have changed from being static and fixed to being dynamic entities [BP01, Bel99, Bel01, Blo07], the view concerning the negotiation of those identities has changed as well. It is always defined as an active process – not a passive context-given script that allows no changes once the situation is set. Ting-Toomey [TT99, 40] defines it as a "transactional interaction process, in which individuals attempt to evoke, assert, define, modify, challenge and/or support their own and others' desired self-images". Another view of negotiation is presented by Davies and Harré [DH90] and adopted by Blackledge and Pavlenko [BP01, 249], where it is defined as "the interplay between reflective positioning, that is, self-representation, and interactive positioning, whereby others attempt to reposition particular individuals or groups". Whatever definition of identity negotiation one examines, one common denominator is the activeness and *inter*activeness of its nature, where it is not only the speaker herself who is in the center of attention but rather the questions of where, when, how and *with whom* this speaker is interacting. Especially phonology should be seen as being "fluid and skillfully deployed by individual speakers" [BP01, 244].

Identity negotiation (or identity construction) in the course of an interaction depends upon many factors (which Davies and Harré term *subject positions* [DH90]), e.g., race, ethnicity, gender, generation, sexual orientation, geographical and political reality or even institutional affiliation [BP01, 249]. These subject positions continuously underlie changes as well and can be created anew in every encounter.

So how does a negotiation of identities come about and which means can a speaker employ? Apart from the possibility of applying various nonverbal changes (as e.g., the way someone is dressed or the choice of the place of conversation), verbal communication (= speaking style) is the most powerful means of identity construction [Bel01, BP01, Cou07, GCC91b]. Bell [Bel01, 141] defines style very suitably as everything a speaker does with a language in relation to other people, and considers it to be an active and interactive process designed primarily for and as a response to the speaker’s audience. According to Bell, these changes in style are not just unidirectional. He distinguishes between audience and referee design, where the former indicates a shift towards a present communicative partner or, in the case of more recipients, the audience. By contrast, the latter indicates an accommodation to a person or group not necessarily present at the moment of speaking (ibid.). What needs to be kept in mind, however, is that the social situation not only unilaterally influences the language used but that the language itself is an active determinant of this situation as well [Cou01, 189]. This can be interpreted as follows: identity negotiation and consequently also style shifting are not only highly dynamic processes which allow for on-line changes but they in fact *demand* the participants in such an encounter to react immediately to the changing situation. Whether or not a speaking partner is able to react in such a way, depends, among other things, heavily upon their linguistic competence, and, as will be proposed later on (see 1.4.2), also on their language talent.

1.2 Native and nonnative speaker competences

Identity is a crucial issue in second language acquisition, especially when it comes to pronunciation. A nonnative speaker´s accent “is part of one´s sense of identity and personality”, and all three aspects underlie a mutual influence, with the accent mirroring both personality and identity but also being formed by them [Maj01, 66]. However, the concept of a native and a nonnative speaker bears a linguistic dimension as well, defined by the speaker´s language ability and knowledge. This fact, combined with an obvious difference in this very language competence, results also in disparate identity options being available for native and nonnative speakers.

1.2.1 The native speaker

Without entering the problematic discussion of who should be called a native speaker – a person born in a certain geographical area or a person being able to participate as a fully-fledged member in a certain linguistic community [PL00, 169] we turn to a description of the "state-of-the-art" competences of such a native speaker[15] that are necessary for a full understanding of the NS-NNS relation in a discourse.

As Davies [Dav03, 205] defines it, the term *native speaker* refers to a "group whose idiolects show certain formal and codified norms", in other words the Standard Language. This standard form, he further elaborates, encompasses a certain ideal model for NNSs to imitate, aim at and also serves as a reference norm to be judged by, where the ideal can reflect one specific person, an élite group or even a text [Dav03, 205]. In Bartsch's [Bar88] terms, the ideal model in the (standard) language is called a *point*, but since that would not allow for any variation within the standard variety, he defines an additional *range* that allows for a sufficient amount of tolerated idiolects [Dav03, 205]. This defines the NS as a person belonging to a speech community, speaking a model variant of their native language or an idiolect within the range of the native language, which is the desired target of the nonnative speaker. Being a member of such a (standard) language group also means sharing a set of norms. It also ensures mutual intelligibility among all members (ibid.).

A native speaker is further capable of telling the difference between his or her own native language and dialects of that language, and has a creative capacity to invent neologisms and to judge them acceptable or not according to the rules of the native language [Dav03, 89]. Davies also hints at an interesting phenomenon here: neologisms invented by a native speaker are never regarded as mistakes, whereas similar creations of nonnative speakers would most probably not be accepted and judged to be errors [Dav03, 90]. This seems to be a further disadvantage for a nonnative speaker[16], since he or she cannot (regardless of their possibly already great grammatical and vocabulary skills) "play" with the language in a creative way

[15] Henceforth also abbreviated *NS*.

[16] Henceforth also abbreviated *NNS*.

to the same extent as a native speaker would be allowed to. This, in fact, demands a very high degree of vocabulary mastery, correctness and contextual appropriateness combined with a firm control of grammar, and NNS usually do not possess the same skills at rephrasing grammatical constructions or paraphrasing mistakenly chosen words and phrases or compensating for the lack of an appropriate word or phrase in a certain context [Dav03].

1.2.2 The nonnative speaker

All the above-mentioned competence areas also affect the nonnative speaker, just, in a different manner. The social dimension in the concept of a nonnative speaker arises from the choice of identity or membership in a certain group or community, affected to a huge degree by attitude [Dav03, 11]. Apart from the competence level in the various language skills and the cultural knowledge needed, the NNS has the option of whether to belong to the native speaker community in which he finds himself or not.

The target to achieve in a second language is usually connected to gaining an appropriate communicative competence, which can be characterized as the knowledge of "how to seek appropriateness and how to recognise it, how to match background knowledge and context in such a way that messages are understood and understandable" [Dav03, 91]. Furthermore, discourse should have the properties of being understandable and coherent. This is, as Davies argues, obviously not only dependent on the correct use of language in terms of grammaticality[17], but also on an appropriate use of language in view of the present situational and discourse context – and of course the dialog partner ("being responsive and accommodative towards the dialog partner" [Dav03, 91]).

Given such a potential error source for nonnative speakers and the fact that using language creatively and inventing neologisms would also probably not be accepted by native speakers, this is an additional source of stress for the L2 learner. The

[17] Phonetic aspects, of course, do play a crucial role in becoming an *understandable* dialog partner in a second language.

NS notices all errors of an NNS and furthermore judges errors uttered by another native speaker as correct, labelling them instead language creativity. This could put the NNS in a situation with several possible outcomes, depending to some extent on their raw linguistic proficiency and to some degree on psychological features and sociological issues of membership. In general, one can say that native and nonnative identities reflect "expert-novice identities" [Par07, 342] in terms of the possessed linguistic and communicative knowledge. Those are usually asymmetrical.

One way of behaving in such an asymmetrical situation is to try to eliminate the disparity and regain a state of equilibrium [Par07]. However, the question comes up as to whether a nonnative speaker really wants to pass for a native speaker (which would pose him as *equal* in terms of status) with all its consequences or maybe feels comfortable with retaining signs of her nonnativeness, and having a foreign accent. The lack of cultural knowledge could be one of the reasons for actually retaining one's foreign accent since it allows for a *safe* position in the interaction[18] [Dav03]. Davies [Dav03, 72] argues for exactly such a safeguarding behavior in native-nonnative interactions:

> (...) non-native speakers may, in practice, prefer to rest at some level of approximation, to choose fossilisation, because it suits them to be outside, not indistinguishable, because then the kind of expectation I have been suggesting is not made of them.

One's accent is definitely a crucial (and sometimes it is even the only) factor giving away one's origin and identifying one as a nonnative speaker (for most speakers probably involuntarily). Although, for some speakers it might be the result of a conscious choice, which Davies [Dav03, 72] denies is only negative.

> This may be the explanation for the foreign accent which many adult immigrants retain, the only sign perhaps of a non-native origin but it would be

[18] being an accepted *alibi* for certain cultural (or grammatical) mistakes. One could imagine (or even have already experienced) a situation in a foreign country where one is mistakingly taken for a native speaker due to perfect pronunciation and then addressed accordingly, which unfortunately exceeds one's linguistic competence in that foreign language and leads to massive communicative problems.

> wrong, in my view, to regard this as necessarily a disadvantage for users since what it can also mean is a choice of identity and they have chosen not to belong to the native speaker community of the speech community they now reside in.

Holmes [Hol92, 258] also provides some evidence that retaining an accent does not need to have negative consequences. It can even be beneficial (e.g. for actors, comedians in their performances, or tourists when needing help in a foreign country), as the reverse situation – sounding too much like a native speaker – can lead to unfriendly, suspicious or even hostile behavior[19].

In the following section we will turn to describing how dialog partners, whether they be native or nonnative speakers, accommodate to each other and negotiate both their identities and their linguistic repertoires (or fail to do so, consciously or unconsciously).

1.3 Accommodation in dialog

So far, we have surveyed the questions of identity and competence of native and nonnative speakers, focusing on the features of the person entering a communicative situation. We will now adopt a rather *processing*-oriented perspective and devote our attention to the processes that occur during such a person-to-person interaction. As we are here describing phenomena at the microsocial level, we will not explicitly refer to macro-level factors[20], but one should nonetheless bear in mind that they are part of the contextual framework for every dialogic interaction.

1.3.1 Communication Accommodation Theory

What was first conceived of as a model of interpersonal accommodation “where a speaker makes certain linguistic adjustments in the direction of his partner as a

[19] e.g., the disapproving reactions toward a French-English bilingual speaker discovered by a French-speaking group to speak perfect French and also perfect English [Hol92, 258].

[20] as e.g., attitudes toward languages and countries, the political situation etc.; see also Davies and Harrés’ subject positions’ [DH90] in 1.1.3.

means of facilitating social attraction"[21] [STG76, 374] has since undergone considerable development to form a "model of relational and identity processes in communicative interaction". This encompasses many disciplines, some of which are not mere linguistics [CJ97, 241]. As opposed to Labov, who argues that the reason for stylistic variation lies in the varying degrees of attention speakers give to their own speech [Lab01, Bel07], followers of Communication Accommodation Theory (CAT) have searched for an explanation in social psychological and motivational processes. Labov's definition, however severely it was challenged in many later publications on accommodation, contained at least one grain of truth: namely the factor of attention. Attention in CAT, however, plays a totally different role and operates at a different point in time than Labov initially proposed, i.e. essential here is the attention directed towards the conversational partner's speech[22](rather than to one's own speech), which seems to be crucial for all later processing steps in accommodation/convergence (compare [Pie01, Gil01, GP07, PGA10]).

Communication Accommodation Theory in its present form puts forward the claim that the reason for individual adjustments in communication lies in the wish to create, uphold, or reduce social distance. CAT is said to provide a means of investigating and explaining (and even, to a certain extent, predicting) the outcome of these changes [GO06, GCC91b, GP97, PG08]. Pitts and Giles define the primary goal of CAT as follows [PG08, 18]:

> Communication accommodation theory is primarily concerned with the motivation and social consequences underlying a person's change in communication styles (verbal and nonverbal features such as accent, volume, tone, language choice) to either accommodate or not accommodate their interactional partners.

The essential tenets of Communication Accommodation Theory are described as follows [GO06, 294]:

[21] back then termed 'speech accommodation' [Gil73, GP75].

[22] see 2.2 for more details.

- A communicative situation is influenced by the present situation, the interactants' initial orientations and goals, and additionally by the socio-historic context.

- Identities[23] can be actively negotiated through accommodation.

- People are already entering the communicative encounter with certain expectations as to the "optimal levels of accommodation".

- Communicative partners can behave in three distinct ways: converge to their partners, diverge from them, or maintain their own style - which are the available communication strategies.

Convergence serves to *decrease* social distance between the partners by means of adapting to one another's behavior in a positive way, so as to become more similar to each other. For this purpose a wide range of elements can be altered in discourse, e.g. speech rate, accents, pause duration and utterance length, and even nonverbal behaviors like e.g., gestures, facial expressions, smiling, etc. [GO06, 295]. Convergence can be furthermore linked to "seeking affiliation, social approval, compliance, and communication effectiveness" [PG08, 19]. Pitts and Giles further specify that communication accommodation can also have a cognitive and affective function. Cognitive purposes comprise a.o. accommodation for enhancing comprehension and preventing misattributions and misunderstandings [PG08, 18]. Convergence may thus be used to reduce linguistic dissimilarities and to become more alike, which in turn affects the speakers' attractiveness ratings, predictability and perceived supportiveness, along with intelligibility and interpersonal involvement, which is subjectively perceived (by the recipient) [GCC91b, 18].

It has been observed in some studies that objective, measurable convergence need not equal the interactants' perception and rating of the direction and level of

[23] Giles and Ogay use the term *social category memberships*; identity, of course, exceeds these solely socially motivated limits. However, the re-positioning of a person within a conversation clearly seems to encompass more than only category membership, as has been argued in 1.1.3 and 1.2, also confirmed by Giles' and Powesland's usage of the term "identity-change" when referring to accommodation [GP97, 233].

accommodation (see e.g., [GCC91b, SGLP01, TGC82]). This *perceptual* or *subjective* convergence of the speaking partner has been shown to correlate with one's need of gaining his or her social approval (compare Larsen, Martin & Giles 1977 [LMG77]). Perceiving someone to be closer or more similar to one's own behavior is said to clearly facilitate "real" convergence, since the target does not seem to be as far away anymore [GCC91b]. Respectively, a misinterpretation of the partner's actual behavior may have negative effects on the communication. Thakerar and colleagues have found evidence for the reversed situation to take place as well: sometimes the speaker's intentions do not match the eventual outcome in accommodation, which they termed *psychological* accommodation [TGC82].

Divergence operates conversely, namely in the direction of *increasing* social distance to the partner (due to an expression of social disapproval or the need to underline one's own distinctness) by means of seeking a particularly distinct manner of behavior (e.g., by insisting upon a regional accent). Divergence can be compared to what Bell [Bel01] termed referee-design, since a movement away from the physically present speaking partner could actually be interpreted as a movement toward some absent third person.

Similarly to divergence is *maintenance*, where one interactant persists in his own speaking style and does not accommodate to the partner. This could result both from an insensitivity to the other's behavior or from a purposeful choice to affirm one's own identity or autonomy in a rather deemphasized way [GO06]. As often stressed, this is usually evaluated negatively [GO06, 295] (compare also [GGJ^{+}95, GCC91b, SGLP01]).

A similarly negative impression can be caused by exaggerated convergence (see e.g., 'overaccommodation' [PG08], 'overshoot' or 'hyperconvergence' [GCC91b], Giles and Ogay 2006 [GO06]). An example of overaccommodation would be patronizing speech by using overall slower speech and simplified grammar (towards older people, see [WN01, GO06]). Such an overshoot might also be associated with foreigner talk, where native speakers underestimate the nonnative speaker's competence and consequently approach him or her with an exaggeratedly simplified speaking style (compare the section on ‘foreigner talk’ later on and, e.g., Ellis 1985 [Ell85]).

Accommodation to a speaking partner has also been characterized in terms of reciprocity, modality and direction [SGLP01, 37f.]. Convergence and divergence can therefore proceed unidirectionally or mutually, the latter indicating both partners changing their behavior according to their interactants’ behavior, the former only a change in one direction, with one of the partners retaining her idiosyncratic style. Accommodation can also be described as multimodal (happening across different modalities, e.g., verbal and nonverbal) or just unimodal (happening at one level of behavior only). In terms of direction, it is possible to differentiate between *upward* or *downward* accommodation. Upward adaptation, on the one hand, is usually connected to striving for a higher, more prestigious variety of speech, whereas downward accommodation refers to a change toward a less valued, possibly stigmatized variant of speech. Shepard and colleagues also point to a distinction between *partial* and *full* accommodation, where *partial* refers to only ‘slight’ convergence and *full* indicates ‘exactly matching behaviors’ [SGLP01, 37][24]. These notions have not been further elaborated, however, as shown in studies in phonetic imitation, no two pronunciations of the same item are ever equal[25] (see e.g. [SF97]). This renders anything termed an ‘*exactly* matching behavior’ and anything thought to be a nearly perfect copy in the phonetic domain pretty much impossible.

What is common to all studies within Communication Accommodation Theory is their focus on underlying social explanations, implying that the process is at least partially of a controlled and influenceable nature. However, it has not yet been definitely stated whether *all* accommodation processes are controllable, and if they are not, which types might be susceptible to social and psychological influences and to what extent. Due to the highly dynamic process of identity construction and its negotiation (see 1.1 and 1.1.3 for details), a closer look at the *mechanisms* possibly

[24] See also [Bou91, GW96].

[25] This is even true despite the two words or phonemes being uttered by the same speaker. For an explanation of ‘noise’ in the imitated speech signal, compare Pierrehumbert 2001 [Pie01] and Chapter 2 with section 2.1.

guiding accommodative behavior in a dialog might also shed more light onto its dynamic make-up.

1.3.2 Interactive Alignment Model

So far, accommodation has been analyzed by asking about the underlying motivations. Answers have been found in social and psychological factors, in research mainly connected to the Communication Accommodation Theory framework. What CAT does not focus on is the clarification of the exact mechanism of accommodation. This has been taken up by Pickering and Garrod [PG04b, PG05, PG06], who have proposed a mechanistic theory of alignment in dialog.

The interactive alignment account proposes that dialog partners are affected by a totally unconscious and highly automatized (biologically-founded[26]) drive to become more alike, regardless of group membership, status differences or social attractiveness. Pickering and Garrod claim that the alignment of interlocutors on many distinct linguistic levels is basically an automatic process. The goals of this process are clearly to simplify both comprehension and production in a dialog situation by building a 'common ground'[27] [PG04b, 170]. The interactive alignment model (IAM) proposes that

> (...) in dialogue, production and comprehension become tightly coupled in a way that leads to the automatic alignment of linguistic representations at many levels. We argue that the interactive alignment process greatly simplifies language processing in dialogue. It does so (1) by supporting a straightforward interactive inference mechanism, (2) by enabling interlocutors to develop and use routine expressions, and (3) by supporting a system for monitoring language processing.

[26] Compare also Kelso's and Oullier's studies concerning human bonding and (social) coordination dynamics, which a.o. explain social 'coordinating' behavior in terms of a biologically founded drive one can find in any kind of human behavior [Kel97, KE06, Kel09, OdGJ$^+$06, OdGJ$^+$08].

[27] The term goes back to Stalnaker 1978 [Sta78].

Pickering and Garrod define alignment at a particular level of speech as the state in which the dialog partners have the same representations at that level. Dialog for them therefore is a coordinated behavior[28], with the underlying representations being aligned to each other. For successful dialogic interaction to happen, they assume an alignment of situation models to be the first step (at least to an approximate degree). A situation model is defined as a multi-dimensional representation of the situation under discussion, with the following basic dimensions: space, time, causality, intentionality, and reference to the interactants or individuals under discussion [ZR98, PG04b]. Alignment at this global level is suggested not to be overtly negotiated but rather to stem from an alignment at lower ('local') levels of linguistic representation, turning it into a bottom-up process [PG04b, 173].

> We propose that this works via a priming mechanism, whereby encountering an utterance that activates a particular representation makes it more likely that the person will subsequently produce an utterance that uses that representation.

As the authors stress, this process is basically resource-free and *automatic*. Garrod and Anderson [GA87] found evidence for alignment at a pragmatic and semantic level and assumed the interactants were only able to stay in a state of balance or 'equilibrium' when the production targets of one partner matched the representations of the other. Pickering and Garrod [PG04b, 173-175] also advocate for alignment on a lexical and syntactic level, with an additional possibility of 'percolation' between the levels, meaning that alignment at one level can lead to alignment at another related level. Clark and colleagues [BC96, CWG86, WGC92] found lexical alignment by showing that dialog partners use the same set of referring expressions in order to refer to particular objects, and, moreover, that these words become shorter and more similar after repetition[29] but change when the dialog partner is switched.

[28] which they compare to, e.g., ballroom dancing, [PG04b, 172].

[29] Similar evidence was found by Bybee [Byb02] for historical changes in lexical development, which will be discussed more detailed in Chapter 2 on exemplar models and processes.

The study of Branigan et al. [BPC00] illustrates the activation of representations by priming at a syntactic level. The authors claim that it is indeed the underlying representations being activated. In other words, it is not only further unspecified production or comprehension procedures that are activated. As Pickering and Garrod [PG04b, 174] state more precisely, this suggests a close relation, or as they put it, “an important parity” between perception and production targets (see also Goldinger 1998, [Gol98]). As Wilkes-Gibbs and Clark [WGC92] claim, priming and subsequent alignment does not occur with all interacting partners to the same extent. When comparing *the audience*[30] to the (directly referred to) *dialog partner*, alignment proved to be stronger for the addressee than for other listeners present (also termed the “side participants”). According to Pickering and Garrod, only the direct addressees need to fully activate their production systems to be always ready to make a contribution to the dialog.

The interactive alignment account foresees alignment at an articulatory level as well. As mentioned earlier, according to the findings of Clark and colleagues [BC96, CWG86, WGC92] and also Bybee (2002) and Fowler and Housum (1987) [Byb02, FH87], repeatedly-used expressions in a dialog tend to become shorter, more reduced and even harder to identify when heard in isolation. Bard et al. (2000) [BAS$^+$00] showed that this reduction not only appeared in the speech of one speaker but of all speakers involved, which led Pickering and Garrod to the conclusion that “whatever is happening to the speaker’s articulatory representations is also happening to his interlocutor’s”[31] [PG04b, 174].

Pickering and Garrod have proposed their own model of comprehension and production processes in dialog: the interactive alignment model (see Figure 1.1 and 1.2), which they contrast with the autonomous transmission account (e.g., Levelt 1989 [Lev89]). As illustrated in Figure 1.1, the information flow between the various levels of linguistic representation in the autonomous transmission model allows

[30] here used in contrast to Bell’s terms not as *all* people present at the moment of speaking *including* the directly-addressed dialog partner, but only the not-directly addressed individuals present [Bel01].

[31] A full discussion of the state of the art of convergence measurement is provided in Chapter 3.

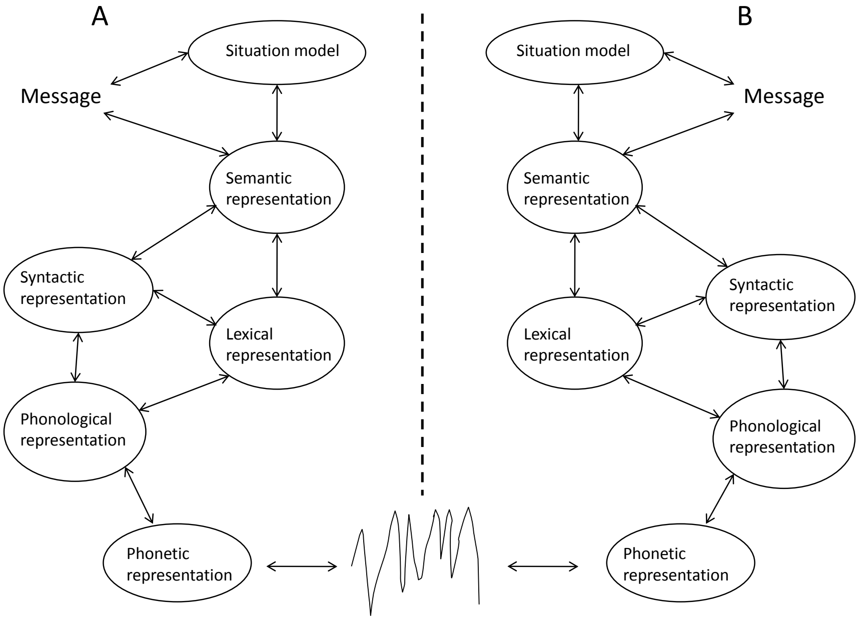


Figure 1.1: Autonomous transmission model, with no links between interlocutors. Pickering and Garrod (2004), [PG04b, 177].

for interaction between the levels, but only speaker-internally. This account is thus rendered void if it is representing dialog, since it does not capture the interaction between perception and production [PG04b]. The interactive alignment model, on the other hand, includes possibilities of interaction not only concerning one individual but also across speaker-listener relations, as shown in Figure 1.2. The horizontal links between the two interactants indicate the ‘channels of alignment’ . The mechanism along which alignment proceeds in these channels is said to be priming, and is assumed to happen in a direct and automatic way, according to Pickering and Garrod, [PG04b, 177]:

> There is no intervening “decision box” where the listener makes a decision about how to respond to the signal. Although such decisions do of course take place during dialogue (...), they do not form part of the basic interactive alignment process, which is automatic and largely unconscious.

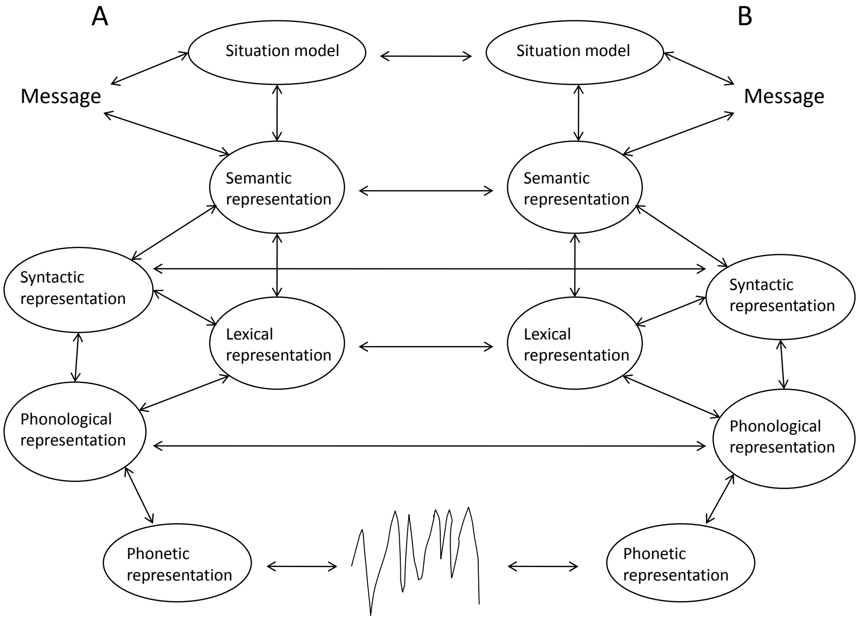


Figure 1.2: Interactive alignment model, links between interlocutors present at all levels. Pickering and Garrod (2004), [PG04b, 176].

Pickering's and Garrod's [PG04b, PG06] theory has raised some criticism though, largely due to their view that no 'intervening' steps are possible in alignment. Communication Accommodation Theory is based on the assumption that accommodation is a means of expressing social attitude. Without the possibility of an intermediate step between comprehension and production allowing one to distance oneself from the dialog partner, no divergence would be possible. As Krauss and Pardo argued in their commentary on Pickering's and Garrod's paper [KP04], in order to capture social processes a *hybrid model* would be required. Such a hybrid account would incorporate alignment deriving from automatic processes in accordance with the IAM, as well as more directed and reflective processes, accounting for socially-motivated changes in the dialog. Newer research is providing more and more evidence for the need of hybrid models. Giles and Ogay [GO06, 294] hinted at but did not elaborate on the fact that social interaction is a matter of balancing the desire of being regarded as the same in some points but at the same time as different in other features. This suggests that the interactive alignment model might account for a more automatic component of accommodation, in the direction of establishing common ground and therefore reducing the distance between interlocutors. It does not, however, rule out the possibility of different components becoming activated in a dialog situation, turning it into a highly dynamic process[32].

1.3.3 Accommodation for increased intelligibility

Accommodation in the CAT model is, on the one hand, a means of negotiating social distance, but on the other hand it can also function as a way of enhancing intelligibility within a conversation. The interactants may want to be (better) understood and use convergence to decrease the linguistic distance from their partners and make the interaction run more smoothly [GP97, GGJ$^+$95, 234]. This can also be compared to the process of gaining 'equilibrium' [GA87] or the establishment of an implicit common ground, advocated by Pickering and Garrod [PG04b], which is a necessary prerequisite for successful communication. Here, however, the focus lies clearly on improving intelligibility by adjusting linguistic properties.

This desire to be more comprehensible affects both interacting partners in a native-nonnative encounter, and not only the NNS as one might think. However, this mutual convergence of both interacting partners is characterized by different features, which are related to the identity of a NS and NNS and the resulting dissimilar status they have in the interaction. Nonnative speakers are often recognizable as speaking with a foreign accent and this accent can interfere with intelligibility in a native-nonnative discourse. A native speaker equipped with the linguistic and communicative competence of the target language will thus probably be assigned a higher status than the nonnative speaker in an NS-NNS dialog, leading to a possible upward accommodation of the nonnative speaker and a downward accommodation of the native speaker.

[32] For more on the dynamics of alignment see Chapter 3.

In order to overcome this inequality, a native speaker often adopts an easier speaking style, so-called *foreigner talk*. Ellis [Ell85, 135] presents a list of many features (subsumed under the headings “interactional modifications” and “input modifications”) a native speaker uses. What is of special interest for us here is his list of modifications concerning pronunciation. This list comprises: slowing down speech, separate word/syllable articulation, more careful pronunciation in general, heavier stress, increased amplitude on words crucial for understanding (standard modifications), and possible vowel insertion to a consonant coda, less vowel reduction or exaggerated intonation (among the non-standard modifications [Ell85, 135]).

Davies also argues that the alternations in speech that NS are capable of and are used by them are meant to simplify the language and thus enable a more efficient decoding of salient features in order to enhance overall intelligibility [Dav03, 48,201]. Ellis [Ell85, 138], on the other hand, suggests other possible explanations for NS to adjust their speech, including regression, matching and negotiation, with the latter being his most favored possibility. In either instance, what the NS most probably does in a NS-NNS conversation is to downgrade his pronunciation, which allows him or her to negotiate the role and status relations at a lower level.

1.4 Individual differences

Giles [Gil01, 218] makes an important point about accommodation and its communicative value, namely that it is *communicatively competent* speakers who adapt themselves to their listeners. These speakers are thus called ‘optimal accommodators’ in the CAT framework. This statement opens the floor for discussion of many related issues, as e.g., what this competence actually means, how it is achieved, and why there are persons who succeed in becoming optimal accommodators while others fail. What the level of such an ideal accommodation (both for native and nonnative speakers) could mean in practice, has already been described in Chapter 1.3. What remains is a discussion of the features turning accommodation and especially convergence at the phonetic level into a highly individual issue, not only dependent on the need to negotiate social distance, but also on the very basic level of linguistic competence. However, as Giles [Gil01, 219] asks, "is having the repertoire sufficient" to become an effective style-shifter? What may thus at a first glance look like a simple correlation between one's proficiency in an L2 and accommodation, should maybe in fact be analyzed with the target one level further down, i.e. amongst the factors accounting for *individual differences*[33] in style, especially when considering second language acquisition[34].

Individual differences were often pushed in the background of the second language acquisition process, and were held 'responsible' for its modification and also personalization. In latest research the direction has turned towards both a situated and process-oriented perspective, from which a subtly different picture emerges, accounting for flexibility and context-dependency of learner variables [D1̈0, MMC09, UD09]. Dörnyei (2010) argues for a model incorporating this considerable amount of variation within learner attributes and proposes moving back from a modularized view to a model where IDs, such as aptitude and motivation, are seen as constructs comprising a number of constituent components themselves ("multicomponential view of L2 ID factors") [D1̈0, 252].

Personality features, as motivation and aptitude, belong to the group of endogenous factors[35] in second language acquisition[36]. We will start with an overview of the latest findings concerning the influence of motivation, and then move on to a discussion about how much variance in the degree of foreign accent can possibly be accounted for by language talent[37].

[33] also abbreviated IDs.

[34] As will be noted in this section and in Chapter 2 on exemplar-based models, some of the factors discussed here are not SLA-specific but rather of a general phonetic nature, related to the processing of newly incoming speech signals, no matter the language, dialect or speaking style that is being considered.

[35] As opposed to external/exogenous factors ('biographical variables' [Bir06]): age of learning onset and L1 background.

[36] Although Major [Maj01, 66] subsumes *all* these factors under the heading "personality of the individual", we prefer to keep personality features (e.g. empathy, ego permeability, self-esteem, risk taking, introversion vs. extraversion) apart from motivation and aptitude.

[37] Language *talent* and language *aptitude* will be used interchangeably.

1.4.1 Motivation

Motivation seems to correlate with all aspects of SLA, including pronunciation. As Major [Maj01] points out, motivation to learn a language and acquire its pronunciation stands in a mutual reinforcement relationship to success in achieving this. Success can therefore strengthen motivation, whereas a failure could weaken it and vice versa. This is also the case where motivation is intertwined with personality factors, namely the BIS-BAS [HR09]. The motivation lost due to failure to accomplish goals or the amount gained through experiencing success depends on one's tendency to be put off by negative experiences (as expressed by the behavior inhibition system – BIS) or animated by positive ones (as reflected in the behavior activation system – BAS).

Motivation is sometimes hard to capture, and various studies used different questions to assess their subjects' motivation to learn an L2 (e.g. by rating the importance of good pronunciation in their professional or private life [Sut76, Moy99]). Gardner and Lambert [GL72] proposed a distinction between two types of motivation – *instrumental* and *integrative*, which fit into a social-psychological framework and take attitudinal factors as its basic underlying variables [DÏ0]. Instrumental motivation, on the one hand, is a type of motivation with specific goals in mind, where the language is used as an *instrument*: e.g. learning the language to get a better job or a job abroad, to communicate with business partners, to get better grades at university, etc. Integrative motivation, on the other hand, denotes a drive toward becoming a member of the target language community, toward *integrating* into the society of L1 speakers. Gardner and Lambert [GL72] have argued that integrative motivation has the stronger influence on achievement in SLA, leading to higher proficiency and a better pronunciation, in order to allow the learner to become 'indistinguishable' from native speakers. However, there is evidence in favor of instrumental motivation having just as strong an effect on second language learning as the integrative type [GM91, GDM92]. Major [Maj01, 67] comments on that fact with the following explanation:

> (...)the distinction between integrative and instrumental motivation is not clear-cut. Integrative motivation can be thought of as the sum of all various instrumental motivations (...); thus the difference between instrumental and integrative motivation is a matter of degree, not kind.

Most studies utilizing motivation as a variable in SLA found it to be significantly correlated with foreign accent in the learned language, e.g. Suter and Purcell [Sut76, PS80]. Flege [Fle95] determined that integrative motivation and a factor called "concern for L2 pronunciation" are both significant variables for predicting accented pronunciation, even if they accounted for only 3% of the variance in the pronunciation ratings of his male subjects (the variable did not seem to affect females). In another study, Flege and colleagues [FYKL99] found a similar result, with even less than 3% of the variance explained by instrumental and integrative motivation. Moyer [Moy99] identified a strong correlation between the degree of foreign accent and the factor "professional motivation" in her English L2 learners of German. However, only one of her 24 subjects was rated to speak within a native speaker range of accent.

Within a more process-oriented perspective, several scientists have proposed that motivation should be seen as a continuously changing construct that never remains stable throughout the learning process [DÏ0]. Dörnyei [DÏ0, 251, emphasis in the original] further clarifies this claim, introducing a new concept of individual differences (IDs), within a process-oriented and situated[38]perspective:

> IDs were usually seen as background learner variables that modified and personalized the overall trajectory of the language acquisition processes, accounting for *why, how long* and *how hard* (motivation), *how well* (aptitude), *how proactively* (learning strategies) and *in what way* (learning styles) the learner engaged in the learning process. (...)we simply cannot fail to realize that the various learner attributes are neither stable nor context-independent, but dis-

[38] a perspective that takes into account the direct learning context and its influence on the learner's disposition [DÏ0, 251].

> play a considerable amount of variation from time to time and from situation to situation.

He also makes a very clear point about the exact nature of the variable *motivation*, which he perceives not as a "monolithic" construct but instead as consisting of several constituent components - which he dubs a "multicomponential view of L2 ID factors". Aptitude is, in his opinion, deconstructable in the same way, being, just as motivation, a "complex, higher-order attribute" [DÏ0, 252].

1.4.2 Aptitude

Aptitude or talent has long been banished from research and curricula, as it implies taking on a variationist perspective with far-reaching consequences for the teaching and learning of foreign languages. It is of course much more comfortable to assume learners to be equally endowed and it also guarantees equal chances at achieving proficiency in an L2 rather than allowing for individual differences, calling for suitable course material and adequate teaching methods [Ske03]. Nowadays, several decades after the first aptitude studies were conducted, no one doubts[39] the existence of an innate aptitude surfacing in the outcome of L2 proficiency – but being distinct from proficiency [Jil09b]. Skehan [Ske03, 187] assumes the following:

> (...) language aptitude is stable in nature, is not susceptible to easy training or modification, and is not environmentally influenced, to any significant degree, at least after the early years. (...) language aptitude is something we are endowed with as a set of cognitive abilities which are either genetic or fixed fairly early in life.

Skehan does not deny the possibility of environmental influences or the impact previously learnt languages have on the acquisition of a new language, but those are not supposed to change the underlying construct of aptitude itself. Another important issue in aptitude research is the question whether there is a distinct talent

[39] Except a few critics, such as Neufeld [Neu79] or Krashen, who have argued that aptitude is only relevant in an instructional framework [Kra81].

for the acquisition of languages or only a domain general aptitude that influences other types of learning as well. To test this, one needs to look for a dissociation of language aptitude and typical measures of general aptitude, as e.g. intelligence (IQ).

Amongst the pioneers in aptitude research were Carroll and Sapon [CS59, Car81], who designed the Modern Languages Aptitude Test (MLAT), and Pimsleur [Pim66], who invented the Language Aptitude Battery (PLAB). Carroll [Car81] proposed that the crucial factors composing aptitude are the following:

1. Phonemic coding ability – the ability of analyzing sounds in a way that allows for their subsequent storage.

2. Associative memory – the ability to associate one type of verbal material to another.

3. Inductive language learning ability – the ability to find structure in, and derive rules from natural language material.

4. Grammatical sensitivity – the ability to determine what function a word has in a phrase or sentence.

Studies that used the MLAT to test aptitude and successful performance of subjects after a training period (usually an intensive language course) found significant correlations of the two factors, mostly within a range of 0.40 and 0.65 [Ske03]. The only other factor with comparable (though still lower) explanatory power for learning success proved to be motivation. Skehan also draws our attention to the fact that Carroll's four sub-components seem to be well justified, since they mirror necessary skills in the SLA process[40]. A cluster analysis of previous aptitude measurement results [Ske86] allowed Skehan to conclude that there

[40] Efficient auditory processing, e.g. matches phonemic coding ability and associative memory, and can be decomposed into the process of memorization and correct later retrieval. However, Skehan notices that associative memory in Carroll's terms is heavily based on a behavioristic account with stimulus-response paradigms, and therefore has not stood the test of time in its original form [Ske03, 192].

seem to be two types of successful L2 learners, namely those relying on their high linguistic-analytic abilities and those who utilize their very high capacity for memorizing language material. There is surprisingly little overlap between the two groups. Skehan [Ske03] draws a parallel between such analytic vs. memory skills and the contrast between rule-based and exemplar-based learning. The former is usually associated with a syntax-dependent learner type, readily working with a rule-based system, while the latter is said to meet the requirements of a rather exemplar-based learner type, which is associated with lexical learning.

This division is in line with a modular view of talent, implying different components, one of which is language talent. Following this modular model, it is not far-fetched to see language talent itself as being further subdivided into subfaculties. One of the most established divisions is the one based on Schneiderman and Desmarais' study [SD88]. They proposed a separate talent for grammar and for accent. This two-fold nature of talent is commonly known as the Joseph Conrad or Henry Kissinger phenomenon, where perfect abilities in grammar and vocabulary have no counterpart in pronunciation, which remains heavily accented [BPS95, Gui90, Jil09b]. One of the neurophysiological explanations for such a distinction assumes that the greater difficulty in mastering L2 phonology (apart from an obviously essential neural plasticity) stems from the need to override already established L1 motor pathways in pronunciation, a problem a learner does not face in the case of vocabulary or grammar [SD88]. It has also been shown that phonetics seem to be one of the skills affected very early by maturational constraints. Another dimension of the specialty of pronunciation talent comes, as argued in chapter 1.2.2, from the tight connection of accent with identity and personality. As Guiora [GBD72, 112] argued, pronunciation is linked to a construct he referred to as "*language ego*", which has the following characteristics:

> (...) language ego too is conceived as a maturation concept and refers to a self-representation with physical outlines and firm boundaries. Grammar and syntax are the solid structures on which speech hangs, lexis the flesh that gives

> it body, and pronunciation its very core. Thus pronunciation is the most salient aspect of the language ego, the hardest to penetrate (to acquire in a new language), the most difficult to lose (in one's own).

While Guiora's latter assumption that one's own pronunciation is the most difficult aspect of language to lose, has been rather rejected[41], the claim that pronunciation acquisition is linked to ego-permeability has been tested and confirmed[42]. That, of course, does not imply that it is pronunciation talent that can be controlled by manipulating the level of ego permeability or self-consciousness, but only the phonetic performance of a person at a given moment.

In spite of overall agreement that language talent is a multi-componential construct, the amount and nature of its sub-components continue to be debated. In contrast to Carroll's original four-component division, or the general distinction between talent for grammar vs. talent for accent, Skehan has proposed a three-component system [Ske03, 201]:

- auditory ability (corresponding to Carroll's phonemic coding ability)
- linguistic ability (corresponding to inductive language learning ability and grammatical sensitivity)
- memory ability (corresponding to associative memory)

Skehan's reformulation of the term 'associative memory' into 'memory ability' is the result of a general twist in memory research. Associative memory is now conceived of as being only one sub-component of general memory. Memory researchers currently agree on three involved steps: encoding, storage and retrieval, while Skehan emphasizes *retrieval* as the one crucial stage in memory and language aptitude (and not *encoding*, which most studies have concentrated on so far) [Ske03, 202]:

[41] Pronunciation shifts, even for the mother tongue, have been reported, for example, by Sancier and Fowler with bilingual Portuguese-English speakers [SF97], and point to the vulnerable character of phonetics and its susceptibility to influences of the language of the surrounding.

[42] Ego-permeability was successfully increased to some extent applying hypnosis (Schuman et al., 1978), alcohol or valium (Guiora's studies in 1972 and 1980) [SHCW78, GBHB^{+}72, GAES80].

"What we need to investigate is the nature of the system which can support rapid access of a very wide repertoire of exemplars[43] so that real-time processing is possible". Many different memory constructs have since been proposed and studied as having a crucial contribution to language (e.g., working memory and language in Gathercole's and Baddeley's work [GB01]; and the relation between attention and working memory in SLA, in Robinson [Rob03]) and language aptitude as well (e.g. Dörnyei & Skehan's work [DS03, Ske02]). Some have even proposed that working memory can be seen as the "central component of aptitude" [MF98] or even the only factor accounting for the predictive power of aptitude tests [McL95]. Robinson deconstructed the memory variable in SLA into the following components [Rob05, 52]. The first two abilities and the first aptitude complex relate to spoken language, while the others relate to written language processing:

- Abilities
  - Phonological Working Memory Capacity (PWMC)
  - Phonological Working Memory Speed (PWMS)
  - Text Working Memory Capacity (TWMC)
  - Text Working Memory Speed (TWMS)
- Aptitude Complexes
  - Memory for Contingent Speech (MCS) – connected to PWMC and PWMS
  - Memory for Contingent Text (MCT) – connected to TWMC and TWMS

Since the role of memory has been shifted toward the retrieval stage of processing, the input stage in Skehan's model is occupied by the phonemic coding ability (see Table 1.1). Phonemic coding ability has often been left aside as being a trivial and

[43] Exemplar acquisition is used by Skehan in the sense of lexical learning and has been compared to acquiring "ready-made 'wholes' " rather than a rule-based generation of items or a "computed performance". See Chapter 2 and 2.1.2 for more details.

Aptitude factor	**Stage**	**Operations**
Phonemic coding ability	Input	Noticing
Memory	Output	Retrieval – ‘computed’ performance – exemplar-based performance

Table 1.1: Aptitude and processing stages, modified from Skehan 2003 [Ske03, 203].

self-explaining component of aptitude[44] and therefore has not received much attention in publications on language talent. However, it is exactly this module which seems to be essential in learning pronunciation [Ske03, 203]:

> This is important in processing input (...), handling the segmentation problem (...), and coping with auditory material in real time, with its coding and analysis, so that it may be passed on to subsequent stages of information processing.

It is also this ability that allows the successful L2 user to decide whether input is noteworthy or can be neglected; input is, firstly, linked to the above-mentioned operation of noticing. It is the one major step in the processing chain on which all subsequent steps rely. Interestingly, Pierrehumbert [Pie01] proposed *attention* as the first step in the process of exemplar gathering (preceding *recognition* and *coding*[45]), a step intimately intertwined with noticing. Noticing will be discussed in more detail in Chapter 2.

Summarizing, phonetic talent seems to be composed of a bundle of abilities, some located at the input processing stage - starting with undisturbed auditory abilities as a premise and the capacity to notice important linguistic information and tell it apart from mere noise or blur - to the more central processing stages of encoding and storage, and ending with the output stage, where stored phonetic information needs to be retrieved from memory. Skehan stresses an important point about the benefits of an exemplar-based route over a rule-based access system at

[44] For example, by Krashen who stated that this component is “simply” connected to the ability to store new sounds of a language in memory and does *not* directly relate to learning [Kra81, 19].

[45] However, there is plenty of confusion about the meaning of those terms. Robinson’s concept of *attention*, for example, includes responsibility for the processes of input encoding, keeping it available in working memory and its retrieval from long-term memory [Rob03, 631].

the output stage in natural conversations, despite learners relying on both methods depending on the task. He states that although the exemplar route might be less flexible and rely on chunks and redundant storage, its advantage lies in fast and convenient access, forming, in his opinion, the basis for both native-like selection and fluency [Ske03, 204]. If we acknowledge the importance of speed of retrieval for fluency in general (presupposing of course efficient and correct storage), its equally important role in conversational speech (where speed and accurate retrieval of suitable forms is expected[46]), seems all the more obvious.

[46] 'Suitable' here refers to a receiver-responsive design of one's own utterance, as described in detail in Chapter 1.3.

Chapter 2

Modeling convergence in a usage-based account

Bybee [Byb06, 711] has proposed that the input “the general cognitive capabilities of the human brain (receive), which allow it to categorize and sort for identity, similarity, and difference” are the specific linguistic events a person encounters. These are then categorized and stored in memory. The change of the description models for linguistic categories, away from traditionally used abstract rules, processes and structures to actual patterns of occurrence of those linguistic categories (see Wade et al., Bybee [WDS$^+$10, Byb02, Byb06]) calls for suitable new ways of describing and interpreting the observed multitude of data. Exemplar-based models provide exactly such a formal means of description, assuming that all the various level categories (be they phonemes, syllables or words) consist of a collection of actually-experienced instances of those categories. The processes underlying perception (or identification) and production, then, only operate on an exemplar level by comparing the items within and between collections. Further specification of occurrence regularities or surface forms of the exemplar categories is not necessary [WDS$^+$10]. Usage-based accounts like exemplar-based models also have the explanatory power to deal with discrete and gradient phenomena (e.g. phonetic neutralization, word frequency- or gender- and speaker-dependent acoustic differences). They are moreover suggested to “provide the most accurate, parsimonious description of linguistic competence and performance” [WDS$^+$10, 1] [Byb02, Byb06, Joh06, Pie01, Pie06].

2.1 Basic mechanisms of exemplar processing

As usage-based accounts have developed, many different approaches have been proposed to explain how exemplars are acquired and stored in memory [Joh97, Pie01, Byb02, Haw03]. Slightly varying suggestions have also been made regarding which exact speaker and situation details are being stored and what form these memories take. The main strands of research will be summarized in the following.

Exemplar theory first emerged as a model in psychology and was further developed for speech processing and subsequently re-modelled by Goldinger (see e.g.

[Gol96, Gol98, Haw03, Joh97, Joh06, Pie01, Pie06]). At present, exemplar-based accounts are being used in phonetics and phonology [Haw10, WDS^{+}10], as well as in semantics, lexicology, typology [Byb02, Byb06], syntax [Bod06] and language acquisition [AST06].

Johnson [Joh97] proposed a model of speech perception where exemplars are seen as associations between a set of auditory properties and a set of category labels, the former defined as output from the peripheral auditory system and the latter as including any classification of possible importance available to the perceiver at the moment of storage in memory (such as gender, speaker name, etc.). When a new item is encountered, the process of categorization involves:

- comparing the new item's auditory properties with each exemplar's auditory properties,
- assigning each exemplar an activation level according to its similarity to the new item – the better the match, the higher the activation level,
- summing up the overall activations of all exemplars of a given category.

The last step serves as a basis to decide whether the newly-encountered item should be categorized as an instance of that category or not. Johnson's exemplar-based model [Joh97] differs from previous perception models in several points. All speaker-specific details are retained in the set of exemplars, which allow for comparing and categorizing new items with reference to appropriate stored exemplars on speaker-specific dimensions. Johnson also added an attention weight parameter to his model that controls the degree of sensitivity to particular auditory properties. It has been suggested that no further speaker normalization processes are needed in this kind of perception model because "the model retains the variability encountered in speech [and thus] it is able to cope with the variability that it encounters in new tokens" [Joh97, 162].

Pierrehumbert [Pie01] has defined each category as represented in memory by a large cloud of remembered tokens of that category – the exemplars. After identifying a new token, it is categorized in a cognitive map such that similar exemplars are close to each other and very dissimilar ones are far apart. The exemplar system then works by mapping points in a phonetic parameter space and the corresponding labels of the categorization system [Pie01, 140].

An important emergent property of exemplar models is related to word/syllable frequency. Given that every linguistic experience is categorized and stored in the exemplar space, more frequent categories will automatically have a larger representation of tokens and less frequent categories will have a less numerous representation. Assuming further that linguistic memory decays and more recent memories will be more vivid than those from several years ago, Pierrehumbert proposes that each exemplar be assigned an associated strength, or, in other words, a resting activation level. Exemplars of newly-stored frequent experiences have higher activation levels than exemplars of temporally remote and infrequent experiences. This plays a crucial role in the classification process of new tokens, since it is not only the distance from any given exemplar in the parameter space that contributes to computing the similarity to a new token but also the strength of that exemplar. After perceptual encoding, the new token is placed in the relevant parameter space, where the computation of distance and the most probable labeling take place. The classification is only influenced by the set of exemplars located in a fixed size neighborhood of the token. The last step consists of calculating the summed similarities to the exemplars for each label present in that neighborhood, with the similarity to the exemplars weighted by their activation level [Pie01, Lac97]. The label favored in this process is the one having more or higher activated exemplars in the neighborhood of the new encountered token. This predicts that high frequency categories that are represented by more numerous exemplars with on average higher resting activation levels will have an advantage in the labeling process.

2.1.1 Units in exemplar models

As pointed out by Bybee [Byb02, 272f], her analysis is based on the assumption that words are the standard units in exemplar models that must be present in memory storage in order for the described changes to happen. Other accounts have posited the syllable, morpheme or even a multi-word string as the unit to be categorized in memory, seeing the introduction of variability as a top-down process proceeding in a hierarchical way towards the lower level. Newer accounts (such as in Pierrehumbert 2006 [Pie06]) suggest that the lowest level of description should be a parametric phonetic map instead of any set of discrete categories.

The Context Sequence Model, an exemplar-based production model recently developed in Stuttgart, however, assumes that frequency effects could be based on the constituent articulations (such as vowels and consonants) composing these frequent sequences. Thus, compared with their less frequent counterparts the differences lie in these lowest-level units [WDS$^+$10]. This does not imply a total negation of higher-level units being present in exemplar-based production models, but it does point to a new account for the described frequency effects, via a model incorporating (acoustic) context. Simulations by Schütze and colleagues [SWWM07] and Walsh and colleagues [WSMS07] showed that syllable length differences in production might be driven by an exemplar-based process involving competition between units at neighboring levels of an organizational hierarchy. First, one chooses a complete syllable exemplar from the memory store. If there are not enough data available at this level, the system turns to the constituent-level information. Incorporating surrounding segment context into the model is assumed to lead to faster and easier retrieval of a suitable segment-level exemplar, since it should be similar to many sequences in memory and the best 'match' in this case would probably be a segment produced originally in the same syllable [WDS$^+$10].

Hence, the context-based production model proposed by Wade and colleagues [WDS$^+$10] suggests that not only the exemplar itself (the segment or syllable) but also the preceding and the following contexts are being considered and, moreover, do play a decisive role in choosing the right token for the actual production. The simulation results have identified a context size between 0.1 s and 0.5 s (the former applying to lower frequency contexts, the latter to high frequency contexts) as useful in trying to emulate human performance. Furthermore, it has been shown that the selection process within frequent contexts is indeed more efficient and quicker, supposedly due to a faster 'recycling' of the segments used, with fewer comparisons needed. As a result those segments displayed stronger variability and more influence of lenition processes than their counterparts in less frequent contexts. It has also been shown that syllable frequency effects do not require the storage of syllables as units. Thus context has proved to be more important than units themselves, since many properties of units which had to be stipulated (e.g. unit strength or frequency) are now emergent [WDS$^+$10].

A similar focus on the role of context can be found in Hawkins' Polysp[1] system [Haw03]. For Hawkins fine phonetic detail present *all over* the sensory signal is important. For her, information from all acquired exemplars (that is already multimodal in nature), is used to extract huge amounts of detailed linguistic[2] and paralinguistic information[3]. However, she assumes that linguistic categories, including phonetic categories, are emergent from our exemplar learning, rather than forming the basis for our learning themselves [Haw03, 398]:

> Phonetic categories (...) are self-organizing, emergent, context-sensitive, dynamic, and plastic throughout life. Given these properties, the mental structures corresponding to a linguistic system can differ between individuals, depending on their experiences.

Thus, what we are dealing with is, on the one hand, stored exemplars, and on the other hand, extracted information, organized in coherent clusters. A listener then takes into account *all* of the speech stream, embedded in context and tries to map it onto linguistic and non-linguistic structural knowledge that seems to be important at the moment of listening. The matching process proceeds probabilistically and the goal is to arrive at meaning as fast as possible and *not* to perform a complete linguistic decomposition of the signal, which is itself unnecessary for an

[1] polysystemic speech perception; Hawkins refuses to call Polysp a model yet and prefers to use 'system' instead.

[2] e.g., syllable structure, stress, phonological weight, word boundaries, grammatical status and segmental identity [Haw03, 389].

[3] e.g., voice quality, emotional and attitudinal information [Haw03, 389].

appropriate understanding [Haw03]. Once meaning (of a whole phrase, not necessarily single words) is accessed, there are two possibilities [Haw03, 389, emphasis in the original]:

> The listener might 'fill in' the rest of the linguistic structure, checking if it fits the memory of the actual signal satisfactorily (...). Alternatively, the listener may simply stop mapping the signal onto formal linguistic structure – hence some parts of any given constructed ('perceived') structure may be more complete than others.

Hawkins thus considers the identification of words and phonemes as simply a by-product of the mapping processes going on between experiencing an acoustic stream and arriving at its meaning, which may in some cases be important, but is not always indispensable for the meaning decoding of a whole utterance [Haw03].

2.1.2 Memory in usage-based models

An important issue within exemplar-based theories is the assumed huge memory load. Unlike accounts that assume a normalization process while listening that overcomes variance in the signal and matches all incoming stimuli to a canonical form, exemplars have been taken to comprise not only detailed acoustic and gender-specific data, but also multiple social indexes [Gol96, Gol98, Pie06], so that normalization is no longer necessary. However, storing every single exemplar encountered in a lifetime with all its additional information in memory is often rejected as being too complex a task for a human brain. Various experimental results, though, point to an astonishing ability of people to remember instances, along with many visual and auditory details (e.g. Goldinger 1998, Johnson 1997 [Gol98, Joh97]).

Inherent to many exemplar-based models [Pie01, WDS^{+}10] is the assumption of gradual memory decay, partially dealing with the issue concerning the too great memory load of detailed episodic information. Exemplars that have been stored at a distant point in time become gradually blurred and their resting activation level decreases compared to recently encountered tokens. Another account of exemplar memory takes it as corresponding not to a single perceptual experience, but to an equivalence class of perceptual experiences. This is suggested by the granularization of the parameter space in which the exemplars are placed: tokens differing in too small detail are assumed to be encoded as identical [Pie01]. Moreover, the exemplar "clouds" undergo a more elaborate process between experience and eventual placement in memory than proposed so far. Not every experience raises attention to the same extent; it has been suggested that people focus on events classified as being "most informative". That, of course, influences the later steps of recognizing and encoding an exemplar[4] [Pie06].

Hintzman's MINERVA 2 model [Hin86], tested a.o. by Goldinger [Gol98], is based solely on episodic traces, denying the existence of e.g. word prototypes. It assumes that all experiences are stored in memory as independent entries, inclusive of situational and contextual details, forming an "episodic lexicon" [Gol98]. Every word is thus represented by a corresponding array of traces in memory (dependent on its frequency), partially resembling each other and therefore redundant. In perception, every heard word activates all similar memory traces, the strength of which depends on the grade of resemblance between the stimulus and the memory traces. Consequently, an "aggregate" of all active traces is sent from long-term memory to working memory as an "echo" [Gol98, 254]. The echo might contain "richer" information than the stimulus, since it also includes conceptual knowledge. The proposed echoes are characterized by two properties [Gol98, 254]:

- *echo intensity* reflects the total activity in memory created by the probe (increasing with greater similarity and frequency).

- *echo content* reflects a unique combination of the probe and the activated traces (since each trace responds to its own degree).

Newer accounts, such as the Context Sequence Model proposed by Wade and colleagues [WDS$^+$10], assume that everything that has been encountered is stored

[4] more details following in Chapter 2.2.

additionally in *full context* and perception and production take as much of this context into account as is needed to find a match. In contrast to static models (such as Pierrehumbert's model [Pie01] described earlier), this is a dynamic model which takes into consideration both contextual and timing details. A temporal match with a stored sequence, therefore, is as important as a spectral match. The form the memory sequence takes is assumed to be of a spectro-temporal nature, with the speech signal divided into 4 or 8 frequency bands. Additionally, both contexts of the element under consideration (the left context has an acoustic nature, while the right context contains linguistic information) are taken into account for defining a "match" [WDS$^+$10]. Hawkins [Haw10] also emphasizes the role of context in exemplar models, which sometimes allows for a semantic analysis of the input without the necessity of an elaborate linguistic (acoustic) analysis.

Two of the currently accepted models of working memory (WM) and long-term memory (LTM) that could fit into usage-based models of information processing and storage are presented in Figure 2.1 and Figure 2.2. Baddeley [Bad03, GB01] proposed that WM has the following constituent parts: phonological loop, visuo-spatial sketchpad, episodic buffer and central executive (Figure 2.1).

The central executive is a "managing device" in this model, directing e.g. attentional[5] resources either to relevant stimuli in its surroundings or to stored information in LTM. The episodic buffer temporarily holds current items from both WM and LTM and allows for information integration. The two slave systems – the phonological loop and visuo-spatial sketchpad – are specialized and hold respectively auditory or visual information. The phonological loop is also said to be crucial in rehearsal of speech stimuli, not only those perceived in an auditory mode but in a visual mode as well (i.e. reading). Working memory has a totally different structure from LTM since it is responsible not only for holding information available and active for a given moment in time, but also for simultaneously allowing the processing of these bits of information [RAC10]. The way the WM works in an individual is therefore said to be related to the ability of thinking and solving problems. The possible rea-

[5] the role of attention will be explained in detail in Chapter 2.2.1.

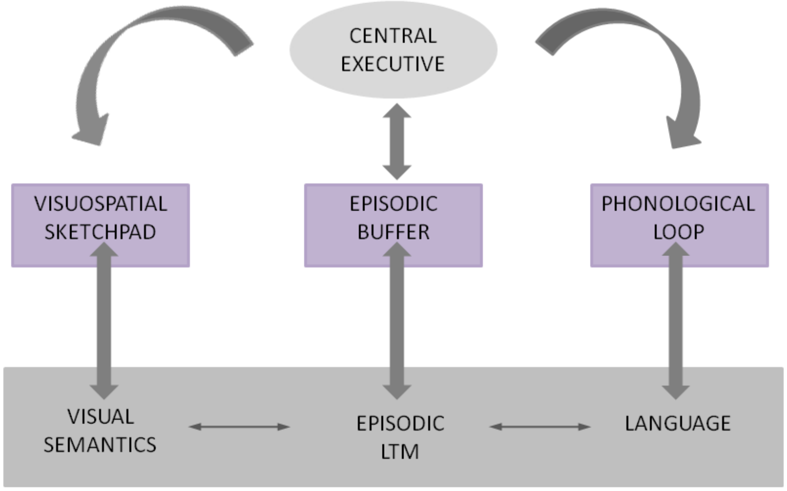


Figure 2.1: Baddeley's revised model of working memory, with the following main components: central executive, the two slave systems – phonological loop and visuospatial sketchpad, and episodic buffer and access to long-term memory. Edited from Baddeley 2003, [Bad03, 7].

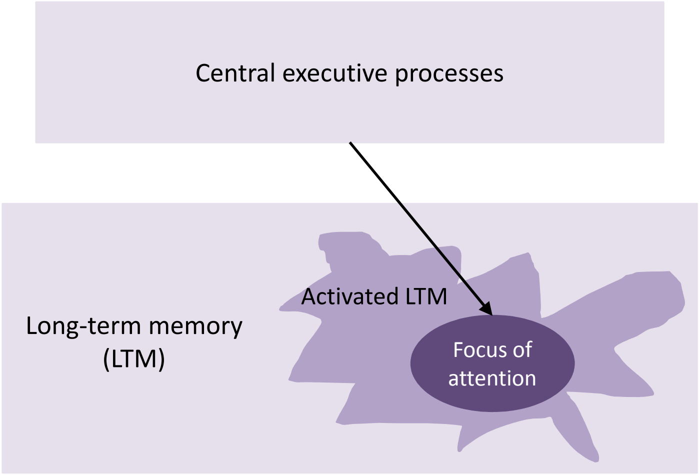


Figure 2.2: Cowan's working memory model [Cow88], which supposes that a certain limited amount of information is held in an active state – the *focus of attention*, rather than only as passively present information in LTM. Edited from Ricker et al. 2010, [RAC10, 574].

sons for IDs in working memory have been explained by Ricker and colleagues as follows [RAC10, 579]:

> One possibility is that individuals who demonstrate higher working memory spans have more efficient executive functions, so that the processing task consumes less attention and leaves more for storage. A second theory posits that individual differences in both processing and storage capacity can contribute to overall differences in working memory performance.

Research on IDs in working memory seems to confirm that WM is indeed a combination of the limited content held accessible and the processing component necessary for proper long-term encoding. Variance between individuals in WM tests also appears to surface in aptitude tests, problem solving and reading comprehension tasks [RAC10].

2.1.3 Frequency effects

Exemplar models of phonological representations that allow for gradual changes in both the phonetic and the lexical dimension have been proposed to account for phonetically conditioned changes in high- versus low-frequency words [Byb02]. Using an exemplar account of speech production, Bybee argues that reductive changes in vowels show a tendency to appear earlier and to a greater extent in high frequency words and phrases. This follows naturally from the assumption that exemplar clouds of high frequency categories show a greater density. Since their strength depends also on their recency, they display on average higher activation levels. Those exemplars then have a higher chance of being chosen for production and any already existing acoustic variation (such as lenition or deletion) can be strengthened or a new mutation can be initiated. Changes introduced at the level of the individual production accumulate over time and, considering the relatively quick re-use of exemplars within high frequency categories, also occur more rapidly. Low-frequency words, in contrast, seem to be affected by quite different changes [Byb02]. Less frequent words have been suggested to bend towards the stronger patterns of language (such as the regularizing of verb patterns), which affect them earlier than their high-frequency category neighbors. Although Shi and colleagues [SGKW05] suggest that some of Bybee's results might have been based on syntactic category differences (function vs. content words) and not on frequency alone, it holds true that the exemplar "clouds" are constantly being subjected to changes and updating processes while language is used [BH01, Byb02, Pie01, WDS$^+$10].

Such an explanation of varying effects contingent on frequency also bears some explanatory power for the difficulty of reaching an exact phonetic target. Random deviations from the acoustic target of one speaker caused by noise in motor control and execution seem to be very likely. Pierrehumbert [Pie01, 145] thus assumes that the process of adding new items to an exemplar pool could be a random sampling with added noise, meaning that recovering an exemplar for production does not guarantee an identical production of that item. Moreover, frequency effects in exemplar models seem to provide a straightforward explanation for social accommodation processes. It has been suggested that speech patterns which are heard recently and frequently automatically guide the typical productions within a speech community, therefore leading to the adaptation of speech patterns[6] [Pie01].

Frequency effects follow naturally from a usage-based account, meaning that:

1. the more exemplars are encountered, the more there are in memory
2. exemplars of newly-stored frequent experiences are assigned higher activation levels than exemplars of temporally remote and infrequent experiences.

This also seems to provide a natural explanation for alignment in dialog as it is understood by Pickering and Garrod [PG04b, PG05]. However, as will be elaborated in Chapter 2.2, the notions of automatic and controlled behavior are more multifaceted than the usually assumed "none-full" dichotomy[7], and a usage-based account seems to be compatible with more than just fully automatic approaches to phonetic convergence. A valuable prediction in relation to convergence and fre-

[6] a more detailed explanation is to follow in Chapter 2.2.

[7] being either fully conscious or completely unconscious.

quency of occurrence is made by MINERVA 2 [Hin86][8], suggesting that perceived high-frequency words are less likely to produce good imitations than their low-frequency counterparts. This has to do with the much higher number of *echoes* that are naturally activated for the former type, leading to a "mixed" output with less probability of an exact match constituting a large percentage thereof. In case of low-frequency words, by contrast, fewer traces present in memory are able to create an output form resembling the original stimulus to a much greater extent [Gol98]. Imitation, of course, does not equal convergence (imitation is a *fully* conscious and controlled action in a controlled setting, whereas convergence happens rather naturally and without full awareness or control) but we can take the predictions as being testable also in a convergence context.

2.1.4 Overspecification, underspecification and full specification of exemplars

When we consider the mechanisms governing exemplar storage described above, it is evident that we are dealing with a type of *overspecification* of exemplars, resulting from an extremely detailed and rich feature indexing. As literally any bit of information, whether it is paralinguistic, non-linguistic or linguistic, can be derived from the signal [Haw03], every single exemplar is highly overspecified. This holds true irrespective of the manner of indexing, since in all cases – even if the indexes are being derived and stored separately at a hierarchically higher level (as in Hawkins' model) – the detailed information must be present and available during exemplar access (for both perception and storage). Hence, what we have, is a huge number of exemplars to choose from, equipped with extremely rich specifications.

Another way of understanding specification is from the angle of the choices we make in speaking. When we find ourselves in a conversational situation, we can theoretically choose any item or phrase from our exemplar pool that fits the right meaning in context. The virtual item which is to be chosen is thus *underspecified* in

[8] described in more detail in Chapter 2.1.2.

its nature. In fact, a speaker is facing a lot of "noise" in his or her memory pool, which appears to complicate and slow down the decision process at first sight. However, a choice needs to be made and it is by no means at random. The decision to choose one exemplar over another is guided by the detailed indexing present, and, as Hawkins claims, an experienced language user makes use of this indexing and picks items that fir the current context. The choice is, of course, mediated by frequency effects and current activation patterns [Byb06, Pie01].

Consequently, the representation one has is simultaneously overspecified in terms of feature encoding and highly underspecified in terms of which exemplar to use for the current production target. Only when the exemplar is accessed and produced in the conversation does it become *fully specified* in the ongoing situational context. The full specification naturally applies only to the speaker at first; the listener still needs to decode the exemplar, which does not always mean a full linguistic decomposition, as Hawkins argued [Haw03]. Therefore, no full specification would theoretically be needed. A focus on linguistic form at this stage, however, could make a huge difference for convergence, as will be argued in Chapter 2.2.

2.2 Exemplar theory and accommodation

The direct link between accommodation and an exemplar-based approach has not been explored yet in the literature. In this section, the multiple points of contact existing between the two theories and the resulting implications at the micro- and macro-scale of linguistic convergence will be presented. Every instance of language change in general has its roots in a 'simple' convergence mechanism in an individual, which is, in fact, anything but 'simple'.

Pierrehumbert [Pie06], drew our attention to a fact that could solve the apparent impossibility of the mysterious *perfect imitation*. What she proposed as a model of the internal variation of an individual's capabilities in a given situation could also be interpreted as a model of inter-speaker variation in the accomplishment of the following processes, which are said to lie between the physical experience of a stimulus and memory storage [Pie06, 525]:

- attention
- recognition
- coding

These processing steps make it evident that what we store in our exemplar memory is not an imprint of our raw experience but the actual experience filtered and modulated by these intermediate stages. If exemplar *storage* proceeds over many levels, it is not far-fetched to assume that exemplar retrieval might also be accomplished as a series of consecutive actions rather than through a straightforward intention-action link. Furthermore, Pierrehumbert mentions the all too obvious physical limitations that impose certain restrictions on what we can sound like in a given moment, despite our willingness and skill to become perfect imitators [Pie06]. Apart from these physical limitations defined by the shape and size of our articulators, the recognition of a target, given that our attention is focused on the stimulus at that moment, could depend on physical constraints imposed by top-down processes as well. Evidence suggests that cortical structures can control the sensitivity of the olivocochlear bundle (OCB), a nerve bundle that transmits neural signals from the temporal lobe back to the cochlea. This feedback loop reacts to bottom-up signals and sends back messages that can tune the cochlea to respond to a certain range of frequency [Sty06]. This mechanism could be responsible for filtering or even rejecting unwanted stimuli at the time of listening[9].

2.2.1 Attention

The generally established division includes a *divided* and a *selective* type of attention, where the former indicates a state of paying attention to multiple stimuli

[9] for instance at a cocktail party, which would allow us to attend to one voice only, despite many distractors present.

simultaneously[10], and the latter to a state of focusing only on one source of information while ignoring others[11]. Current theories of attention also tend to see it as a complex construct rather than a single mechanism applicable to all cognitive processes [LV02, 261]:

> The term ‘attention’ applies to many separable processes, each of which operates within a different cognitive subsystem and in a manner that reflects the structure and processing demands of that cognitive subsystem.

This would, for instance, indicate a differently operating attentional mechanism for visual and auditory tasks, which is why Treisman’s widely accepted Feature Integration Theory (FIT)[12] for visual stimuli is said not to translate well into auditory perception that is enriched with one dimension not immanent to visual scene perception – a dynamic time pattern [Sty06, 132].

For this reason, feature integration in auditory perception has been suggested to come closer to the FIFA model[13]. While the finding that frequency is a more salient cue in attracting and maintaining attentional focus than location comes as no surprise, sometimes faster reactions to conjunctions of features than to individual ones in the auditory domain have left researchers puzzled[14] [Sty06]. Woods et al. [WA01] explained it in terms of the facilitatory interactive feature analysis, which Styles summarized as follows [Sty06, 132]:

> (...) the processing of individual features interacts, particularly in the case of auditory stimuli, such that when attention is focused on the more discriminable feature of frequency, this improves feature processing of other features at the same location.

[10] e.g., driving a car while talking on the phone.

[11] as shown, for instance, in dichotic listening tasks (concurrent listening to different stimuli presented in both ears) and the Stroop effect (task in which subjects need to name the color of the ink of printed words, e.g., the word ‘red’ printed in yellow ink; see [Sty06, Mat09]).

[12] Treisman and Gelade [TG80] suggested that attention can operate in a *divided* mode, where all parts of a scene are processed simultaneously, at a rather low level with parallel access; or in a *focused* mode, which requires serial processing for each consecutive item in a scene and thereafter identifying which features belong together and determining their exact location.

[13] facilitatory interactive feature analysis [WA01, HH83].

[14] since it runs contrary to findings in visual perception.

Woods and Alain conclude from their ERP study that auditory features most certainly undergo an exhaustive analysis which proceeds in parallel in two distinct dimensions - space and time. Clearly, whenever more target features were present, processing of all combined features was enhanced and recognition facilitated [WA01]. This would suggest easier recognition of a target the better (or fuller) specified it is in the signal, provided, naturally, that the features are correctly identified.

Attention also has a very close relationship to memory, as has for instance been shown in the Atkinson and Shiffrin model of memory [AS68]. According to them, working memory is responsible for both processing and storing incoming stimuli. As a consequence, more demanding processing takes away from the resources allocated to storage, which then cannot be equally effective. A similar conclusion has been reached by Lavie and colleagues [LHdFV04] who could demonstrate in their experiment on visual attention that a heavy load on working memory decreased the ability to ignore distractors, meaning that the filtering out of unwanted information became impaired. Styles [Sty06] conceives of this relationship not as simply happening between memory and general attention but as being specifically tied to *conscious* attentional control. Since Atkinson and Shiffrin's discovery, basically all memory models have taken into account attentional control as one of their crucial elements[15]. Gathercole and Baddeley [GB01], for example, place attention in the *central executive* component of their working memory model (see Figure 2.1 in Chapter 2.1.2). The central executive is the most important element in Baddeley's model and is responsible for allocating attention to the currently most relevant processes [GB01, 4]:

> Its functions include the regulation of the information flow within working memory, the retrieval of information from other memory systems such as long-term memory, and the processing and storage of information. The processing resources (...) are, however, limited in capacity.

[15] e.g. Broadbent's or Baddeley's models [Bro84, Bad86, GB01].

Therefore, the efficiency of the system in dealing with specific functions is dependent on the demands concurrently placed on the whole central executive.

The notion of *early* vs. *late* selection has drawn another division within models of attention. Broadbent's [Bro58, Bro84] *filter theory*, for instance, is an early selection model, since it foresees only strict serial processing, with the filter modelled as a structural *bottleneck* letting through only parts of the original stimuli. The filter also controls what becomes consciously known – all information that has been filtered out cannot enter consciousness. Deutsch and Deutsch [DD63], on the other hand, proposed a late-selection model, where the "decision stage" (concerning which message is worth our attention) is actually much later (only after full processing) than it is in Broadbent's model. Treisman [Tre69] entered the discussion of early vs. late selection models with a modified definition of the filter, moving away from the notion of an "all-or-nothing" mechanism. Instead, Treisman proposed a filter that reduced the strength of currently irrelevant or unwanted stimuli – decreasing thereby their salience - but did not block them out completely. Styles has described this mechanism at the level of lexical access as follows [Sty06, 27]:

> Different words have different thresholds depending on their salience and probability. If the attenuator has the effect of reducing the perceptual input from the unattended channels, then only when words are highly probable or salient, will their thresholds be sufficiently low for the small perceptual input to make the dictionary unit fire. Thus the attenuator can account quite neatly for breakthrough of the unattended at the same time as providing almost perfect selection most of the time.

In this model, the influence of the *attenuator* starts early, before memory access. Others, such as Norman [Nor68], have suggested that selective attention kicks in only after parallel access to semantics, which is preceded by extensive automatic and unconscious processing of both attended as well as unattended information. Attention is, in his opinion, also gradable. Selection, however, depends on assigned pertinence values, which are calculated from the input and its context *after* semantic access.

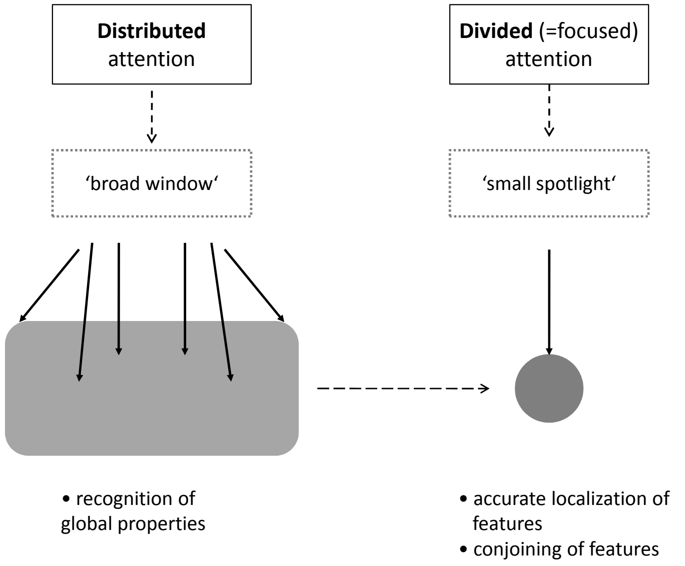

Distributed attention
Divided (=focused) attention
'broad window'
'small spotlight'
• recognition of global properties
• accurate localization of features
• conjoining of features

Figure 2.3: Illustration of the functions of Treisman's two types of attention. While distributed attention allows taking on a global perspective and noticing the *presence* of elements, divided attention is specialized in concentrating on one element only, allowing a thorough analysis. Distributed attention is said to monitor ongoing actions in an automatic way; divided attention requires conscious control, [Tre93, Tre99, Mat09].

In newer accounts, as in the previously mentioned Feature Integration Theory[16] for visual perception, Treisman specifies that features are characteristics allowing for a so-called 'pop-out'[17]. The features are organized hierarchically[18] and the time of selection could be either early or late. The two types of attention in FIT are called *distributed* vs. *divided* attention[19] (see Figure 2.3 for an illustration).

A very interesting point is made by the author with respect to the possible outcome of inattention [Tre99, 108]:

> (...) attention is needed to bind features together, and (...) without attention, the only information recorded is the presence of separate parts and properties.

[16] FIT, see e.g. Treisman 1993 or Treisman 1999 [Tre93, Tre99].
[17] i.e., those which are immediately noticed/spotted.
[18] with features being more or less important or salient.
[19] Treisman uses the term 'divided attention' to mean 'focused attention'.

As mentioned earlier, FIT is not the best theory when it comes to explaining all mechanisms of auditory perception. However, one of the characteristics, namely the possibility of a misplaced recombination of features and the creation of so called *illusory conjunctions*, may hold as well in some form for auditory perception. Hawkins [Haw10, 60] makes a very strong claim regarding the nature of “auditory objects” and argues that the processes underlying listening can in many ways be compared to seeing, thereby not excluding the creation of illusions:

> (...)when sensation meshes with expectations, listeners believe they perceive ‘real’ linguistic objects in spite of possibly severe variation and degradation in the acoustic signal. (...) perceived linguistic units, including distinctive features, are ephemeral (and illusory) ’auditory objects’, which are created by the listening brain using domain-general processes that underpin meaningful behaviour.

This train of thought brings us back to the advantages inherent in an exemplar-based model of language, namely the explanatory power it has for dealing with variation in the speech signal. Reflecting upon the “fuzzy” nature of category boundaries in speech, Hawkins still defends the existence of discrete units in speech, while suggesting that perception actually actively *creates* those units in an absolutely subjective and context-dependent way. Contributing elements in this process are the actual physical sensation (mediated by familiarity and expectations), the context of the stimulus, and – necessarily – attention as well [Haw10]. How exactly attentional mechanisms are involved in this process will be described in detail in Chapter 2.2.2, dealing with the recognition of exemplars in speech.

2.2.2 Recognition

In her work on illusory aspects of auditory perception, Hawkins [Haw10] refers to the Adaptive Resonance Theory (ART) [Gro03, Gro05]. Adaptive Resonance Theory models the link between the incoming stimuli, attentional mechanisms and the role

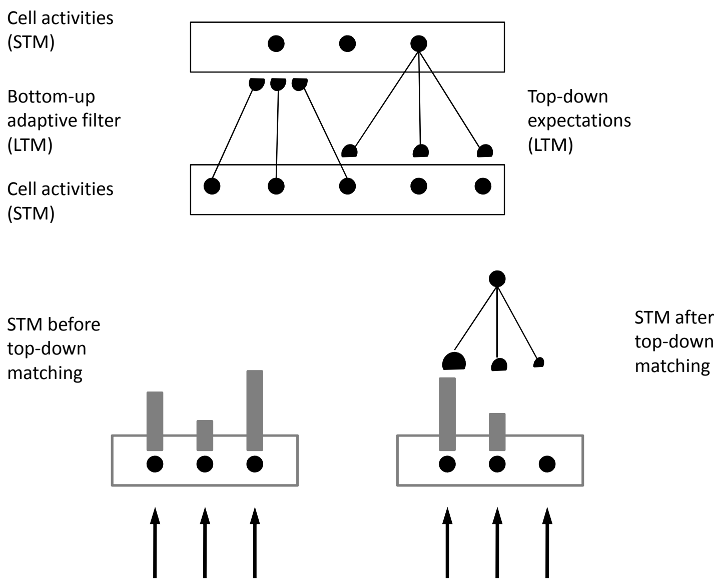

Cell activities (STM)
Bottom-up adaptive filter (LTM)
Cell activities (STM)
Top-down expectations (LTM)
STM before top-down matching
STM after top-down matching

Figure 2.4: Upper part: Grossberg supposes that lower-level activations (short-term memory – STM – in the model) send signals to a higher level. Long-term memory traces (LTM) – the adaptive weights – then reinforce these signals, modulating their activation at the higher level. From this level top-down activations are triggered and a matching procedure between the two processing levels begins. Lower part: the outcome of the matching procedure between top-down (LTM) activation and STM signals – the greater the size of the respective hemidisk, the stronger the learned memory trace (LTM) is within that pathway. Edited from Grossberg 2005 [Gro05, 653].

of expectations, and finally, previously aquired knowledge. Grossberg [Gro03] proposes that any kind of conscious auditory percepts (including speech) emerge from resonant states of the brain. The development of such a resonance is introduced by the interaction of bottom-up physical events with top-down expectations[20] learned prior to the present experience (see Figure 2.4 for an illustration).

The top-down information then either strengthens (reinforces) those bottom-up features which are consistent with the learned prototype, or it weakens (inhibits) those not compatible with the stored pattern. The interaction of the bottom-up sig-

[20] In Grossberg's terms also called *prototypes*.

nal and the top-down modulation creates an "attentional focus" for those features consistent with past experiences. Once a feature (cluster) achieves this top-down reinforcement, it restarts the resonance cycle [Gro03, 425]:

> (...) the selected cells (..) resonate with amplified and synchronized activities. Such a resonance binds the attended features together into a coherent brain state. Resonant states, rather than the activations that are due to bottom-up processing alone, are proposed to be the brain events that represent conscious behavior.

Grossberg [Gro03] holds on to the idea that expectation can modulate what a person perceives, as is evident in instances of phonemic restoration. Moreover, the described activations are suggested to be the basis for the brain's fast learning processes without losing or overwriting previously learned information [Gro05][21]. The essential properties of the Adaptive Resonance Model are subsumed as follows [Gro03, Gro05, Haw10]:

- Bottom-up signals trigger top-down expectations which are matched against the incoming data.
- Both the bottom-up and top-down pathway hold so-called 'adaptive weights' (long-term memory traces modifiable by experience).
- Bottom-up input can, given sufficient strength, activate cells by itself.
- Top-down modulation cannot cause a cell to fire by itself; it can only act as a prime or sensitizer and thereby change the reaction threshold of this cell.
- A match is achieved when a cell receives convergent signals from both pathways.
- A mismatch, meaning that a large bottom-up input is met with only low or no top-down feedback at all, can inhibit cell firing.

[21] for details on the learning process in ART, see Chapter 2.2.4.

Consequently, attention is seen as a means to achieve modulatory priming and matching through a mechanism called *top-down modulatory on-center off-surround network*[22]. Supporting evidence for the existence of such an on-center off-surround network and the aiding effect of attentional feedback has been provided by studies in neurophysiology[23], a.o. within the auditory cortex[24], and for speech perception and word recognition [GBC97, GM00].

One essential aspect of the debate which has not been addressed in this context so far has been brought up by Hawkins [Haw03][25]. She poses the question *what* it exactly is that we need to recognize in an incoming stream of speech. According to her line of argumentation, it is fine phonetic detail, since all other formal linguistic categories (such as phonemes or words) are usually superfluous as they are emergent and not always necessary for extracting meaning. It is therefore mainly meaning extraction which communicative partners are interested in, and meaning is present all over the signal in fine-grained phonetic detail [Haw03]. This is not to say that the message is analyzed in the absence of any linguistic structure, but that identification takes place *probabilistically*, as [Haw03, 391]:

> (...) each structural element is identified relative to others in the environment and to the listener's expectations derived from exemplar memories. The listener aims at meaning, not a complete linguistic description, so he or she will accept the most probable meaning as soon as the overall evidence matches the expected sound pattern well enough.

It seems vital here that the actual degree to which the formal identification of linguistic structure proceeds is assumed to be listener-specific and contingent on his or her experience. Moreover, Hawkins' *Polysp model* assumes that experienced listeners use contextual information to identify the patterns of the current signal

[22] Cells in the on-center receive positive feedback, while cells in the off-surround receive top-down inhibiting signals [Gro05].

[23] e.g., Luck and colleagues [LCHD97].

[24] in particular for feedback from the auditory cortex to the medial geniculate nucleus (MGN) and the inferior colliculus (IC) [ZSY97].

[25] compare also Chapter 2.1.1.

according to the style and accent currently being used rather than just drawing from a “canonical” pool for pronunciation [Haw03, 392]. Experience should thus change the way listeners process an auditory stream, and it makes no difference whether it concerns a foreign language, a new accent, or even just a novel speaker.

2.2.3 Automaticity and consciousness

Both Grossberg and Hawkins [GBC97, Gro03, Gro05, Haw10] addressed the issue of consciousness in perception in a very straightforward way, by saying that what the listener becomes consciously aware of is the result of the resonance between the lower and higher levels of processing, and not the initial physical stimulus itself. However, one cannot avoid the question of how much control over attentional mechanisms a person has and if control always means being conscious of the ongoing processes. Since automaticity and consciousness have been central to the debate between Interactive Alignment Theory and Communication Accommodation Theory – and the stands taken by both sides have been said to be mutually exclusive[26] – a closer look at how attentional control and consciousness interact is needed.

Think back on how it was to learn to drive a car: it took huge amounts of concentration and time to handle all the different actions simultaneously and still pay attention to what was going on on the road in front of us, remembering all the necessary traffic rules we learned beforehand. Certainly it cost us a lot of effort. But, with more and more practice, we paid less and less attention to our arm and leg movements, which seemed to get automatized. Stepping on the brake in case a sudden obstacle appears on the road, should in fact require *no* conscious thinking, only a reflex-like reaction. It appears, then, that the more practice we have had, the less conscious thinking was needed [Sty06]. However, does that mean that no conscious knowledge about how an action is performed indicates that a person has no *control* over it?

[26] see Chapter 1.3.

Most models of executive control agree that the more skilled a person is at a certain task, the less interference it should cause. This is true for all "strategic bottleneck models"[27], as e.g. Meyer and Kieras's executive process interactive control model (EPIC) [MK97]. Their computationally-oriented model expects two *skilled* tasks to be successfully handled concurrently, since neither draws on central processing capacities[28]. When an *unskilled* action needs to be performed, however, the central executive may assign priority values to one of the tasks and postpone some processing stages. The EPIC model is, in contrast to most other models, not based on limited general processing capacities but assumes that the limitations arise at the output level for carrying out actions[29] [MK97].

Norman and Shallice [NS86] propose another model of control in information processing: *"automatic"* control and *"controlled"* control. The two types of control are distinct in that the latter requires attention, while the former does not (see Table 2.1 for more details). Such a distinction between two types has become known as the *Two process theory of attention*. Such an approach has also been advocated by Posner and Snyder [PS75, 81], who drew a line between:

> Automatic activation processes which are solely the result of past learning and processes that are under current conscious control. Automatic activation processes are those which may occur without intention, without any conscious awareness and without interference with other mental activity.

The conscious processing system mentioned by Posner and Snyder is assumed to be of limited capacity. Thus, performing two or more simultaneous tasks takes away from overall resources. Furthermore, Norman and Shallice propose a detailed list of situations in which deliberate attentional control is needed.

[27] Models which allow the executive control to set the "bottleneck" either early or late, depending on what task is given priority [Sty06].

[28] i.e., they do not demand control processes.

[29] as e.g., speaking, listening, seeing, or moving limbs.

'Automatic' control	**'Controlled' control**
1. carried out without awareness 2. initiated without conscious intention 3. attention is automatically drawn to a stimulus 4. such actions should not cause visible interference or competition	1. is deliberate 2. is conscious 3. allows only a limited amount of data 4. actions can cause interference
–> does *not* require attention	–> requires attention

Table 2.1: Features of the two types of control in information processing, as proposed by Norman and Shallice 1986 [NS86].

Such tasks usually [NS86, 2]:

- have a planning or decision-making component
- require problem solving
- have not been learned properly or are new
- are classified as dangerous or difficult
- demand suppressing a strong habitual answer

A possible parallel to exemplar perception and re-usage could lie in the authors' final three suggestions. More deliberate attentional control could therefore be needed in cases where encountered exemplars are not sufficiently familiar or have been previously stored wrongly, thereby impeding their recognition. Tasks considered very difficult are also said to activate conscious mechanisms (or at least should do so) to allow for better control. On the other hand, the level of excitement – or in Kahneman's or Revelle's terms[30] the level of *arousal* accompanying tasks or situations labelled difficult – can also influence attentional mechanisms. Whereas a normal level of arousal combined with motivation for the task is assumed to increase performance, too high a level of excitement might cause an attentional breakdown. The last issue on Norman and Shallice's list could apply to a situation where, when

[30] [Kah73, Rev93].

in dialog with a nonnative speaker, a speaker needs to simply overcome speaking in a way he or she normally does. In fact, though, the more experienced a person is and the better she manages to direct the attentional resources to the relevant items and recognizes and stores them properly, the more exemplars from one's "own pool' of exemplars there are to draw from. In contrast, the more difficulties someone has in the exemplar acquisition process, the more attention is demanded on inhibiting habitual answers, in this case, choosing the most frequent own exemplars. Success in suppressing such a strong habit might be a factor determining how good an accommodator a person can be. Actions that theoretically proceed only stimulus-triggered and do not require awareness, can nevertheless be consciously controlled, if necessary.

Various experimental results suggest a parallel between the controlled/automatic and conscious/unconscious condition [CM85, Mar80]. Styles interprets unconscious processing not to be open to strategic manipulation, while conscious processing is [Sty06]. To begin with, one needs to find a proper definition of consciousness. However, as is often the case in highly debated areas of research, every single study presents a slightly different definition of a conscious state. In Farah's terms [Far94], consciousness can be considered as "a state of integration among distinct brain systems". According to some models, consciousness is even seen as an almost separate faculty that can e.g. be disconnected from perception and action (DICE[31] model by Schacter and colleagues [SMM88]).

However, the two process theory of attention mentioned above, which supposes a clear distinction between automatic and conscious control modes, has also faced criticism [Neu84]. Neuman argues that it is extremely difficult to prove that tasks running *automatically* do not demand attention. Practice also does not always prevent interference in tasks. For these reasons Neuman suggests that automatic processing is not totally uncontrolled, but is rather controlled *below* a conscious and aware level. He pleads for automaticity to arise only in cases in which the surrounding conditions, including both the processing mechanisms as well as the external

[31] = **d**issociated **i**nteractions and **c**onscious **e**xperience.

situation, favor it [Neu84, 282]:

> A process is automatic if its parameters are specified by a skill in conjunction with input information. If this is not possible, one or several attentional mechanisms for parameter specification must come into play. They are responsible for interference and give rise to conscious awareness.

This would suggest that accommodation processes can be controlled well below the level of conscious awareness, just as Neuman expects for access to long-term memory and the processes of forgetting [Neu84]. Styles [Sty06] shares Neuman's view on this and concludes that the majority of all information processing happens below a consciously aware level, since we are not able to reflect on them. This does not necessarily imply that it is not controlled, just as it happens, for instance, with lexical choice in language production. We are aware of the effects rather than of the intermediate processing steps.

Wegner [Weg03] approaches the conscious-unconscious issue from yet another angle. His concept of action control foresees two paths - one actual causal path grounded in an *un*conscious cause of action (the original trigger for the action) and a second unconscious path leading to a thought about the action which in the end gives rise to the *illusion* that we caused our action with our thought (see Figure 2.5 for an illustration).

He backs his ideas up with evidence from such phenomena as visual form agnosia or the alien hand syndrome, where action indeed seems to be separated from conscious control [Weg03, Sty06]. Following Wegner's line of thought, one could assume that even though we think we are consciously manipulating the outcome of our actions, their original source lies somewhere else entirely and the impulse for the action was in fact the trigger of our thinking about it, and not the other way around.

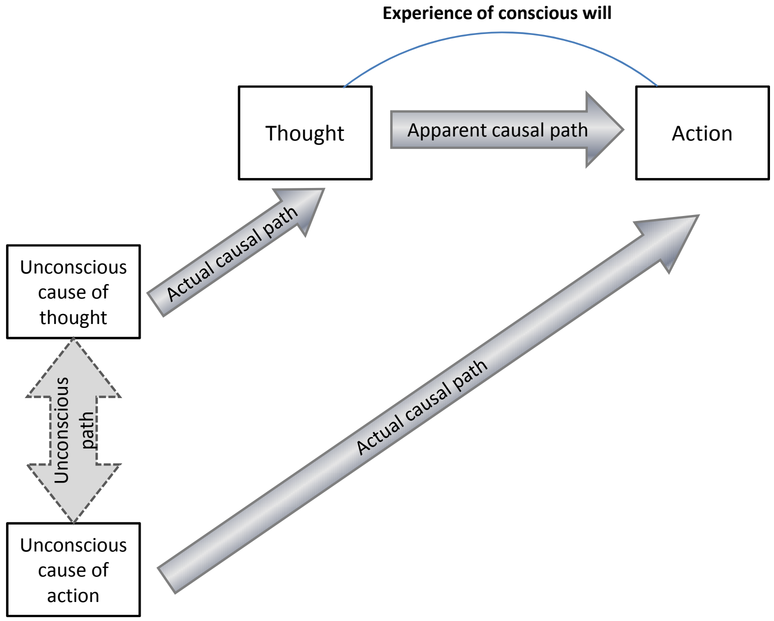


Figure 2.5: Wegner's model of two paths leading to the "illusion" of a direct connection between conscious thought and the cause of an action. Edited from Wegner (2003) [Weg03, 66].

2.2.4 Coding

Learning processes are, in mechanical terms, encoding processes. Grossberg's ART model[32] provides an interesting description of how we can link attention, encoding and learning, as it is understood in terms of storage in long-term memory. Grossberg [Gro05] proposes a solution to the problem of overwriting of old information by suggesting that neural representations can only be modified by incoming stimuli which match them to a sufficient degree. Only in the case of a close enough match can resonance be achieved, and thereby, learning. What the ART model foresees is a *fine-tuning* of existing representations, and no overwriting by outliers. A crucial process here is the so-called *vigilance control*, which can modulate the level at which resonance occurs, being concurrently the level at which learning happens [Gro05, 659]:

[32] for an introduction see Chapter 2.2.2.

> (...) a learning individual can flexibly vary the criterion of how good a match is needed between bottom-up and top-down information in order for presently active recognition categories and their top-down expectations be refined through learning.

Two general levels of vigilance control have been suggested in the ART model [Gro05]:

- coarse matches – attention is focused on general and abstract information
- fine matches – attention can even be focused on individual exemplars, therefore learning is more specific and concrete.

In the case that no sufficient match is achieved, any neural activity of the currently active top-down exemplar[33] will be inhibited and the search will proceed to another exemplar until a match is found – or a totally new exemplar is learned. The described memory search is mediated by corticohippocampal interactions [Gro05]. Hawkins [Haw10, 76] further specifies that it is the ability to narrow down the focus of attention (equal to achieving a "fine match") which forms the basis of perceptual learning, making it "(...) adaptive for speech perception, allowing plasticity and general short- and long-term adaptability, for example to unfamiliar accents".

And so this highest level of vigilance is thought to be the one responsible for exemplar learning, since it allows for acute discrimination and access to details. The model includes vigilance as an individually modifiable component and Hawkins assumes that listeners may consciously[34] direct attention to fine phonetic detail that, while it is not necessary for understanding, is still present in the signal. A *skilled perceiver*, thus, should, according to Hawkins, be able to perform fast and appropriate attentional shifts between different levels of linguistic information[35]. It seems, however, possible that some people cannot be classified as such gifted perceivers

[33] Grossberg uses the term *prototype* but mixes it frequently with the terms *exemplar* and *representation*; hence, to avoid inconsistencies, the term *exemplar* will be used throughout.

[34] or, along the line of thought presented in Chapter 2.2.3, "controlled" but not necessarily "consciously".

[35] for more details on Hawkins' view of exemplar processing, see Chapter 2.2.2.

and might encounter problems with vigilance control, which hinders resonance and learning of exemplars at a fine-grained acoustic level. Performing appropriate shifts, defined by which cues or *anchor points*[36] need to be paid attention to in non-native speech, also requires some reorienting [PGA10] and may not be equally manageable for all individuals. This *orienting* is, in turn, not only essential for facilitating detection of relevant information but also for the processing in short-term memory and subsequent storage in long-term memory [PGA10].

2.2.5 Retrieval

For correct[37] retrieval of phonetic exemplars from memory, it is vital that all preceding steps – allocating attention, recognition, and coding – be performed effectively. Accessing long-term memory for exemplar retrieval also requires attention, e.g. in the form of the executive control in Baddeley's model [Bad86, Bad03]. Executive control directs attention to the relevant long-term memory traces, triggered by its current sensory input, and holds the information in the episodic buffer. The two possible ways in which actions (and therefore also speaking) can be carried out are described as follows [GB01, 6, emphasis in the original]:

> Well-learned or "automatic" activities are guided by schemas that are triggered by environmental cues.(...) However, when novel activities are involved, or when the environment presents an urgent or threatening alternative stimulus, the higher-level Supervisory Attentional System (SAS) intervenes to control action.

According to Gathercole and Baddeley, the SAS model developed by Shallice [NS86] may correspond to one of the many functions the central executive bears but cannot be uniquely identified with it. Nevertheless, executive control seems to

[36] see Hawkins 2010 [Haw10].

[37] i.e. appropriate, in terms of the communicative situation, the language, the speaking partner, the topic, etc. See also Chapter 1.3.1.

be involved in many mechanisms apart from action selection and control, and in all probability is an independently operating mechanism[38] [GB01].

Norman and Shallice's model of deliberate vs. no-awareness situations presented in Chapter 2.2.3 also concerns behavior control [NS86]. As already mentioned, they assume the existence of *schemata* in long-term memory that can be activated by environmental stimuli, which is the bottom-up pathway. Any conflicts between active schemata are routinely managed by the "contention scheduling" mechanism. However, a second pathway might take over control, namely the top-down thread, which acts according to current higher goals. This top-down modulation by the SAS is achieved by applying more excitation or inhibition directly to the schemata[39] [Sty06].

Another problem in exemplar retrieval is addressed by Norman and Bobrow [NB75], who developed a model for resource limitations that can surface in degraded performance. Norman and Bobrow [NB75, 45] consider all of the following *resources*:

- processing effort
- different forms of memory capacity
- communication channels

The most important distinction in their model concerns the notions *data-limited* vs. *resource-limited* performance. With a single task at hand, *data-limitation* can include low-quality input, as for instance in noise-degraded speech. It can also be located in memory when the entry being currently searched for is not available or is not fully available. Data-limitation cannot, by its very nature, be overridden. When, however, performance changes after assigning more resources, performance is then said to be *resource-limited*.

[38] e.g., task coordination, planning or conscious awareness.

[39] an example of such a probable intervention of the SAS is the inhibition of a word-reading response replaced by a color-naming response in the Stroop task.

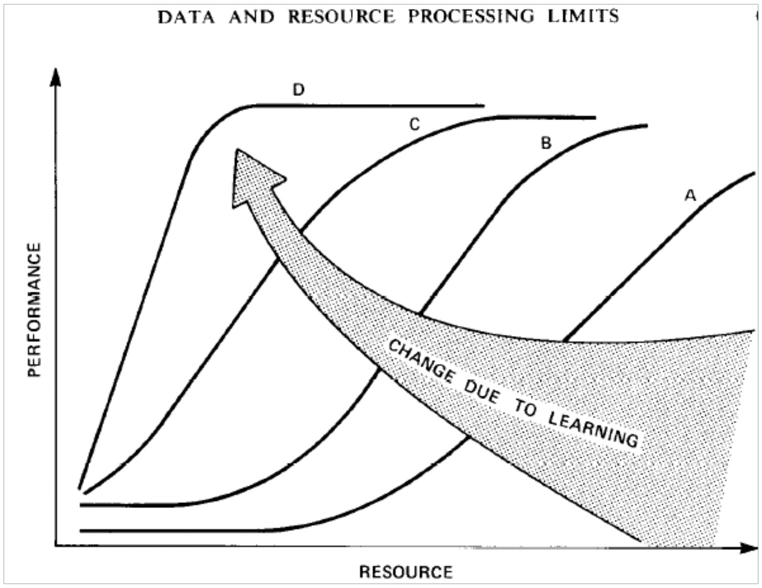


Figure 2.6: Illustration of changes in the performance-resource functions induced by learning. Performance with the same invested resources rises from curve A to curve D. Norman & Bobrow (1975) [NB75, 61].

It is expected that all processes are, up to some point, resource-limited and data-limited. In cases where more tasks have to be combined, the following scenarios are possible:

1. two (or more) tasks share resources: an increase in resource allocation to one task increases performance for task A, but simultaneously decreases performance for task B.

2. if two tasks do not seem to be affected by this complementary relationship, they either do not share resources *or* are data-limited in nature.

Norman and Bobrow have also illustrated the way learning can influence resource-dependent performance (see Figure 2.6). In a re-analysis of LaBerge's experiment [LaB73], they concluded that

> when only a single, expected task is tested, then both well learned and newly learned processes will be in the data-limited portions of their operations: hence, both will appear to give equal performance. Under conditions of dis-

> traction, however, the newly learned process can be driven to the resource-limited region, whereas the well learned process will often stay within the data-limited region. Presumably, severe attentional distraction will force even the well learned process towards the resource-limited portion of its operation.

The described benefits from learning processes again point to the fact that adaptation to a wide range of features is possible and even *expected* in normal everyday communication. Hawkins [Haw03] assumes that the reason for adaptation to all kinds of new situations, accents and individuals throughout life – and therefore the reason for such a change in the distribution of exemplars – is a corresponding change in input. Thus, a considerable amount of (fruitful) experience with a specific variant of speech not only enhances our recognition of relevant patterns leading to faster meaning access, but also allows us to retrieve those exemplars appropriate to the current situation – i.e. non-canonical forms but situationally-colored ones – at a much higher speed and much more efficiently. This is also consistent with Pierrehumbert's view of speech adaptation in and between bigger communities [Pie06].

2.2.6 Talent and exemplar processing

The process of exemplar acquisition is anything but straightforward. Ample evidence in support was presented in Chapter 2.2. Pierrehumbert's [Pie06] intermediate stages of noticing, recognition and coding have served as a basis to describe the multiple mechanisms standing in between the mere physical experience of a stimulus and its subsequent re-usage in production. It has also become evident that one mechanism seems to be especially powerful when it comes to the explanation of possible individual differences in exemplar processing and phonetic convergence – namely attention. Hawkins, moreover, establishes the ties between formal linguistic thinking and a more ID-oriented version of language competence when she says [Haw03, 389]: "Because categories are self-organizing and emergent, each individual develops somewhat different mental representations of language".

It follows that every individual is faced with a slightly different language "corpus" and consequently also stores a slightly different picture of language. Holding on to that thought of varying input, and bearing in mind the previously described processing stages and their cognitive components, we might speculate that each individual is also somehow differently endowed when it comes to the deployment of these processing mechanisms. Not every person seems to have equal control over the attentional and working memory mechanisms [CES^{+}05, Rob03, Sty06] that could surface in processing difficulties of fine acoustic detail necessary for acquiring a native-like pronunciation.

In Chapter 2.1.1 it was hinted at that, according to some usage-based models, no full linguistic analysis of an incoming stimulus is usually necessary for correct meaning extraction [Haw03, Haw10]. However, such a full analysis focusing on acoustic detail is possible and could, according to the ART model, be controlled via a fine-tuning of the vigilance level in every individual and this should be situation-dependent. The following description is one of many possible variants of how and where such an individual difference might come into play. Nevertheless, its purpose is to point to places in the processing chain where aptitude might cause crucial differences (see Figures 2.7 and 2.8).

Figure 2.7 presents an illustration of a standard attention allocation following the general lines of Hawkins's and Grossberg's suggestions, but without going into the details of the resonance mechanisms described in the ART model [Gro03, Haw03]. Here, further access to fine-grained acoustic information proceeds without difficulty. Figure 2.8, on the other hand, displays a situation where attention allocation is hindered at the moment of accessing fine phonetic detail. Once a coarse pattern matching is accomplished, and meaning access becomes possible, attention might be immediately redirected to the following incoming sequence of stimuli[40]. This withdrawal of attentional resources could hamper proper storage of

[40] pattern matching, as described by Hawkins [Haw03, Haw10], is assumed to happen for longer stretches of speech, and is only very rarely broken down to the word or phoneme level if not necessary/appropriate.

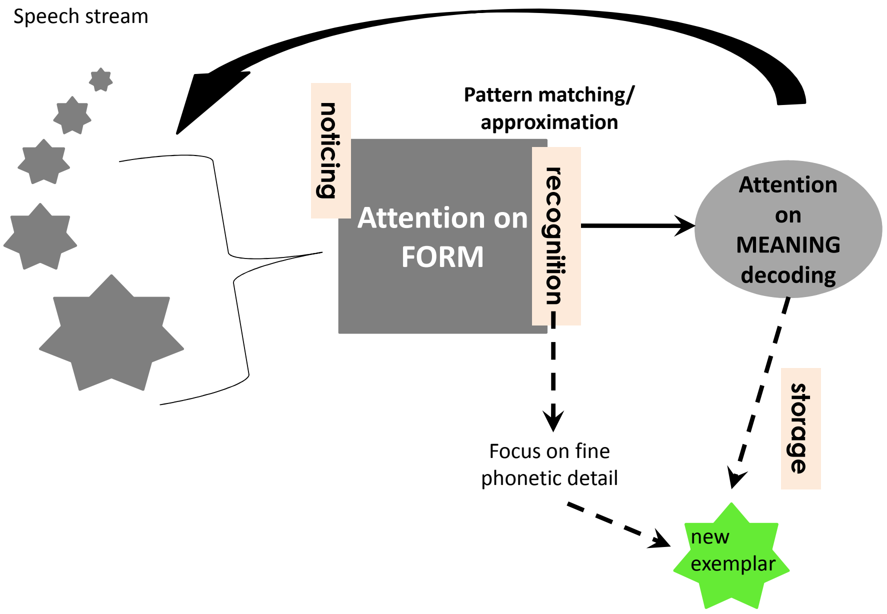


Figure 2.7: Illustration of how attention could be directed in speech perception in skilled individuals who are able to narrow down focus to fine phonetic detail and can subsequently store detailed acoustic exemplars of currently perceived speech, including e.g. dialectal or accent characteristics.

situation- and speaker-specific exemplars for later re-usage[41].

What has been laid out so far concerns a supposed *early* redirection of attention, happening before exemplar storage. In this case, the problem a less talented speaker would face is simply the lack of situationally-appropriate exemplars to choose from. As a result, such a conversational partner naturally would not have the right exemplar pool to choose from in order to converge at an acoustic level. However, a second situation seems conceivable as well: there could be a rather *late* attention redirection located at the retrieval level of exemplars for speech production. After the conceptual phase and meaning access, the search for both a suitable linguistic and phonetic form begins. This search is normally said to be guided by the resting

[41] though probably not totally block it, since one could assume that a "stripped down" version of the exemplar containing only coarse acoustic information is stored instead.

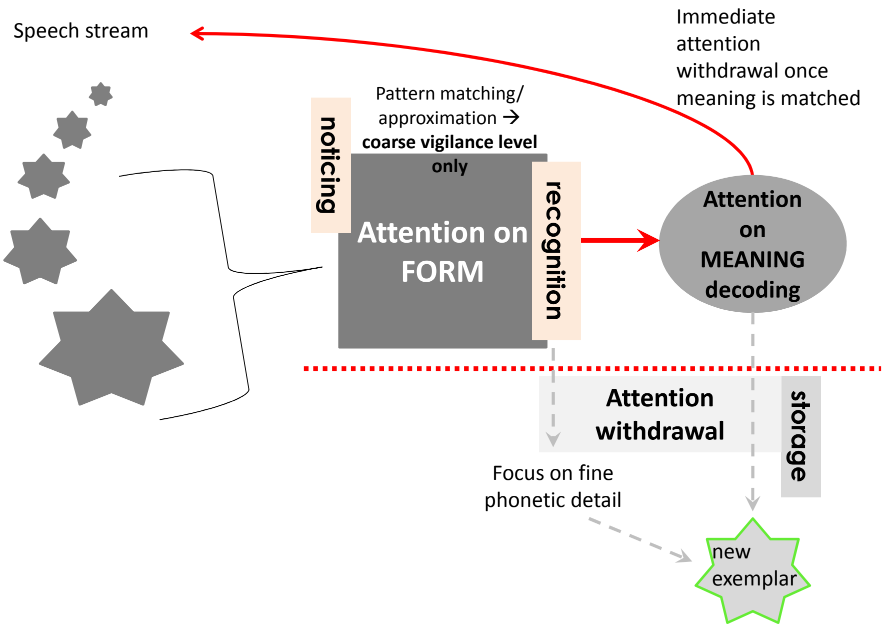

Speech stream
Immediate attention withdrawal once meaning is matched
Pattern matching/ approximation →
coarse vigilance level only
noticing
Attention on FORM
recognition
Attention on MEANING decoding
Attention withdrawal
storage
Focus on fine phonetic detail
new exemplar

Figure 2.8: Illustration of how attention could be redirected in speech perception in less skilled individuals who are only able to operate at a coarse vigilance level, withdrawing attention from the stage of fine phonetic detail analysis once pattern matching allows meaning access. Such a withdrawal inhibits proper storage of an acoustically fully specified exemplar.

activation levels of exemplars (latest and frequent items have higher activation levels; see Chapter 2.1.2 and 2.1.3). Even if the storage part has been accomplished successfully and there are many exemplars with a fairly detailed indexing, a redirection of attentional resources at this stage of retrieval would hinder the chance of finding a good match, and could, for instance, stop at the first match fitting the meaning requirements. Consequently, one's own most frequently used exemplar or just any other exemplar fulfilling the rather coarse phonetic form requirements could be chosen for production.

The presented suggestions thus foresee that less talented individuals could experience problems either early or late in the convergence loop, or even a mixture of both. The redirection of attention in these cases is by no means meant to be an all-

or-nothing mechanism. A gradable version seems to be in the realm of possibility. Further processing shortcomings located at the working memory level in perception and production are, of course, possible as well[42] [GB01, Obe02, RAC10].

2.2.7 Outlook on exemplar models and SLA

One central question about second language learning mechanisms remains still unanswered in exemplar-based models. Although some attempts have been made to clarify whether pure exemplar models suffice or whether a necessary abstraction process needs to be incorporated [Gol07, McL07, NO03], the latter models, though more advanced, still place any abstraction mechanisms within the declarative component of memory. Ullman's work [Ull01b, Ull01a, Ull04], on the other hand, provides yet another perspective to deal with the dual nature of the learning process and the biological makeup of our memory system. A closer look into current exemplar models evokes the impression that they might have so far looked only at the declarative side of learning, thereby overlooking another crucial component, *procedural memory*.

Based on McClelland and colleagues' complementary learning systems (CLS) approach [MMO95, NO03], Goldinger presents an extended model of exemplar theory [Gol07, 49] in which "detailed episodic traces and holographic, abstract traces combine to create behavior in real-time, allowing perceptual or memorial data to appear more or less "episodic"(...).

The complementary learning systems approach foresees dual-processing in recognition memory, with two neural circuits involved: the medial temporal lobe cortex (MTLC) and the hippocampus [MMO95]. The latter is responsible for the fast memorization of detailed events, while the former circuit is involved in the much slower processing of statistical regularities in the data [NO03]. In the CLS model, the rapid processing of specific items in great detail in the hippocampus perfoms the basis for specific "recall" decisions, while the data in the MTLC allows

[42] see Chapter 2.1.2.

"familiarity" judgments. The detailed procedure put forward in the model is described by Norman and O'Reilly as follows [NO03, 613, emphasis in the original]:

> The hippocampus assigns distinct (*pattern-separated*) representations to stimuli, thereby allowing it to learn rapidly without suffering catastrophic interference. In contrast, neocortex assigns similar representations to similar stimuli; use of overlapping representations allows neocortex to represent the shared structure of events and therefore makes it possible for neocortex to generalize to novel stimuli as a function of their similarity to previously encountered stimuli.

The Hippocampal Network stores events rich in details, which first need to pass through cortical structures (see Figure 2.9). Goldinger points to the fact that every bit of information reaching the hippocampus is not in its "raw" form but instead already comes in some degree of abstraction, mediated through the cortex. Furthermore, representations – or *abstractions* – in the Cortical Network naturally build up over time and become more stable as more similar events are projected back from the hippocampus [Gol07].

Through these reciprocal connections between the hippocampus and cortex, recent experiences can affect the early perception of incoming stimuli as they reach the cortex. As Goldinger notes, selective attention is also operating at this level, since prior traces can be used to improve performance. The creation of representations in the internal loop from cortical to hippocampal structures is said to unite long-term memory with real-time perception [Gol07]. Goldinger illustrates the practical meaning of his approach with an example of a perceptual learning situation: a listener quickly adapts to any encountered non-standard pronunciation of a specific token[43] in all other words that are affected as well. At the same time this perceptual adaptation holds only for this concrete token and for this unique speaker. It does not negatively affect perception in general [Gol07, 54], as "the ab-

[43] e.g., a mispronounced /s/ due to lisping.

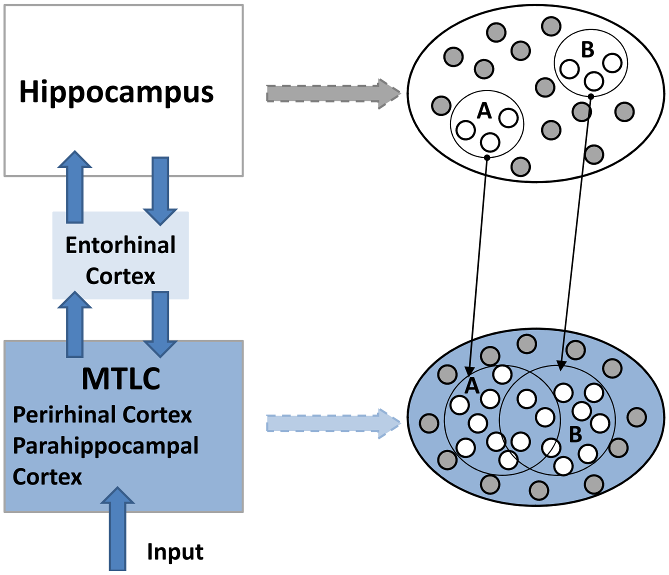


Figure 2.9: Complementary learning system (CLS). General structure presented on the left, feature separation in the hippocampus vs. feature overlap in the medial temporal lobe cortex on the right side. Adapted from Goldinger [Gol07, 51] and Norman & O'Reilly [NO03, 612].

stract lexicon is required to interpret an odd segment; episodic memory is required to both generalize and delimit the effect".

McLennan challenges some of Goldinger's claims by addressing more specific questions as to the coexistence of both abstract and episodic representations in the model [McL07]. He suggests that it should be clarified which type of representation is the dominant or default type and whether the approach can account for more fine-grained timing effects in processing speech[44]. His findings on the processing of foreign-accented speech suggest not only that it proceeds more slowly, but that it also seems to be accompanied by detectable talker-specific effects, thus evidencing the usage of episodic representations. McLennan also points to *population* differences, in that the reliance on both types of representations might also differ in L1 versus L2 learning. The assumption that episodic features are of more central value

[44] It has been suggested that talker and rate-independent factors influence processing earlier than talker and rate-specific cues [ML05].

in L2 perception than in L1 perception has been entertained as a potential solution [McL07, Ull01a]. However, these insights lead to the concession that the exact share both representations have in the processing of the L2 depends upon many exogenous[45] and endogenous factors[46]. This brings us back to the issue of individual differences.

Ullman's model of memory in language encompasses more than just the declarative memory circuit, in contrast to the CLS [Ull01b, Ull01a, Ull04]. While Norman and O'Reilly, and McClelland and colleagues [NO03, MMO95], do not refer to procedural memory at all in their CLS model, Ullman argues that both networks – the declarative and procedural memory – are vital for L2 acquisition. Mechanisms of abstracting information from episodic traces, which are part of the declarative network by Goldinger [Gol07], are sourced out to Ullman's procedural component of memory [Ull01b]. As will be shown, the claims of the declarative-procedural model (henceforth *DP model*), though challenging, are in principle not running against exemplar models but could form a valuable extension of usage-based theories of speech.

According to Ullman, grammar and lexicon are tied to two different memory systems or circuits: the *procedural* (PM) and the *declarative* memory (DM) system [Ull01b]. The two memory systems and their crucial characteristics are given in Table 2.2.

In general terms, Ullman [Ull01b] proposes a functional distinction between *declarative memory*, which is linked to the mental lexicon, and *procedural memory*, which is responsible for the acquisition and usage of grammar and rules. Declarative memory stores all words with a unique[47] phonological form and a unique meaning, as well as unpredictable word forms[48] and all other information categorized as distinctive, as comprised in affixes or idiomatic phrases. Procedural memory, on the other hand, is involved in the learning of the new and in the control-

[45] e.g., the degree of similarity between L1 and L2.

[46] e.g., the proficiency of the speaker.

[47] i.e., an *underivable* form.

[48] as e.g., irregular past tense verb forms.

Declarative memory	**Procedural memory**
tied to the mental lexicon	tied to grammar/rules
the mental lexicon stores unique words/expressions	learns new/controls established sensori-motor and cognitive habits, skills and procedures
explicit	implicit
fast, real-time learning	gradual learning
information stored not informationally-encapsulated	probably informationally-encapsulated

Table 2.2: Basic features of the DP model [Ull01b, Ull04].

ling[49] of already established sensori-motor and cognitive habits, skills and procedures [Ull04]. Its primary purpose, therefore, is the *combination* of lexical forms, phonological representations, and syntax and morphology[50] into complex forms [Ull01a, Ull04].

The basic tenets of Ullman's model [Ull04] allow us to draw a parallel between his declarative system and the system in which exemplars are stored. It seems straightforward to assume that the described mental lexicon is the place where the *unique* exemplars are kept and where access to them is granted for perception and production. What exemplar-based models have not yet discussed is the other side of the coin – if and how we acquire and store the rules necessary to combine our material into bigger chunks, and, probably more importantly, how exactly we retrieve the context-bound exemplar we need in a given moment. These processes have not yet been given a ground for discussion within usage-based models even though they seem to be essential where individual endowment might come into play. Ullman describes procedural memory as working as follows [Ull04, 237]:

> Functionally, the system may be characterized as subserving aspects of the learning and processing of context-dependent stimulus-response rule-like re-

[49] i.e., both the representation and the usage.

[50] e.g., all regular plural or past tense endings in English.

> lations (...). The system seems to be especially important for learning and processing these relations in the context of real-time sequences – whether the sequences are serial or abstract, or sensori-motor or cognitive (...).

The fact that real-time processing and rule-like relations seem to be two of the crucial functions of PM might allow speculations about their contribution to exemplar retrieval. Ullman and Pierpont [UP05] argue that Specific Language Impairment (SLI) in some patients might be caused by deficits in procedural memory brain structures[51]. One of the observed symptoms in many SLI patients is word retrieval difficulties [UP05], pointing again to a possible involvement of procedural memory structures in real-time exemplar access. Another argument in favor of the dual nature of the memory circuits involved in language has been presented by Ullman who points to the multiple interfaces of declarative and procedural memory circuits, e.g. in selecting information from declarative memory. Basically, this means that "(...) brain structures which underlie procedural memory also perform *context-dependent selection* and maintenance (in working memory) of knowledge stored in declarative memory" [Ull04, 243, emphasis N.L.].

A closer look into the working principles of procedural and declarative memory can thus provide further hints as to how exemplar models could incorporate it in its core assumptions. The two types of memory, for example, also differ in terms of their accessibility – DM allows explicit, conscious access, while PM is said to be implicit and not overtly accessible or consciously controllable. In terms of learning speed, DM favors a fast real-time acquisition, and PM only a gradual learning process, although the knowledge acquired might be the same or analogous. The information present in the declarative system can be shared amongst other mental systems; the rules within procedural memory, however, remain unaccessible and not manipulable by other mental systems (i.e. it is informationally-encapsulated) [Ull04].

The search for an entry (or the search for an exemplar) according to the DP model begins by looking up stored (ready-made) items in declarative memory, a

[51] A theory they call *Procedural Deficit Hypothesis* (PDH).

match automatically blocking further computation in procedural memory. Only if no match is found is a complex form computed using existing rules. However, even complex forms need not be computed from subunits each time but can instead be stored in declarative memory[52]. This depends on two factors: word/unit *frequency* AND *individual memory and learning ability* [Ull01b, 720]:

> (...)the successful computation of a form by the procedural system should inhibit the memorization of that form in declarative memory, therefore decreasing the likelihood of memorizing regular forms. However, any regular form can, in principle, be memorized. The likelihood of memorization should increase with factors such as the *frequency* with which the item is encountered or *individual variation* in learning abilities of the declarative memory system.

Rules in procedural memory can be derived from existing entries in declarative memory by an *associative memory* component which can, e.g., infer possible regularities within irregular listings as bring-brought-brought –>buy-bought-bought. Ullman argues also that frequency effects appear only for irregular (idiosyncratic) stored items and not for regular rule-computed items (e.g. regular -ed past tense vs. irregular forms). Bearing in mind the possibility that storage can occur also for regular forms, as seen above, this distinction does not seem to be categorical.

Individual variation becomes evident when focusing on the complementary relationship the two memory systems share. Higher usage or dependence on one system can inhibit the proper functioning of the other. Clinical evidence further suggests that a break-down of one system can also trigger a functional enhancement of the other [Ull04, UP05]. After discussing compelling neuroimaging evidence, Ullman draws the following conclusions [Ull04, 244]:

> This suggests that individuals vary with respect to their relative dependence on the two systems. Moreover, this relationship changed over the course of learning. These experiments (...) strengthen the view that early in learning declarative memory can play a particularly important role compared to procedural

[52] A similar dual-route ("multi-level") exemplar model has been proposed by Walsh et al. [WMWS10].

> learning, and that over time this balance shifts to the opposite direction. Thus, with increased dependence on procedural memory for a given function, there may be a decreased dependence on declarative memory, even if that system played a role initially in the same function.

In summary, Ullman supposes that the changing dependence on the two memory systems during learning holds also for the learning of *languages*, which could be a possible starting point for answering many open questions in SLA[53]. This view fits McLennan's approach to changing dependency on either type of representation (abstract or episodic) in first or second languages in an CLS-based extension of exemplar theory [McL07]. Furthermore, both the DP and the CLS model allow for individual variation in the usage of both memory/processing types[54], which could in theory alter the way people go about retrieving ready-made exemplars with rich details, finding rather abstract matches derived from detailed indexing in a cortical network, or even computing items on-line with the help of learned rules. Adding an individual's memory capacity and memorizing ability to this (which is supposed to influence the amount of information present in declarative memory), a picture emerges where there is much room for individual differences, including both pronunciation and convergence mechanisms. Obtaining a more precise account of how the assumed presence of two different underlying memory structures and/or processing types and their functional dissociations and interdependencies could be incorporated into exemplar theory, as well as shedding more light on the individual differences resulting from there, could form important future tenets of advanced usage-based models.

[53] as, e.g., the origin of differences between first and second language acquisition, also in terms of pronunciation acquisition.

[54] Note that the CLS speaks of changes within two types of processing *within* declarative memory, while the DP model allows shifts *between* the two types of memory.

Chapter 3

Measuring phonetic convergence - state of the art

Phonetic convergence has been studied under a multitude of aspects, all of which assume that some concrete front-end analysis[1] is necessary. Under such an approach, the input a listener receives is broken down into a set of features which could all show convergence when re-used in production. It is within these features that convergence should be considered (be it VOT, formant values or f0). Assuming such an analysis occurs, there is still one problem left to tackle: which features are the ones the listener relies on most? Is this choice universal for all listeners or is there variation? Previous studies, as will be presented in more detail in this chapter, have assumed either the existence of such a concrete feature (or a feature set) underlying change within convergence or they relied on impressionistic perceptual judgments without making such an assumption. Since exemplar-based models do not assume any automatic feature-extraction before storing episodes, it is unnecessary to pin down convergence to any single feature. Chapter 4 on the methodology and data analysis of this study will introduce a more holistic approach for measuring phonetic convergence.

Convergence studies can be generally divided into two groups: the first group of experiments focused on measurements of single features, amongst which were voice onset time (VOT), various pausing characteristics (pause length, duration and frequency), general speaking rate, formant values and f0 curves. The second group tried to capture convergence by using perceptual judgments or computer simulations of possible behavior.

3.1 Parameters

Starting in the early 1960s, even before the rise of a coherent theory of accommodation, an interest in phonetic convergence effects in dyadic encounters arose. The studies of Matarazzo and colleagues [MWSW63, MW67] were the precursors

[1] In such cases, it is presupposed that the tested phenomenon will be observable for a certain number of individual features, therefore only those features of the signal are analyzed.

of convergence studies and, as most of the other *early studies* did, they dealt with prosodic features of speech[2].

Matarazzo and colleagues [MWSW63] measured utterance duration in their 1963 study, just as Cappella and Planalp did in 1981 [CP81], though the latter used finer-grained measurements[3]. The subsequent experiment of Matarazzo's team [MW67] focused on response latencies, which were also the dependent measure in Street's interview study [Str84]. Further measures belonging to the broader prosodic domain have been pause duration, turn taking, speech amplitude [Nat75b, Nat75a], speech rate and various f0 measurements[4] [FTD$^+$89], investigated in the studies mentioned above and also in the work of Lieberman and Street [Lie67, SSVK83, Str84].

Although Gregory and colleagues turned towards spectral parameters in their measurements, as given by the calculation of long term average spectra (LTAS) [Gre83, Gre86, Gre90, GW96, GG02], their original region of interest for convergence within higher frequency ranges [Gre83, Gre86] shifted towards focusing on frequencies below 500 Hz in their later work [Gre90, GW96]. A frequency band below 500 Hz though, is very likely to carry f0 information, once more capturing convergence in intonation patterns, not the spectral properties of the signal.

Newer studies, such as Nielsen's and Babel's work [Nie07, Nie08, Bab09], have opted to track phonetic convergence with more fine-grained measures in the spectral domain, for example voice onset time (VOT) and the amount of voicing in vowels [Nie07, Nie08], and vowel formant values [Bab09]. Smith [Smi07], apart from looking mostly at convergence in prosody, also studied final vowel devoicing and the addition of schwas in her data. Delvaux and Soquet [DS07] concentrated on a more detailed picture of vowel properties, i.e. vowel duration, MFCCs[5] and the first three vowel formants (F1, F2 and F3). Their study, however, was an imitation study and not a classical example of convergence in dialog. We will be returning to

[2] as captured by pitch, amplitude and speech tempo.

[3] time series regression on duration-related parameters.

[4] e.g., mean f0, max f0, min f0 and f0 range.

[5] Mel frequency cepstral coefficients.

this in more detail in Chapter 3.2. Some recent studies on phonetic convergence, e.g. Kim et al. [KHB11], Pardo [Par06] or Namy and colleagues [NNS02], instead of measuring any discrete acoustic features, have relied on perceptual judgment tests using the AXB paradigm or perceptual judgments of accents (Willemyns et al. [WGCP97]). Yet another way of investigating convergence has been the usage of computer simulations, as in the study of Wedel & van Volkinburg [WvVed], in which the convergent behavior of two groups in contact was simulated. The groups were characterized by two sets of features – a distinctive and non-distinctive one – and, in the course of the simulation, the non-distinctive features exhibited convergence while the distinctive set diverged.

3.2 Setting

In studies on adaptation phenomena various elicitation techniques and paradigms have been used to obtain the data. A considerable number of experiments relied upon the repetition of words or longer fragments of speech [ACGSY11, Bab09, Bla49, BMH10, DS07, NNS02, Nie07, Nie08], based for instance on (a modified version of) Goldinger's shadowing paradigm[6] [Gol98] or on word games, as e.g. the dominoes game used by Bailly & Lelong [BL10]. As Natale [Nat75a] pointed out in his 1975 paper on convergence of vocal intensity, measuring convergence in a word or sentence repetition paradigm is not like measuring behavior during natural conversations. He explicitly refers to Black's study [Bla49] on vocal intensity, the results of which he considers not generalizable to standard interactive communication in dialog due to numerous flaws in the experimental design[7] that were non-representative of a normal dialog. As mentioned in Chapter 2.1.3, the imitation effect is not the same as convergence, and shadowing-based paradigms lead more to the former type of effect. The benefits of repetition or shadowing designs are clearly the perfect control over the data used for perception and production,

[6] see also Chapter 2.1.3.

[7] e.g., using headphones which eliminate normal sidetone and the presence of a limited amount of items the subjects had to react to – only five-syllable utterances.

and therefore it is easier to make comparisons when single phonetic features are measured within the data set. The generalizability of the results to a fully natural conversational context and their attribution to naturally appearing convergence mechanisms, however, remains unclear.

Another type of data used for measuring convergence comes from *interviews*, in either a laboratory [WGCP97, Nat75b, Nat75a] or a quasi-natural setting, as e.g. interviews conducted by Larry King in his talk show [GW96]. Willemyns and colleagues' study investigated the adaptation to different accents in a job interview situation. The participants/applicants were informed that their interview had a two-fold purpose and that they were participating in a linguistic experiment separate from applying for a real job [WGCP97]. Therefore, the situation could be considered almost natural. A unique type of data was used by Gregory and Gallagher [GG02]. They analyzed recordings from 19 debates of US presidential candidates from the 1960s until the present. Notably, free or semi-free conversations [CP81, Gre83, Gre86], with an occasional hidden experimenter appearing as a normal interactant[8] [Nat75a] (seemingly the best scenarios for tapping into convergence), were used rather rarely in comparison to the other dominant elicitation techniques. One of the possible reasons for this could be the lack of control over the elicited data. Therefore, the chances of performing more fine-grained acoustic measurements, or even perceptual judgments of similarity, would be worsened.

A step away from mere word repetition towards more naturalistic scenarios for studying convergence without totally losing control over the data has been the usage of Map Tasks [ABB$^+$01]. Map tasks involve two participants engaging in identifying the right path on a map. One person has the role of the instruction giver; the other must follow the instructions. The instruction giver, of course, possesses a map with not only the correct path, but also the relevant landmarks to describe the location. Those landmarks are only partially present or presented in a modified form[9] on the follower's map. Map tasks in this form were used in the convergence

[8] In Natale's study, a trained experimenter took part in the recording session acting as a "normal" test subject [Nat75a].

[9] Some landmarks have changed names or appear twice on the map.

studies of Pardo and Smith [Par06, Smi07]. Although map tasks have substantial benefits in that they provide certainly more naturalistic data than shadowing experiments while still allowing for some control over the linguistic content via the landmark manipulation, they also have several flaws. Since the speaker roles *Giver* and *Follower* are assigned right from the beginning of the task, it is very likely that a disparity arises in the turn taking and the amount of speech uttered by both participants. Our evaluation of a map task corpus collected for a different type of research [Cla07] seems to confirm that the speaking time of the instruction follower is usually shorter than their partner's and is very limited in terms of linguistic variation of syntactic structures and vocabulary used, including even the target landmarks. Map tasks might thus not elicit a balanced type of data from both speakers with equal amounts of speech and comparable quality[10], both of which are vital for assessing convergence between the speakers. A few current studies have already taken up the investigation of convergence with quasi-spontaneous [KHB11][11] or even fully spontaneous speech data [LORC11].

With the exception of a small number of studies dealing with children's speech, as e.g. Street and colleagues [SSVK83] and Oviatt and colleagues [ODC04], convergence has been predominantly measured in adults. One paper reports on the adaptation of parents to their preverbal children, arguing this to be an example of a reaction toward the needs of the interaction partner, however, triggered not by speech itself and without any verbal feedback [FTD$^+$89] (compare Bell's *audience design* in Chapter 1.1.3). Street and colleagues were able to find evidence for convergence in three-year-old children who were interacting in a play setting. The effect, however, was reported to be relatively unstable [SSVK83]. Oviatt and colleagues tested and found evidence for the prosodic adaptation of children aged from seven to ten to animated personas with different TTS-generated[12] voices.

[10] *Quality* here means a richness of utilized syntactic structures, morphological variation and vocabulary.

[11] Kim and colleagues [KHB11] use the same data elicitation technique as described in the current study – the Diapix game. For more details, see Chapter 4.1.3.

[12] text-to-speech-synthesis.

3.3 Convergence effects

Phonetic convergence has been investigated under many different aspects and has been linked to several internal and external factors that might influence its mechanisms. The following sections review some of these correlations and provide a commentary on the resulting implications.

3.3.1 Gender and dominance

A few studies found a gender effect in their data, Namy and colleagues and Pardo [NNS02, Par06] among them. Pardo and Namy's studies, however, report opposite patterns for speaker gender. Whereas Namy in her shadowing experiments found evidence for female talkers exhibiting more convergence [NNS02], Pardo found men to converge more than women [Par06]. Pardo comments on Namy's findings, discarding their hypotheses about women being generally better detectors of phonetic detail[13] and therefore better accommodators[14]. Instead, she suggests that attentional mechanisms might play a greater role here. Bailly and Lelong identified same-sex pairs in their study as converging more than mixed pairs [BL10], while Willemyns and colleagues' results point to a more complicated picture, where gender effects were mixed with other effects, for example, accent type [WGCP97][15]. The great majority of other studies have not found men and women to differ significantly in their degree of convergence.

Perceived dominance, status or group attachment/bias were features that have been investigated in many studies. In most cases it has proved to be somehow correlated to convergence [Bab09, BG77, Gil73, GW96, GG02, Par06, WvVed]. Bourhis and Giles' and Wedel and Van Volkinburg's studies dealt with the notion of greater in-group/out-group phenomena. Wedel reports from a computer simulation that features not relevant for group identification are most likely to underlie

[13] They are said to display greater "perceptual sensitivity" [Par06, 2389].

[14] compare Namy et al. [NNS02].

[15] Men diverged from an interviewer speaking with a cultivated Australian English accent, but did not do so with an interviewer using a broad Australian accent [WGCP97].

convergence, while group-distinctive features rather diverge amongst two groups in contact [WvVed]. Babel [Bab09] found convergence to her Black model talker to correlate with a pro-Black bias tested in an Implicit Association Task.

Gregory and Webster performed a factor analysis on LTAS for Larry King and his guests in the show and compared the data with subjective ratings of the status (popularity) of the respective guests. Results showed that less famous guests converged to their hosts and that Larry King in turn, converged to talkers with a judged higher status (the authors named these factors *dominance* and *deference* [GW96]). Another analysis of Gregory and collaborators [GG02] dealt with the correlation between the degree of convergence of presidential candidates in debates preceding the elections in the US and the factor of social dominance. Convergence was, in most cases, related to perceived dominance, which, in hindsight, proved an effective way to predict the outcome of elections.

Other factors which influence convergence and fall into line with predictions of the CAT framework, are *social desirability* and *social attractiveness*. The studies of Natale and colleagues [Nat75b, Nat75a, Nat76] used the Marlowe-Crowne-Scale to test the social desirability of the subjects. These studies have shown positive correlations between a greater need for social approval and the degree of convergence. Street [Str84], however, has related the convergence of speech latencies to mutual ratings of the social attractiveness and competence of his subjects.

3.3.2 Magnitude and persistence of the effect

As Cappella and Planalp argue, the magnitude of convergence effects is quite small, despite the fact that it remains detectable even for pairs of strangers [CP81]. This assumption runs against previous data, which has offered a picture with rather strong effects [MWSW63, MW67, Nat75a]. Cappella and Planalp confidently claim that the interspeaker effect is never as strong as a speaker's own consistency [CP81, 126]: "(...) it is unlikely that the influence between speakers is large enough to challenge the usual finding of individual consistency within conversations or across conversations with the same partner".

Nielsen [Nie07] also suggests that single-feature analysis might yield less robust results than a more global perceptual evaluation of the data, and that some effects require great statistical power to obtain valuable results. In her modified version of an imitation-design, she found subjects imitating lengthened VOTs but not shortened VOTs, which she explained by citing the influence of linguistic knowledge on the imitation. She sees this as a partial argument against experience-based models. The preservation of phonemic contrast and, therefore, divergence from the model might, however, be comparable to Wedel and van Volkinburg's results on simultaneous convergence and divergence [WvVed]. As has already been mentioned, computer simulations have revealed that group-distinctive features tend to be underlined through diverging from other group models simply to keep them distinct[16]. What works on a macro-scale for groups in contact should also be applicable on the micro-scale, where a mechanism preventing the loss of a phonemic contrast might result in divergence rather than convergence. This solution does not necessarily imply that exemplar models do not hold, since the specification of an exemplar for a production context is a very complex process and is, as was argued in Chapter 2.2, much more than the result of a simple input-output mechanism.

Cappella and Planalp also found conflicting data, including huge variation between pairs of speakers, where for some features (such as mean pause duration) most pairs displayed neither convergence nor divergence but rather maintenance of their own behavior. They suggest that the huge variation and mixture of divergence and convergence in the data should be accounted for by showing the exogenous and endogenous factors behind it, including a.o. personality or attraction. Another conclusion drawn by Cappella and Planalp from the small size of convergence effects is the suspected lack of conscious awareness of the speakers' mutual influence on each other. They go even further, claiming that not consciously being aware of

[16] The Australian vowel shift can be interpreted as a great example of diverging in order to preserve phonemic contrast in a language on a large scale. Once one vowel has been changed, the others needed to become subject to language change as well for the system to work.

the other's influence excludes more complex cognitive judgments[17] from having an influence on the convergence effect. This strong conclusion, however, appears unwarranted in the light of current views on (sub)consciousness and awareness, explained in detail in Chapter 2.2.3.

Gregory and Webster [GW96] pointed to an interesting relation occurring within dyadic encounters: convergence was found to operate within a state of *balance* between the partners. In cases in which one interactant fails to reach a satisfying level of convergence, the other seems to compensate for this failure by converging himself. The authors commented on this phenomenon with the famous words: "If the mountain won't go to the prophet, the prophet must go to the mountain" [GW96, 1237].

In terms of the persistence of convergence effects, Delvaux and Soquet [DS07] as well as Pardo [Par06] report post-test findings for their data, indicating that convergence[18] within a conversation carries over to items beyond the scope of the respective interaction. Pardo used landmarks from the map task for her pre- and post-tests that were embedded in carrier phrases and then read by the subjects before and after performing the map task. Delvaux and Soquet used the same sentences prompted via ideograms on a computer screen for all parts of their experiment, with the subjects performing the task alone (without the influence of the reference speaker) in the post-test condition. The authors reported data for one subject who started the post-test by pronouncing adapted target vowels and only gradually returning to her own pronunciation[19] [DS07]. Another interesting phenomenon is the apparently very rapid accommodation toward the reference speaker. In some cases this happens in the course of only a few encounters of the same token [DS07]. An unexpectedly quick adaptation toward the conversational partner was given as a possible explanation for the lack of evidence for convergence in the cultivated accent condition in Willemyns' study [WGCP97]. They suspected a possible ceiling

[17] As, e.g., attractiveness, empathy or competence [CP81].

[18] Delvaux and Soquet refer to the longer-lasting post-effect as "mimesis" in contrast to "imitation", which they consider to be only limited to the moment of speaking. Mimesis is supposed to involve updating existing representations with new variants [DS07].

[19] Within about 6-10 minutes.

effect, where subjects started with a higher accent variant instead of slowly adjusting to it.

3.3.3 Summary

Although many questions concerning the mechanisms of convergence have already been answered, the applied methodology so far has not yet covered many detailed aspects of segmental pronunciation, since most studies have focused on the broad area of prosody. In those cases in which prosody was not in the foreground, the elicitation techniques were mostly based on shadowing/repetition designs or the focus was laid on perceptual judgments of similarity, and not on taking acoustic measurements of segmental details. Many authors have acknowledged the existence of extraverbal factors influencing convergence, but, apart from a few personality measures and social status differences, no strictly *individual* differences have been investigated.

As will be shown in Chapter 4, the methodology applied in this thesis does not build on a parameter-tracking front-end but rather allows for non-subjective acoustic measurements instead of perceptual judgments only. The elicitation technique has improved upon several aspects compared to map tasks, and the measurement is of a global nature instead of single-feature tracking, as the identification and the judgment of its relevance is an extremely difficult – if not impossible – task.

Chapter 4

Methods and data analysis

The following chapter contains a description of the experimental procedure and the applied methods of data analysis. The individual sections will provide information about the technical details and highlight the crucial differences between the currently applied and previously used methods in convergence research.

4.1 Design of the study

The experimental set-up was chosen especially to fulfill the needs of the basic research questions, namely:

- How do convergence phenomena surface in *native-nonnative dialogs*?
- Which laboratory conditions are as close to a *natural scenario* as possible?
- How to (covertly) encourage the nonnative speakers to perform well in their second language and express themselves in a clear way[1]?

The stated research questions required subjects with two different first languages, an elicitation technique that would yield quasi-natural spontaneous speech, and the provision of an encouraging setting for high linguistic performance that would at the same time deliver enough material for conducting the planned acoustic analyses on the word-level. The subsequent planning of the study will be laid out in more detail in the following sections.

4.1.1 Subjects

The experiment was designed to provide a detailed picture of phonetic convergence within native-nonnative conversations. Two English native speakers and 20 native speakers of German were recruited for the study. The native speakers of German were high-proficiency users of English as an L2 at the time the study was conducted.

[1] compare communicative reasons for adaptation in dialog in Chapter 1.3.2 for the general mechanisms, and Chapter 1.3.3 for more details.

The first English native speaker – **T** – was male, 33 years old and spoke with a General American accent (henceforth abbreviated as *GA*). The second English native speaker – **J** – was female, 57 at the time of the study, and was a Southern Standard British English speaker (abbreviated as *SSBE*). A male and a female speaker were deliberately chosen to look into the question of gender-effects in phonetic adaptation in dialog[2]. It was also deliberately decided to have an American and a British native speaker as dialog partners for the NNS in order to see whether any accent preferences of the NNS can impact convergence.

The German native speakers were between 20 and 42 years old, ten female and ten male speakers coming from the greater region of Stuttgart in southern Germany. All had the same EFL[3] background, including the following:

- All started learning English as a foreign language at school in fifth grade
- None has stayed in an English-speaking country for longer than 3 weeks
- All were tested as proficient users of English in a comprehensive test[4]

The German test subjects were chosen from the subject pool of a preceding project "Language talent and brain activity – the neural basis of pronunciation talent" funded by the German Research Council DFG[5] conducted at the University of Stuttgart and the University of Tübingen. Within this project all subjects have been extensively tested on their phonetic abilities and accordingly categorized into three groups [DR09]:

1. highly phonetically talented
2. normally talented/standard performance in pronunciation

[2] These have been reported in the literature. Compare Chapter 3.3.1 for details.

[3] English as a foreign language.

[4] Carried out within the project "Language talent and brain activity – the neural basis of pronunciation talent" funded by the German Research Council (DFG) DO536/6-1 and AC55/7-1; more details in the following and in Jilka 2009 [Jil09a, Jil09b].

[5] German title "Zerebrale Korrelate der phonetischen Fremdsprachenbegabung", projects DO536/6-1 and AC55/7-1.

3. rather untalented in terms of pronunciation

For the current study speakers were chosen from these groups, but with an emphasis on members of groups 1 and 3, in order to ensure a reliable comparison of the behavior of highly talented versus non-talented speakers in the study.

The English native speakers participated in the recordings with all 20 German subjects and were informed about the research questions. They were explicitly instructed not to reveal the purpose of the study to the German participants and were asked not to adapt their speaking style in any way to their conversational partners[6]. None of the German subjects was informed about the real reason for the study; they were simply asked to participate in two task-oriented dialogs with native speakers of English, the goal of which was collaborative work on a "spot-the-difference" game[7].

4.1.2 Experimental set-up

The experiment consisted of five parts: the pre-test, the first dialog, the mid-test, the second dialog, and the post-test. Table 4.1 gives an overview of the procedure.

Block 1	**Block 2**		**Block 3**	**Block 4**		**Block 5**
pre-test	**dialog 1**		mid-test	**dialog 2**		post-test
	Diapix	summary		Diapix	summary	
read	spoken		read	spoken		read

Table 4.1: Sequence of the five test blocks.

Blocks 1, 3 and 5 contained a reading task, with the same speech material each time. The reading list consisted of words from the Diapix target words[8] and filler words[9]. The German participants (NNS) were asked to read the list before the first dialog, between the two dialogs, and after the second dialog. The English native

[6] i.e., they were asked to maintain their own speaking style as far as possible, compare Chapter 1.3.1 on CAT.

[7] See Chapter 4.1.3 for details on the experimental task.

[8] See Chapter 4.1.3 for a detailed account of the elicitation technique and the target words.

[9] See Appendix A.

speakers (NS) were asked to read out the list once in the beginning of the recording sessions to ensure a model for the list comparison with the NNS.

Blocks 2 and 4 were the dialog sessions between the English native speakers and the German subjects. Every German subject took part in both dialogs with NS1 and NS2, one at a time. The instructions in the dialog sessions consisted of solving a Diapix-task – a picture matching game with two pictures differing from one another in ten details[10]. The participants were instructed to cooperate with their conversational partner to find all changed or missing items and to use only English throughout the conversation. Once all items had been identified (or once it was reported that no more items could be found), the NNS were asked to give a short summary. Noteworthy here is that the NNS had to make a sudden switch from dialog to monolog speech, though the dialog partner was still present in the sound attenuated booth at the time the summary was given.

The whole recording session was performed en bloc, on one day. The timing interval between block 1 and block 2, and 2 and 3 was approximately 1-2 minutes; the break between block 3 and 4 was approximately 15-20 minutes; block 5 followed after only a-one-minute break. The dialog recordings, including the summary task, had a length of approximately 12-20 minutes[11], and each reading of the word list took about 1-2 minutes. The whole recording session, including all five blocks took approximately 90 minutes for each NNS participant.

4.1.3 Elicitation technique

In the search for a reliable elicitation technique for quasi-spontaneous speech data, it was deliberately decided against the usage of Map Tasks [ABB$^+$01]. Although widely used for corpora collection in phonetics and even for convergence research [Par06, Smi07], it proved to provide very biased data[12]. In order to acquire a bal-

[10] See Chapter 4.1.3 for a detailed description of the Diapix, and Appendix B for the respective pictures.

[11] The length of the conversation seems to have depended on the respective NS partner, with the dialogs of the SSBE speaker *J* being on average longer.

[12] See Chapter 3.2 for details.

anced amount of speech data from both speakers consisting of a wide range of complex utterances, it was decided to use Diapix in this study [BBC+07, vEBBB+10]. Diapix is a picture matching or “spot-the-difference” game comparable to those often found in newspapers and magazines. Each set contains two pictures which differ from one another in ten details in the following ways:

- items have different names
- items have different colors/shapes
- items are located at a different spot
- items are completely missing

For the two dialogs two sets of pictures were used, a different one for each of the conversations in order to ensure that the NNS really have to focus on solving the task at hand. Since the two English native speakers took part in all 20 conversations with the German subjects, they eventually came to know the location and nature of all target items. They were explicitly asked not to reveal their knowledge and act as if they were seeing the pictures for the first time, thereby pretending to be interested in a real purposeful interaction with their conversational partners in which they share the same knowledge status.

The two sets of pictures used for the recordings were the shop scene and the farm scene (see Figure 4.1 and Appendix B for all pictures used). The pictures were provided in DIN-A4 size and were laminated to reduce any noise from rustling paper during the recordings. The subjects were additionally given colorful pens to mark or cross out the items already found.

The Diapix task requires an intensive interaction between the two partners to identify all target items, and has no predefined talker roles of instruction giver or follower, as is the case in a Map Task. Both speakers are free to describe what they see and ask the other questions at any time. This technique allows the collection of balanced amounts of speech data with a wide range of utterance types and more complex responses to questions (than, e.g., in Map Tasks) while maintaining

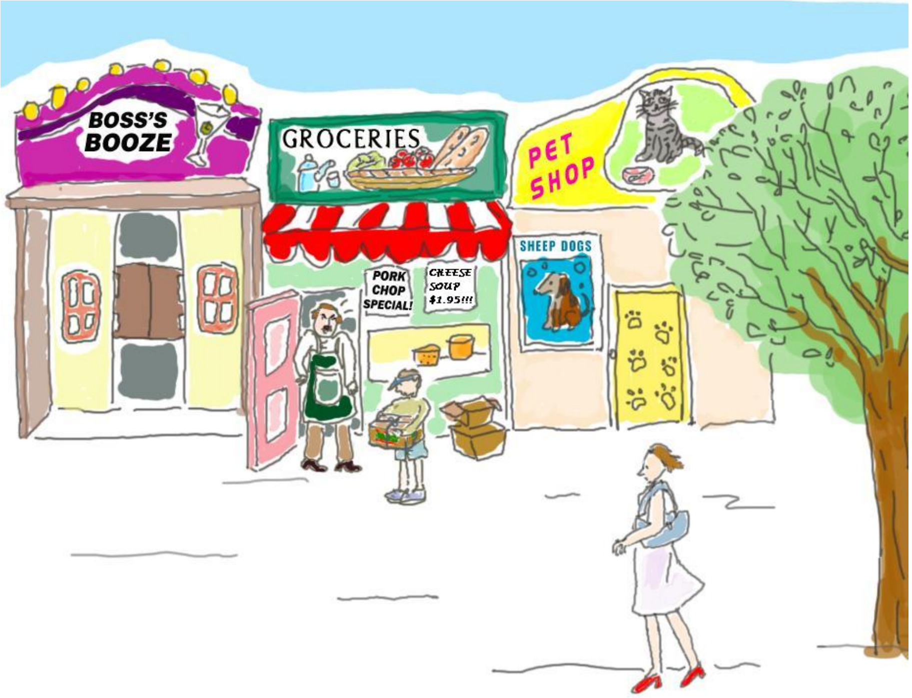


Figure 4.1: The first picture of the Diapix shop scene [BBC+07, vEBBB+10]. The remaining three pictures can be found in Appendix B.

a balanced talker role relation [vEBBB+10]. Status differences between the talkers resulting from the native (expert) vs. nonnative (learner) identity or any other possible feature, however, remain present in the Diapix setting. The target words used later for the acoustic analysis of convergence were mostly the words appearing within the changed/missing items, as e.g. *tomato, shop, shoe, cat, dog*[13], and other content words frequently used by the speakers. The target words being located at the spots of the changed/missing items in the picture sets ensures a sufficient number of repetitions by both speakers during the conversation. This was necessary for carrying out the acoustic analyses described in Chapter 4.2.

4.1.4 Recordings

The recordings were carried out in a sound attenuated booth at the Institute for Natural Language Processing (IMS) in Stuttgart. During the dialogs both subjects

[13] These can be seen in Figure 4.1.

were placed in the same sound attenuated booth and were separated by a padded wall placed in the middle of the room. The participants could not see each other. In order to ensure undisturbed communication despite the particular attenuating characteristics of the chamber, participants wore headphones to hear each other clearly. They were also able to hear the experimenter and follow any instructions given for the tasks.

Two head-mounted *AKG C520* microphones and two *AKG K271 MkII* headphones were used for the recording. The dialogs and reading tasks were initially digitized to a 48 kHz stream with two separate channels, which was downsampled later on to 16 kHz for further signal processing.

4.2 Data analysis

4.2.1 Target word extraction and labeling

After downsampling the recordings to 16 kHz, the speech material from all experimental blocks was further processed to break down the speech stream into the analysis level of words. Since automatic alignment of conversational speech with varying speed and colloquial expressions proved to be difficult and highly error-prone, the subsequent correction stage would have been too time-consuming. Instead it was decided to manually extract all target words from every dialog – for both talkers – and from the three instances of the word lists from all participants.

The dialog target words were always labeled with the speaker's initials and given the markers *early*, *late* or *sum*[14], depending on the stretch of time in the dialog they were taken from. All items labeled *early* were taken from the first third of the dialog, the *late* items were uttered in the last third of the dialog, and the *sum* items stemmed from the summary the NNS participants were asked to give at the end of each Diapix dialog.

[14] *sum* standing for "summary", further explanation is given in Chapter 4.1.2.

The extracted target words were furthermore evaluated according to the following criteria and accordingly labeled:

- *a* – (unnaturally) intensively aspirated
- *b* – belonging to/part of a phrase
- *c* – coarticulated with preceding/following segment or phrase
- *f* – fast
- *g* – glottal/creaky voice
- *h* – (unnaturally) high pitch
- *i* – intonation (unnatural)
- *l* – loud
- *L* – laughter
- *m* – missing sound/mispronounced
- *n* – nasal
- *o* – long
- *p* – plop sound/noise
- *q* – quiet
- *r* – interrupted by another person talking
- *sh* – short
- *u* – unclear

If the extracted target word was judged to be too unnatural or unclear due to the listed phenomena and a reliable comparison to another item was not likely, this particular target word was not taken into consideration for the amplitude envelope analysis[15].

4.2.2 Amplitude envelopes

Measurement methods used to investigate phonetic convergence have centered around either individual features, such as formant values or VOTs[16], or more coarse-grained variables over longer stretches of speech (such as LTAS). The disadvantage of the first method is that it is very hard to determine where exactly (i.e. for which acoustic feature) convergence will surface. The adaptation of one speaker to another might, for instance, be only marginal and barely detectable at all for just one small acoustic feature like a vowel formant. Taking a more global measurement, which comprises several or even all features present in the speech signal, should make it possible to capture convergence happening literally “everywhere” in the acoustic signal. More coarse-grained measures are able to provide exactly that, but unfortunately they are based on longer stretches of speech and do not cover more precise comparisons where one would actually still be able to tell which element resembles which more closely after convergence occurred. What is therefore needed is a fairly precise measurement (operating, for instance, at word level) that need not be pinned down by any specific features.

These requirements are met by the slowly varying amplitude envelopes which reflect the amount of energy present in the separate frequency bands of the acoustic signal.

Two crucial things can be gained through the usage of amplitude envelopes: on the one hand, no single feature tracking is necessary, and on the other, the measurement can be done at word-level without the need to resort to higher levels of speech like sentences or even whole utterances as a means of comparison, as these would blur the picture. Amplitude envelopes can be seen as [WDS$^+$10, 231] “rep-

[15] Common cases for such an exclusion were laughter, interruption by the speaking partner, noise, unclear pronunciation or an extremely creaky voice.

[16] Compare Chapter 3.1 for details.

resentations that more faithfully encode the speech signal as it unfolds over time without making specific assumptions about what types of cues might be extracted or which regions of the signal are the most important".

The calculation of amplitude envelopes thus allows us to capture not only a momentary static image of the speech signal and the extraction of one specific feature, but also enables a comparison of speech signals as they unfold in time, adding an important dynamic component. It has already been concluded that the information contained in amplitude envelopes is present in the auditory system, and it is also enough to build intelligible speech, with properties at least approaching those of natural speech as long as enough frequency bands are given (in order to provide at least nominal spectral resolution[17]) [WDS$^+$10, LDT99, SZK$^+$95]. As has already been mentioned, a great advantage of amplitude envelopes lies in their lack of underlying abstract dimensions in the speech signal that would make it necessary to assume front-end analyses [WDS$^+$10]. In addition, their transparent and compact form facilitates their usage for convergence measurement.

Wade and colleagues [WDS$^+$10] assume that envelopes might represent a part of the information stored and used by humans in speech perception and production. They claim that the signals are stored as linear time sequences, which enables a simple comparison of two such signals by using a cross-correlation function (see Chapter 4.2.2.4 and Appendix C) [WDS$^+$10].

A Matlab script (see Appendix C) was used to calculate the amplitude envelopes for all manually extracted target words and to calculate and return their match values. The scores showed the degree of similarity of the two signals. The components of the script will be described in the following subsections.

[17] the number of frequency bands the signal is divided into can be changed according to the precision requirements of the measurements. It proved to be sufficient to use four frequency bands for the current analysis. As few as three to four frequency bands have been shown to be enough to allow for proper speech recognition. See [SZK$^+$95, LDT99] for more details.

4.2.2.1 Working loop script

After reading the file names provided in the working path, the first loops[18] extract the respective name tags, for the target word and the subject name[19] for the first and the second wave file (*extract name tag*):

- example file name for a dialog item: *bench_ABJ_1oearly*[20]
- example file name for a word list item: *bench_AB_1*[21]

The loop contains a filter for dialog items and read list items, which allows the comparisons to be grouped between words from the two dialog partners; or from the read speech produced by both the NS and NNS in the word list reading task (*filter dialogs*); or a cross-comparison of dialog and word list items. The filter can be switched on or off. The following loop calls the script *sampleMatch_nl* for producing a final output of the comparison results – the *match values*. The output is saved as a Matlab *mat*-file and is also exported to a csv-format, readable with a standard Microsoft Excel package.

4.2.2.2 Sample match script

The sample match function[22] returns the match between two given wave files (*dat1* and *dat2*) and is linked to the *sound to envelope* and *matchVal* functions, both of which will be described later.

The script starts with setting the values of the parameters *root mean square KRMS* to 0.03; the sampling rate of the original wave signal *KFS* is set to 16000Hz and the envelope sampling rate *KENVFS* is 500Hz. A first amplitude normalising procedure is then introduced (*normRMS*) and the lowest and highest cutoff frequencies of the frequency bands are defined – 80Hz for the first band (*band_lo*) up to

[18] See Appendix C.1 for the script.

[19] i.e. the coded initials of the speaker in the respective list or dialog

[20] The last part of the string after the second underscore contains the sequential number of the item, any special symbols used (defined in Chapter 4.2.1) and time information (early, late or summary).

[21] The number indicates the list number – respectively 1, 2 or 3.

[22] See Appendix C.2 for the script.

Input argument	Explanation
wd1, wd2	the filtered and normalized signal
KFS	the original sampling rate (16000 Hz)
band_lo, band_hi	the low and high cutoff of the frequency bands
4	the number of frequency bands
60	cut-off frequency for the envelope filter
abs(hilbert)	the absolute value of the outcome of the Hilbert transform
KENVFS	the envelope sampling rate (500 Hz)

Table 4.2: The table comprises the input arguments used for the extraction of the amplitude envelopes in the *sample match* script. Compare Appendix C.2.

7800Hz for the highest (fourth) band (*band_hi*). The sound files then undergo high-emphasizing to give more weight to their lower-amplitude higher frequency range (*hiEmph* function). The coupled procedures of high-emphasis filtering and amplitude normalizing return the new normalized and filtered variables *wd1* and *wd2*. The next step is the extraction of the amplitude envelopes (*env1* and *env2*), using the sound-to-envelope function[23] and the input arguments[24] given in Table 4.2.

The comparison of the two envelopes proceeds twofold – separately for every frequency band (displayed in the rows for *x1, x2, x3 and x4*) using the first part of a Dynamic Time Warping function[25] and using a cross-correlation function contained in the *match_val* function (see Chapter 4.2.2.4 on the envelope match script). The second method requires a further normalisation of the envelopes to fit the results into the [0-1] range.

[23] described in detail in Chapter 4.2.2.3 (*sound to envelope script*).

[24] The Hilbert transform, given in Table 4.2, returns a discrete-time analytic signal, composed of a real part (=the original signal) and an imaginary part (containing the Hilbert transform). It is used for the calculation of instantaneous attributes of a time series, hence also amplitude. The function is implemented in the Signal Processing Toolbox of the Matlab Software Package [Mat11].

[25] The *simmx.m* routine calculates “the full local-match matrix i.e. calculating the distance between every pair of frames from the sample and template signals” [Ell03, simmx.m].

Input argument	**Explanation**
s	original signal
iFsOrig	original sampling rate (Hz)
fTotFreqRange	total frequency range to be considered (*hi* must be smaller than *iFs*/2)
iNumBands	number of frequency bands within *fTotFreqRange*
fEnvCutOff	cut-off frequency (Hz) for the envelope filter
fhEnvMethod	handle of function to extract the (raw) envelope
iFsNew	new sampling rate (should be bigger than twice the *fEnvCutOff*)

Table 4.3: The table comprises the used input arguments for the calculation of the envelope *e* in the *sound to envelope* script, compare Appendix C.3.

4.2.2.3 Sound to envelope script

The sound to envelope function[26] calculates the amplitude envelopes from the defined frequency bands and returns the envelope *e*, the band center frequencies *CFs* and the band-separated original sound *sB*. It takes the input arguments specified in Table 4.3.

The script starts with calculating the log-spaced low and high cutoff frequencies, the center frequencies and initializes the band-separated signal *sB*. A low order *Butterworth filter* (*butter*)[27] of the *ftype*[28] '*bandpass*' with the syntax – *[b,a]=butter(n,Wn,'ftype')*[29] is then applied to filter the signal into the appropriate bands. The raw envelopes calculated from the band-passed signal (new *sB*) in the manner described above are then low-pass filtered, using a fourth-order Butterworth filter. The second applied filter function returns the variable *e*, which is the low-pass filtered, band-separated envelope of the original signal (see Figure 4.2).

[26] See Appendix C.3 for the script.

[27] A Butterworth filter "is characterized by a magnitude response that is maximally flat in the passband and monotonic overall" [Mat11, But30]. It can function as a lowpass, highpass, stopband or bandpass filter. The filter applied here is of the 'bandpass' type. The higher the order, the steeper the slope of the filter curve.

[28] i.e. *function type*, predefined in the Matlab code.

[29] *n* being the order, in this case – 2; and *Wn* being the cutoff frequency. See Appendix C.3 for the input argument.

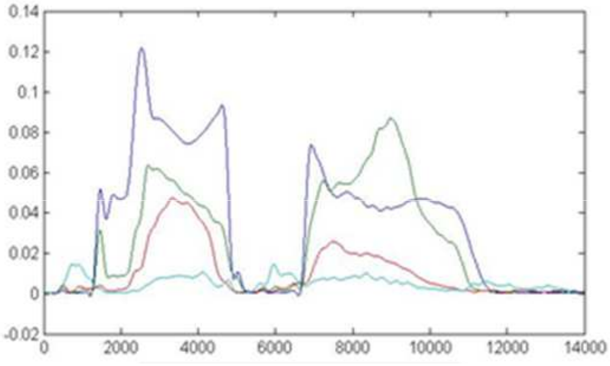


Figure 4.2: The amplitude envelopes for the four frequency bands in the word *tomatoes*. The Y axis represents amplitude and the X axis represents time in samples.

4.2.2.4 Envelope match script

The envelope match script[30] returns the final *match value matchVal* of the two amplitude envelope signals *e1* and *e2* using a cross-correlation function.

After defining which envelope is longer in the *length_difference*-loop and returning it as the variable *maxLag*, the function compares the shorter and the longer envelopes, separately for every frequency band of the signal, to return an estimate of similarity between the two (the variable *matchSum*). The used cross-correlation function *xcorr*[31] is a built-in implementation in Matlab. The final result *matchVal* is the maximum value of the cross-correlation output *matchSum*.

4.2.2.5 Function evaluation script

The function evaluation script[32] tested the two methods of calculating match values – DTW-based simmx vs. cross-correlation – on a small test set with two identical wave files, two completely distinct wave files, and three pairs of similar items, all

[30] See Appendix C.4 for the script.

[31] The cross-correlation function here takes the syntax ‘c = xcorr(x,y,maxlags)’ and returns the cross-correlation sequence over the lag range [-maxlags:maxlags] [Mat11].

[32] See Appendix C.5 for the script.

uttered by the same speaker. Test results indicated that both methods return values in the expected ranges, with 1 for identical signals, very low/close to zero values for completely distinct signals, and high values approaching .80-.90 for fairly similar signals. Further results will be discussed in Chapter 5 on the results of the dialog analysis.

Chapter 5

Results: dialog speech

The first results section will describe the data analyses of all dialog measurements. All reported analyses were calculated using SPSS 19 [Inc11]. Phonetic convergence was measured and analyzed for both relevant directions: nonnative speakers toward native speakers and vice versa, in order to capture not only the nonnative speakers' adaptation but also its possible interactions with the behavior of the native speakers. Additionally, a measurement of self-consistency was introduced, to test how close to their own pronunciation our German subjects and the English native speakers stayed.

The following hypotheses were formulated regarding the dialog speech data:

1. **Hypothesis A1**
   Nonnative speakers converge to their native speaking dialog partners.

2. **Hypothesis A2**
   Talented speakers converge more to their partners than less talented ones.

3. **Hypothesis A3**
   Female subjects converge more in dialog than male subjects.

4. **Hypothesis A4**
   Talented subjects will show more perturbed self-consistency values than less talented ones in the dialog.

5. **Hypothesis A5**
   The level of adaptation to the conversational partner persists after the switch from dialog to narrative.

6. **Hypothesis A6**
   The English native speakers converge to their nonnative speaking partners.

7. **Hypothesis A7**
   At the end of the dialog the speakers reach a level of convergence balance which is higher than at the beginning.

Evidence supporting or rejecting these hypotheses will be presented in Chapter 5.1 and Chapter 5.2. In order to test the presented hypotheses, the following match value measurements were obtained from the native and nonnative speakers in the dialogs for the times indicated[1] – the *match values* were taken for:

- Set 1 –*X* early[2] vs. *Y* early[3]
- Set 2 – *X* late vs. *Y* early
- Set 3 – *X* late vs. *Y* late
- Set 4 – *X* early vs. *Y* late
- Set 5 – *X* summary vs. *Y* late
- Set 6 – self-consistency[4] of *X* early vs. early in both dialogs
- Set 7 – self-consistency of *X* early vs. late in both dialogs

A comparison of **Set 1** and **Set 2** will provide information about the convergence of our 20 German subjects in both dialog conditions. **Set 1** and **Set 3** will provide a value for the mutual convergence or *balance level* of both conversational partners in the course of the dialog, while comparing **Set 1** and **Set 4** reveals whether the native speakers *J* and *T* converged to their interactants (in spite of being told to maintain their own speaking style). Putting the values for **Set 5** into relation with **Set 3** can reveal whether a direct switch in speaking styles, away from dialog

[1] Tables with an overview of the mean match values in all tested conditions are presented in Appendix D (Figure D.1, Figure D.2, and Figure D.3).

[2] X standing for the German subject; nonnative speaker.

[3] *Y* standing for either *J* – the British English speaker, or *T* – the American English native speaker.

[4] i.e., the measurement indicating how true to their own *read* speaking style the subjects stayed in the course of the whole experimental session, measured as the comparison between an early and late set in the dialog, and, in case of the nonnative speakers, also in the summary.

toward a narrative, leads to a change in the level of convergence. Finally, a comparison of **Set 6** and **Set 7** allows a closer look at the self-consistency of the nonnative speakers throughout the dialogs.

5.1 Results of nonnative speaker performance

The first focus of the dialog results chapter lies on the behavior of the nonnative speakers – the German subjects. The analysis of their accommodation during the dialogs is tied to five of the hypotheses: **Hypothesis A1** about the general convergence of the nonnative speakers toward the English interactants; **Hypothesis A2** investigating the impact of *talent* as an individual difference; **Hypothesis A3** on the influence of gender; **Hypothesis A4** concerning the self-consistency of the NNS throughout the experimental task; and **Hypothesis A5** related to the persistence of convergence after a style change from dialog to monologue.

5.1.1 Convergence toward English native speakers

The following sections are concerned with the analysis of the comparisons between the **Sets 1, 2, 3** and **Set 5**, which are crucial for determining the phonetic convergence of the nonnative speakers toward the native speakers. The self-consistency data of the nonnative speakers (**Set 6** and **Set 7**) will be presented in Chapter 5.1.2.

Tables 5.1, 5.2 and 5.3 show the t-test statistics for the whole group of German subjects in their convergence toward the English native speakers within the dialog. Table 5.1 gives the means and standard deviations for the values **Set 1** and **Set 2** for the dialogs with native speakers *J* and *T*. The mean values for **Set 2** (the late comparison in the dialog) are higher in both cases. The two sets of values are also significantly, though only mildly, correlated (see Table 5.2).

The results of the paired samples t-test in Table 5.3 point to significant changes between the values given in **Set 1** and **Set 2**, with a significance of p<.01 in both conditions (p=.005 for condition *J* and p=.003 for condition *T*). The mean dif-

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set1J	.7328	20	.03468	.007755
	Set2J	.7591	20	.04953	.011076
Pair 2	Set1T	.7255	20	.02364	.005286
	Set2T	.7489	20	.03961	.008857

Table 5.1: Mean, standard deviation (SD) and standard error mean for **Pair 1**: the comparison of **Set 1** and **Set 2** for matches with native speaker *J*; and **Pair 2**: the comparison of **Set 1** and **Set 2** for matches with native speaker *T*. Set 1, as defined previously, indicates the mean match value of early items of the NNS and NS, while Set 2 shows the mean match values of late items of the NNS vs. early items of the NS. All values are calculated for the complete set of German subjects - 20 speakers.

		N	**Correlation**	**Significance**
Pair 1	Set1J & Set2J	20	.662	.001
Pair 2	Set1T & Set2T	20	.627	.003

Table 5.2: The correlation between the paired match values for Set 1 and Set 2.

		Mean	**SD**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
Pair 1	Set1J – Set2J	-.0263	.03715	.008308	-3.169	19	.005
Pair 2	Set1T – Set2T	-.0235	.03088	.006905	-3.401	19	.003

Table 5.3: Paired samples t-test for **Set 1** and **Set 2** *J* (Pair 1) and **Set 1** and **Set 2** *T* (Pair 2). The significance level of the changes is given in the last column.

ference of values (Set 1 - Set 2) is negative, showing an increase of match values between the sets. It can therefore be concluded that there is a significant increase in match values, signifying convergence of the nonnative speakers toward the native speakers between an early and late point in the dialog. The results of the analyses shown allow a rejection of the null hypothesis $A1_0$, which stated there are no differences between Set 1 and Set 2. **Hypothesis A1**, therefore, is correct: nonnative

		convergence_XT
convergence_XJ	Pearson correlation Sig. (2-tailed)	.696** .001

Table 5.4: The Pearson correlation and its significance for the amount of convergence (**convergence_XJ** and **convergence_XT**) of the nonnative speakers in dialog *J* and dialog *T*. All values are calculated for the whole set of N=20 subjects.

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set1J	.7269	10	.02386	.007746
	Set2J	.7301	10	.04715	.014912
Pair 2	Set1T	.7180	10	.02015	.006372
	Set2T	.7178	10	.01974	.006245

Table 5.5: Mean, standard deviation (SD) and standard error mean for **Pair 1**: the comparison of **Set 1** and **Set 2** for matches with native speaker *J*; and **Pair 2**: the comparison of **Set 1** and **Set 2** for matches with native speaker *T*. All values are calculated for the subset of 10 less talented speakers.

speakers do converge to their native speaking dialog partners.

The amount of convergence of the nonnative participants toward the native speakers in both dialogs is correlated (Table 5.4). This is to say that a subject who converged strongly in one of the dialogs was quite likely to do so again in the second dialog. Nonnative speakers who, on the other hand, failed to converge or diverged in one dialog, most likely did the same in the second dialog as well. Figure 5.1 displays a scatter plot of the nonnatives' convergence in both dialogs, with an added reference line.

As far as the relation of the talent component and convergence stated in **Hypothesis A2** is concerned, separate t-tests for the two groups were carried out (Tables 5.5-5.8). The descriptive statistics and the paired samples test results are given in Table 5.5 and Table 5.6 for the less talented group, and in Table 5.7 and Table 5.8 for the talented group. The means for Set 1 and Set 2 for both dialogs (J and T) given in Table 5.5 show only a minimal change of the mean value in the *J* condition

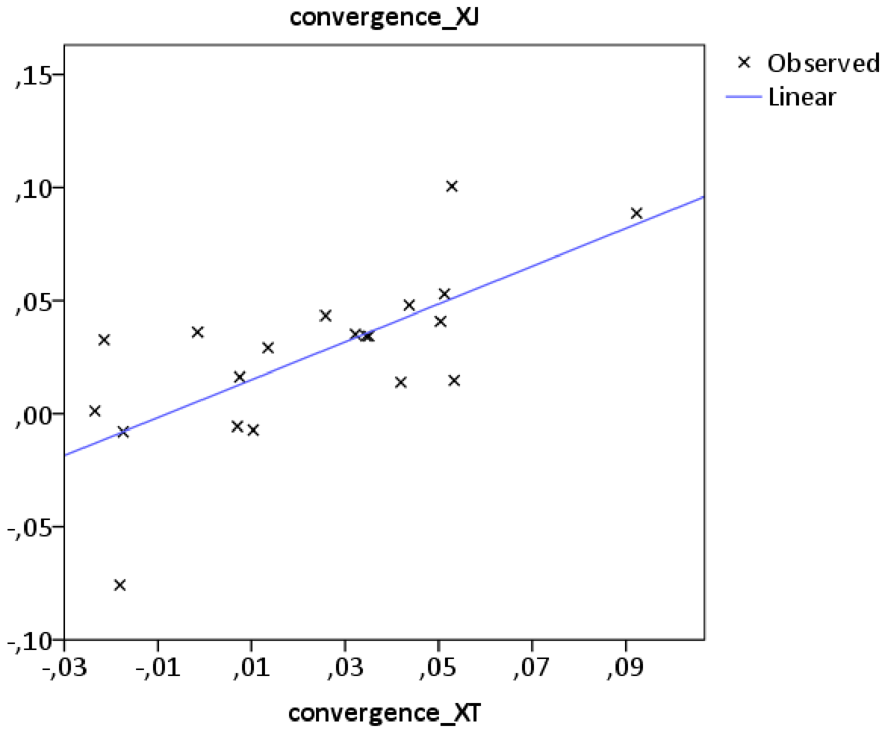


Figure 5.1: Scatter plot of the degree of the German subjects' convergence for both dialogs. The linear relationship of the two data sets is significant at the .01 level. The reference line is an estimated linear regression curve for the two data sets *convergence_XJ* and *convergence_XT* ($R^2 = .484$, std. error=.027, beta coefficient .696).

		Mean	SD	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	Set1J – Set2J	-.0033	.03245	.010263	-.322	9	.755
Pair 2	Set1T – Set2T	.0002	.02054	.006496	.026	9	.980

Table 5.6: Paired samples t-test for Set 1*J* and Set 2*J* (Pair 1) and Set 1*T* and Set 2*T* (Pair 2) for the subset of less talented speakers. The significance level of the changes is given in the last column.

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set1J	.7387	10	.04351	.013758
	Set2J	.7880	10	.03314	.010479
Pair 2	Set1T	.7330	10	.02547	.008053
	Set2T	.7801	10	.02765	.008743

Table 5.7: Mean, standard deviation (SD) and standard error mean for Pair 1: the comparison of **Set 1** and **Set 2** for matches with native speaker *J*; and Pair 2: the comparison of **Set 1** and **Set 2** for matches with native speaker *T*. All values are calculated for the subset of 10 talented speakers.

		Mean	**SD**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
Pair 1	Set1J – Set2J	-.0494	.02613	.008264	-5.972	9	.000
Pair 2	Set1T – Set2T	-.0471	.01865	.005899	-7.990	9	.000

Table 5.8: Paired samples t-test for Set 1*J* and Set 2*J* (Pair 1) and Set1*T* and Set 2*T* (Pair 2) for the subset of talented speakers.The significance level of the changes is given in the last column.

and almost no change in the *T* condition. This is supported by the outcome of the t-test in Table 5.6 for the group of *less talented* speakers: the changes between an early and a late point in the dialog for this group are not significant (p=.755 for *J* and p=.980 for *T*).

The comparison of **Set 1** and **Set 2** for the *talented group* portrays a reversed picture (Table 5.7 and Figure 5.8). Here, the differences in mean match values for **Set 1** and **Set 2** are positive and visibly larger than in the *less talented* group (approx. .05 in both conditions[5]). The paired samples t-test shown in Table 5.8 confirms a significant change between **Set 1** and **Set 2** (p<.000 for both *J* and *T*) for the group of *talented speakers*. Their match values with the native speakers

[5] The *mean* values in Table 5.8 are negative, since they were calculated as the difference of "**Set 1-Set 2**" and the latter set contained the higher values. The negative number nevertheless points to an increase of values in time, not a decrease.

Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
Time	Sphericity assumed	.007	1	.007	15.967	.001
Time*Talent	Sphericity assumed	.005	1	.005	12.215	.003
Error (Time)	Sphericity assumed	.008	18	.000		

Table 5.9: Repeated measures ANOVA – within-subject effects for Set 1 and Set 2 (factor **Time**) in condition *J*.

Source	**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
Intercept	22.256	1	22.256	9054.965	.000
Talent	.012	1	.012	4.929	.039
Error	.044	18	.002		

Table 5.10: Repeated measures ANOVA – between-subject effects for the factor **Talent** in condition *J*.

increased between an early and a late point in the dialog.

Tables 5.9 and 5.10 display the within- and between-subjects effects of a repeated measures ANOVA for the *J* condition with the within-subjects factor *Time* and between-subjects factor *Talent*. *Time* is defined as the two measurements - Set 1 and Set 2 - taken respectively at an early and a late point in the dialogs. The within-subjects results displayed in Table 5.9 show a significant main effect for *Time* (F=15.967, p=.001) and a significant effect for *Time*Talent* (F=12.215, p=.003). The test of between-subjects contrasts confirms a weak significant effect for the factor *Talent* in the *J* condition (F=4.929, p<.05), indicating that the change in match values of the nonnative speakers in this condition can be at least partially attributed to the phonetic talent of the subjects.

The same repeated measures ANOVA was performed for the *T* condition (Table 5.11 and 5.12). Here, the within-subjects Table 5.11 presents a strong main effect (F=28.640, p=.000) for *Time*, and also a strong effect for the combined factor of

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Time	Sphericity assumed	.006	1	.006	28.640	.000
Time*Talent	Sphericity assumed	.006	1	.006	29.057	.000
Error (Time)	Sphericity assumed	.003	18	.000		

Table 5.11: Repeated measures ANOVA – within-subject effects for Set 1 and Set 2 (factor **Time**) in the condition *T*.

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	21.739	1	21.739	23837.221	.000
Talent	.015	1	.015	16.397	.001
Error	.016	18	.001		

Table 5.12: Repeated measures ANOVA – between-subject effects for the factor **Talent** in the condition *T*.

*Time*Talent* (F=29.057, p=.000). The between-subjects effect of *Talent* for the *T* dialogs (Table 5.12) is also significant (F=16.397, p<.01). These results confirm that there is a strong influence of the *talent* factor on the increase in match values, i.e. convergence in the dialogs in this condition. The two groups of talented and less talented subjects differ significantly in their behavior between an early and a late point in the dialogs with native speaker *T*. Figures 5.2 and 5.3 show the change in mean match values between the early and late points in the dialog (early= **Set 1**, late= **Set 2**) for each of the dialog conditions. The two talent groups are depicted as separate lines in the diagrams (the solid line for the talented speakers and the dotted line for the less talented speakers).

The evidence presented here allows us to reject the null hypothesis $A2_0$ denying that *Talent* is an influencing factor in convergence in nonnative-native dialogs. Thereby **Hypothesis A2** is **confirmed**: talented speakers indeed converge more to their partners than less talented ones.

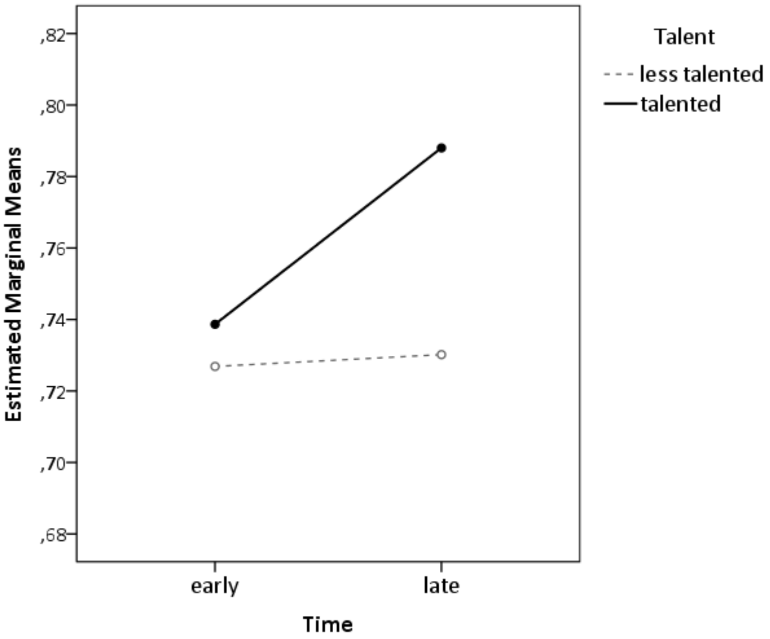


Figure 5.2: The degree of change in the match values between Set 1 (early) and Set 2 (late) of all 20 German subjects compared to native speaker *J*.

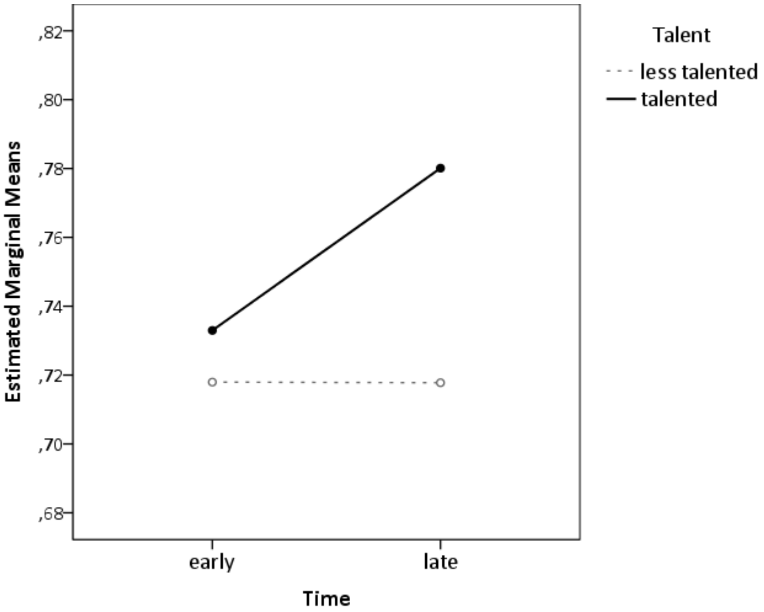


Figure 5.3: The degree of change in the match values between Set 1 (early) and Set 2 (late) of all 20 German subjects compared to native speaker *T*.

		N	Mean	SD	Std. Error Mean	Min	Max
conv_XJ	Female	10	.0216	.04348	.01375	-.08	.09
	Male	10	.0310	.03122	.00987	-.01	.10
	Total	20	.0263	.03715	.00831	-.08	.10
conv_XT	Female	10	.0215	.03773	.01193	-.02	.09
	Male	10	.0255	.02409	.00762	-.02	.05
	Total	20	.0235	.03088	.00690	-.02	.09

Table 5.13: Mean, standard deviation (SD), standard error mean, and minima and maxima for *conv_J* and *conv_T*.

		Sum of Squares	df	Mean Square	F	Sig.
conv_XJ	Between Groups	.000	1	.000	.310	.584
	Within Groups	.026	18	.001		
	Total	.026	19			
conv_XT	Between Groups	.000	1	.000	.082	.778
	Within Groups	.018	18	.001		
	Total	.018	19			

Table 5.14: The table shows the results of the one-way ANOVA analysis for the group factor *Gender*. The convergence of the nonnative speakers in condition *XT* and condition *XJ* is shown in separate rows. *X* stands for the NNS. The convergence here is calculated as the difference between Set 1 and Set 2.

Hypothesis A3 states that female speakers converge more than male speakers toward their native speaking partners. A descriptive analysis of the mean values for all female and male speakers is given in Table 5.13. Men show slightly higher mean values than women in both conditions, while women have a greater range of match values in both dialogs, pointing to greater variance. Table 5.14 displays the results of a one-way ANOVA for the amount of convergence between female and male speakers. The values *conv_XJ* and *conv_XT* were calculated as the difference between **Set 1** and **Set 2** in the two dialog conditions, and thus stand for the mean amount of convergence of the German subjects toward the native speakers. The analysis did not confirm any significant differences between the two groups in either condition (for *J*: F=.310, p=.584, for *T*: F=.082, p=.778).

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set 7J	.7989	20	.04408	.009856
	Set 6J	.8258	20	.03370	.007536
Pair 2	Set 7T	.8095	20	.04303	.009622
	Set 6T	.8488	20	.03056	.006833

Table 5.15: Mean self-consistency values of the nonnative speakers in both dialogs. The comparison of **Set 7**: early vs. late items and **Set 6**: early vs. early items of the nonnative subjects.

		Mean	**SD**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
Pair 1	Set 7J – Set 2J	.0398	.057570	.012873	3.309	19	.006
Pair 2	Set 7T – Set 2T	.0656	.055264	.012357	5.308	19	.000

Table 5.16: Paired samples t-test for the self-consistency values of the nonnative speakers (Set 7) and their convergence in the respective dialog (Set 2). The self-consistency of the subjects was significantly higher than their convergence.

The null hypothesis $A3_0$ that women and men converge to the same extent cannot be rejected and must be upheld. Therefore, **Hypothesis A3**, which states that female subjects converge more to their conversational partners than male subjects, **cannot** be **confirmed**.

5.1.2 Self-consistency of the nonnative speakers

When comparing the mean self-consistency values of the nonnative subjects' pronunciation to their convergence (Table 5.15), self-consistency with values around .80 are significantly higher than the convergence in both dialogs, indicating that the speakers retained their own pronunciation more than they accommodated to their conversational partners (p<.01 in both conditions).

A more detailed analysis of the self-consistency values taking into account the statement of **Hypothesis A4** (talented subjects will show more perturbed self-

	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Set 7J – Set 6J	Intercept	26.396	1	26.396	9024.84	.000
	Talent	7.744E-6	1	7.744E-6	.003	.960
	Error	.053	18	.003		
Set 7T – Set 6T	Intercept	27.501	1	27.501	12746.03	.000
	Talent	.001	1	.001	.449	.511
	Error	.039	18	.002		

Table 5.17: Repeated measures ANOVA analysis for the self-consistency measures of the nonnative speakers, as seen from the angle of the group factor *Talent*. Self-consistency here is calculated as the difference in match values between items from **Set 7** and **Set 6**. The group factor **talent** is non-significant for the variance the subjects display.

consistency values than less talented ones in the dialog) are given in Table 5.16 and Table 5.17. The paired samples t-test in Table 5.16 shows decreasing self-consistency values between the early and late measurement in both dialog conditions – *J* and *T*. The performed t-test confirms that the differences between both measurement points are significant (t=-4.848, sig.=.000 in condition *J*, and t=-4.727, sig.=.000 in condition *T*), indicating that the match values of the target word utterances decreased significantly when we compare their late and early versions within the dialog. The between-subjects effects for the group factor **talent** are reported in Table 5.17. As shown, there is no significant difference between the two talent groups concerning the change of their self-consistency values throughout the dialog ($p>.50$ in both conditions).

Figure 5.4 and 5.5 show the self-consistency values of both talent groups in the dialogs. The results of the repeated measures ANOVA in Table 5.17 could not prove any significant differences between the conditions, therefore the null hypothesis $A4_0$ has to be upheld. **Hypothesis A4**, which states that talented subjects should

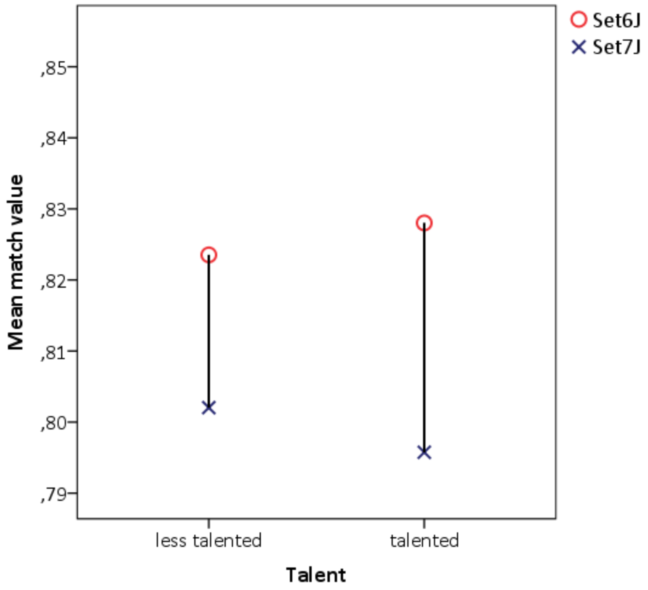


Figure 5.4: Mean match values of **Set 7J** and **Set 6J** of all 20 German subjects according to the group factor *Talent*. The differences between the talent groups are not significant.

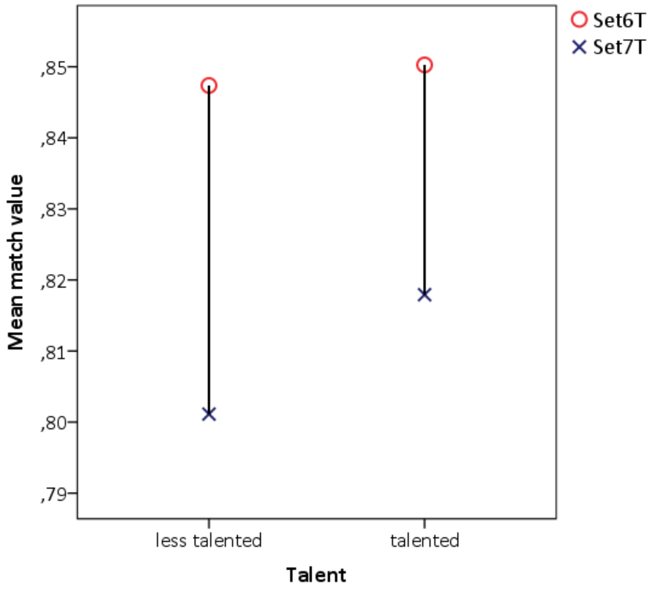


Figure 5.5: Mean match values of **Set 7T** and **Set 6T** of all 20 German subjects according to the group factor *Talent*. The differences between the talent groups are not significant.

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set 3J	.7516	20	.03102	.006937
	Set 5J	.7110	20	.03125	.006987
Pair 2	Set 3T	.7700	20	.03477	.007775
	Set 5T	.7218	20	.04920	.011001

Table 5.18: Mean, standard deviation (SD) and standard error mean for the comparison of **Set 3** with late items of both speakers and **Set 5** in both dialogs – containing summary items of the NNS and late items of the NS. All values are calculated for the whole set of N=20 subjects.

show lower[6] self-consistency values (caused by their higher convergence) can**not** be confirmed.

5.1.3 Speaking style switch from dialog to monologue

After the dialog task was completed, every nonnative participant was instructed to summarize the identified differences in the task. This took place with the native speaker still present but not actively participating in the conversation. **Set 5** was calculated as a comparison of items coming from this summary part of the NNS and those coming from a *late* point in the dialog. This is put into relation with the match values obtained in **Set 3**, which contains *late* items from both conversational partners, representing the mutual balance level obtained at a late point in the dialogs. This allows us to test the statement in **Hypothesis A5**, as to whether the level of adaptation to the conversational partner persists even after a switch from dialog to monologue style has taken place.

Table 5.18 shows the mean values for **Set 3** and **Set 5** in both dialog conditions. The values for the summary matches (**Set 5J** and **Set 5T**) are roughly .04 to .05 *lower* than the values for the late matches (Set 3J and Set 3T). A paired samples test confirms that the match values from both sets differ significantly from one another (t=6.009, p=.000 for *J*, and t=4.872, p=.000 for *T*), i.e. that the convergence of

[6] The lower the self-consistency values, the more perturbed the speakers' own way of speaking is.

		Mean	SD	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	Set 3J – Set 5J	.0406	.03023	.006759	6.009	19	.000
Pair 2	Set 3T – Set 5T	.0482	.04420	.009884	4.872	19	.000

Table 5.19: Paired samples t-test for **Set 3** and **Set 5** in both dialog conditions. The difference between the dialog and summary match values is highly significant in both cases.

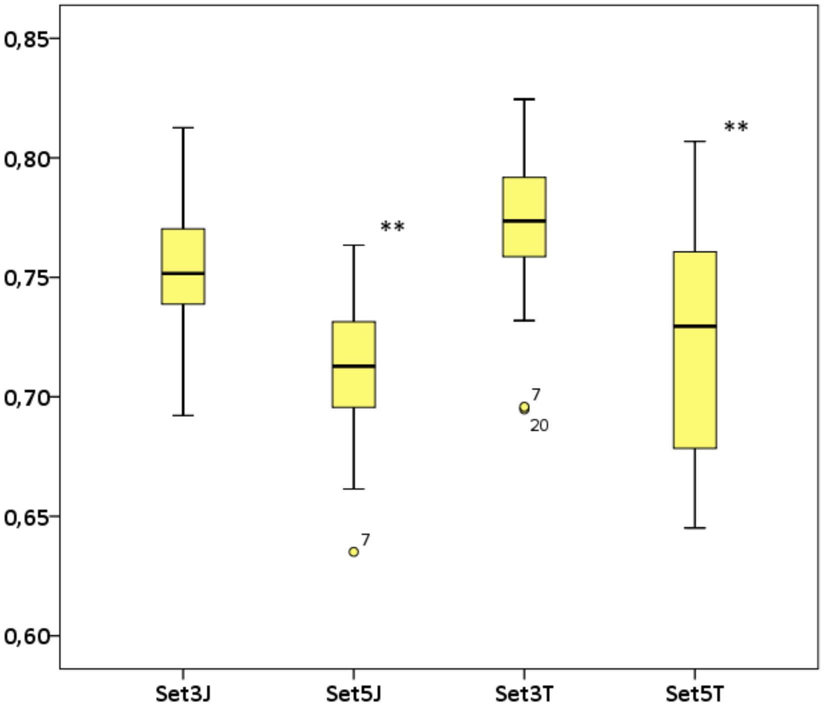


Figure 5.6: Mean match values of **Set 3J**, **Set 5J**, **Set 3T** and **Set 5T** of all 20 German subjects. The differences between late-late match values and the summary-late match values are significant (indicated with a double asterisk).

the nonnative speakers toward the native speaker decreased when they produced the summary at the end of the dialog session (Table 5.19). The boxplots in Figure 5.6 represent each of the tested sets in the *J* and *T* conditions separately. The significant drops in match values are marked with a double asterisk above the relevant box in the diagram.

Figures 5.7 and 5.8 show the change in mean match values between the non-

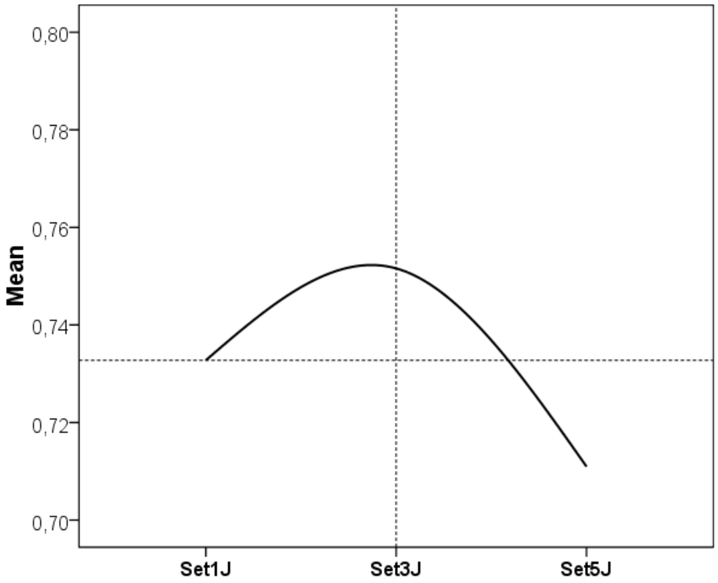


Figure 5.7: Fitted curve of the change of mean match values in dialog **J**. Measurements taken for **Set 1J** (early-early), **Set 3J** (late-late) and **Set 5J** (summary-late) including all 20 German subjects. The difference between **Set 3J** and **Set 5J** is significant.

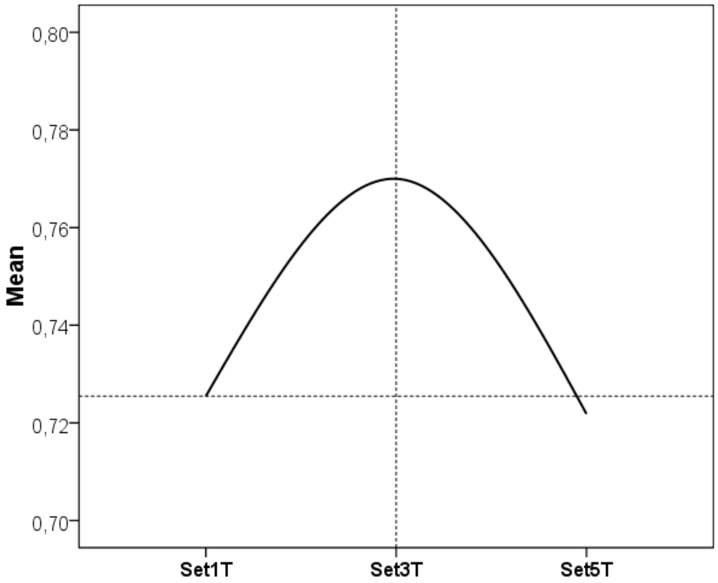


Figure 5.8: Fitted curve of the change of mean match values in dialog **T**. Measurements taken for **Set 1T** (early-early), **Set 3T** (late-late) and **Set 5T** (summary-late) including all 20 German subjects. The difference between **Set 3T** and **Set 5T** is significant.

native speakers and the native speakers starting from *early*, through *late* and to the *summary* part of the dialog. The values rise from early to late, and thereafter decrease to a relatively low level that is comparable to or below the starting point. The difference between the early (**Set 1**) and the late (**Set 3**) measurement are statistically significant, as was stated earlier. The difference between (**Set 1**) and (**Set 5**) is only significant for the dialogs of native speaker *J* (t=-2.687, p=.015, Figure 5.7), indicating that the values dropped even below the initially measured level of the early match values. In the case of native speaker *T*, the mean match values settled around .72 for both the early and the summary measurements and the difference here is not significant (t=-.451, p=.657, Figure 5.8).

The presented analysis results do not allow us to reject the null hypothesis $A5_0$, as there is a significant difference between **Set 3** and **Set 5** in both dialog conditions. **Hypothesis 5**, which states that the level of adaptation to the conversational partner persists after the switch from dialog to narrative has taken place, must therefore be **rejected**.

The next section is concerned with the analyses of the native speakers' behavior in the dialogs: the adaptation toward the nonnative speakers and the possible interactions of the convergence levels of both dialog partners.

5.2 Results of native speaker performance

After having analyzed the behavior of the nonnative speakers in the dialogs, we turn now to the data of their conversational partners and compare the native speakers' match values at the beginning and end points of the dialogs. The analyses presented here will serve to test **Hypothesis A6** and determine whether the native speakers did in fact converge to their German dialog partners, despite being told to maintain their own speaking style. In addition, **Hypothesis A7**, which stated that both interactants' convergence contributes to the establishment of a mutual *balance level* at a late point in the dialog, will also be examined.

5.2.1 Convergence towards nonnative speakers

The descriptive statistics in Table 5.20 present the means for **Set 4** (the match values of *late* items of the native speakers along with *early* items of the nonnative speakers) and **Set 1** (the comparison of both *early* items). Calculating the difference between these two sets will lead to an estimate of the native speakers' convergence toward the NNS in the dialogs.

The paired samples test in Table 5.21 shows a difference of means of .020 for native speaker *J* and .045 for native speaker *T* across all dialogs. This difference

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	Set 4J	.7531	20	.02317	.005181
	Set 1J	.7328	20	.03468	.007755
Pair 2	Set 4T	.7714	20	.04660	.010420
	Set 1T	.7255	20	.02364	.005286

Table 5.20: Mean, standard deviation (SD) and standard error mean for the comparison of **Set 4** with late native speaker items vs. early NNS items, and **Set 1** with early items of both speakers, in order to determine the amount of convergence of the native speakers. All values are calculated for the whole set of N=20 subjects.

		Mean	**SD**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
Pair 1	Set 4J – Set 1J	.0203	.02867	.006410	3.166	19	.005
Pair 2	Set 4T – Set 1T	.0459	.04727	.010569	4.342	19	.000

Table 5.21: Paired samples t-test for **Set 4** and **Set 1** in both dialog conditions. The difference in match values for the native speakers between an early and late point in the dialogs is significant in both cases.

is statistically significant at a level of p<.01 for both NS, indicating that the native speakers converged to their conversational partners as well.

The differences for convergence toward *talented* or *less talented* speakers were not significant[7]. There was a tendency, however, in the direction of less adaptation towards more talented speakers, as shown in Figure 5.9 and Figure 5.10. Less convergence by the nonnative subjects was often met by more convergence on behalf of the native speaker, though the relationship is not statistically significant[8].

Figure 5.11 presents box plots of the mean convergence values for both native speakers. Native speaker *T* produced on average more convergence than native speaker *J* but also showed considerably higher standard deviations from the mean value (compare Table 5.21).

In order to define the possible interactions of the native speakers' convergence with other factors, the following correlations were tested: **conv_J, conv_T, conv_XJ, conv_XT, gender, talent, Set 1J** and **Set 1T**. Only one comparison yielded a statistically significant result: the inverse correlation of native speaker *J*'s convergence and the match values from **Set 1J** (Table 5.22). In other words, the amount of convergence of native speaker *J* is inversely tied to the height of the match values obtained from the comparison of *early–early* items in **Set 1J**. The higher the match with the nonnative speaker at the beginning of the dialog, the lower the convergence of NS *J* at a late point in the dialog. Neither the factor *talent* nor *gender*

[7] Repeated measurements ANOVA for the effect *Time*Talent*: F=2.067, p>.05 for NS *J*; F=1.153, p>.05 for NS *T*.

[8] More details on the matter of a mutual balance level are presented in Chapter 5.2.2.

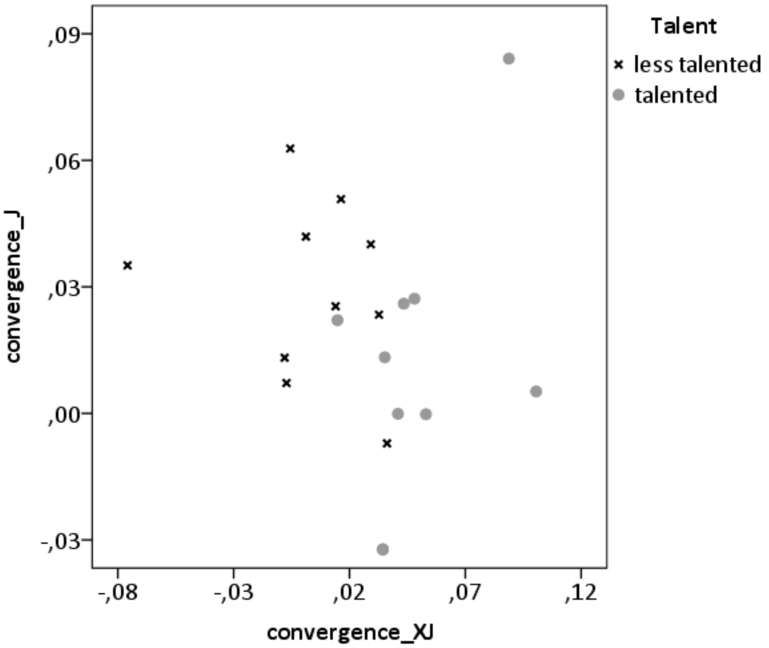


Figure 5.9: The mean match values of **convergence_J**, calculated as the difference of **Set 4J**, **Set 1J**, and **convergence_XJ** of all 20 German subjects.

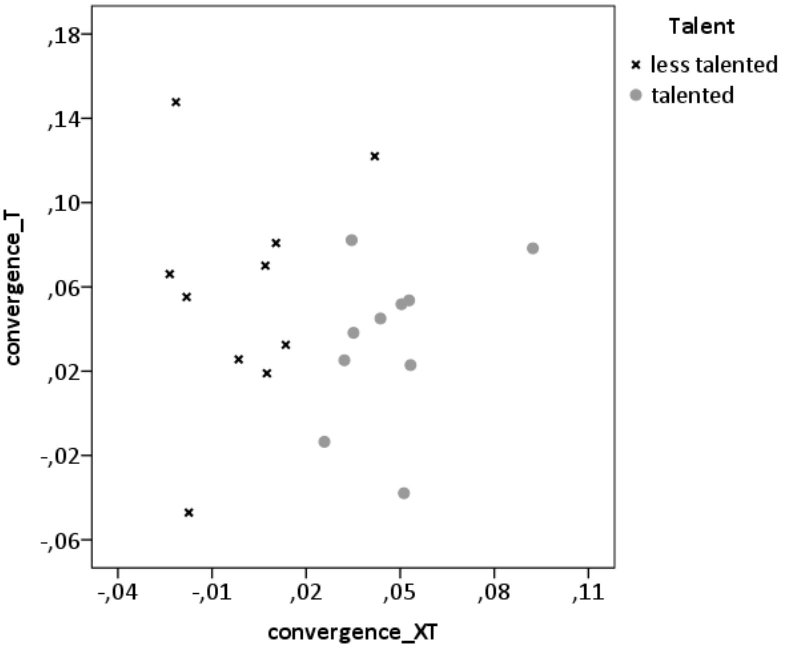


Figure 5.10: The mean match values of **convergence_T**, calculated as the difference of **Set 4T**, **Set 1T**, and **convergence_XT** of all 20 German subjects.

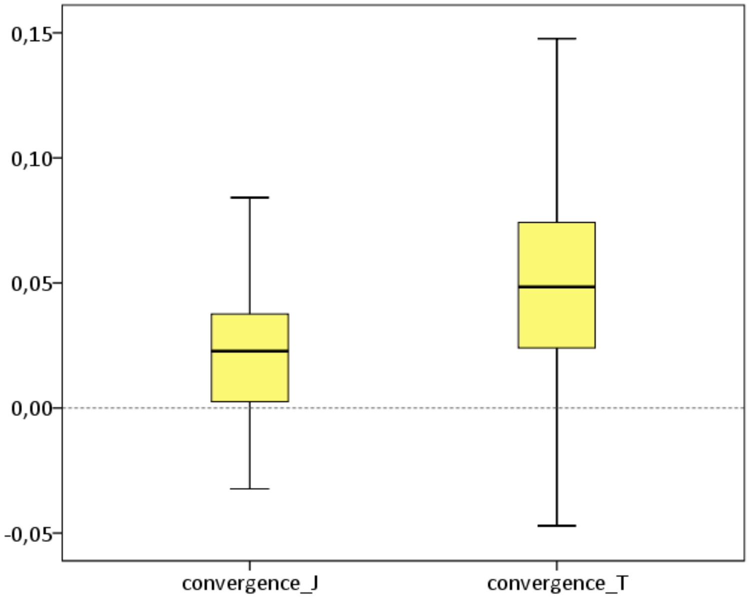


Figure 5.11: Box plots representing the mean match values of **convergence_J** and **convergence_T**, calculated as the difference of **Set 4T**, **Set 1T** for all 20 nonnative subjects.

		conv_XJ	**conv_XT**	**Gender**	**Talent**	**Set 1J**	**Set 1T**
conv_J	Pearson correlation	-.097	–	-.358	-.322	- .748**	–
	Sig. (2-tailed)	.658	–	.121	.167	.000	–
conv_T	Pearson correlation	–	-.033	-.097	-.245	–	-.278
	Sig. (2-tailed)	–	.890	.684	.297	–	.235

Table 5.22: The correlation between the native speakers' convergence and multiple factors. The double asterisk ** indicates a correlation which is significant at the 0.01 level (2-tailed).

have a comparable influence in the case of this native speaker. For native speaker *T* neither of the comparisons produced a significant correlation.

The results of the paired samples test presented in Table 5.20 and Table 5.21 allow us to reject the null hypothesis $A6_0$ that the native speakers did not converge. Due to the rejection of the null hypothesis, **Hypothesis A6** can be **accepted**. The native speakers did converge to their nonnative conversational partners between an early and a late point in the dialogs.

5.2.2 Mutual level of balance in convergence

It has already be shown that both nonnative and native speakers alter their pronunciation, with the respective interactant's pronunciation as the probable target. The convergence of the nonnative speakers (*convergence_XJ* and *convergence_XT*) was calculated as the difference between **Set 2** and **Set 1**; the convergence of the native speakers (*convergence_J* and *convergence_T*) as the difference between **Set 4** and **Set 1**. The calculated difference between **Set 3** and **Set 1** is a measure of the **magnitude of achieved balance**, indicating the direction (positive vs. negative values) and the degree (low or high values) of the mutually reached balance level.

A natural consequence of both interactants converging or diverging would be a changed level of balance at a late point in the dialogs compared to the early point, represented by the match values in **Set 1**. This late level of convergence can be calculated as the comparison of the following two sets[9]:

- **Set 3** – late vs. late items
- **Set 1** – early vs. early items

Table 5.23 shows the means and standard deviations of the values in **Set 1** and **Set 3** for both dialogs and all speakers. The mean values for the late comparison in **Set 3** are generally higher than the early dialog comparison contained in **Set 3**. The values of both sets are also positively correlated (Table 5.24). In the case

[9] The comparison was carried out for both dialog conditions *J* and *T* separately.

		Mean	N	SD	Std. Error Mean
Pair 1	Set 3J	.7516	20	.03102	.006937
	Set 1J	.7328	20	.03468	.007755
Pair 2	Set 3T	.7670	20	.03477	.007775
	Set 1T	.7255	20	.02364	.005286

Table 5.23: Mean, standard deviation (SD) and standard error mean for the comparison of **Set 3** and **Set 1** for matches with native speaker *J* and native speaker *T*. Set 1 indicates the mean match value of early items of the NNS and NS, while Set 3 signifies the mean match values of both late items. All values are calculated for the complete set of German subjects - 20 speakers.

		N	Correlation	Significance
Pair 1	Set 3J & Set 1J	20	.777	.000
Pair 2	Set 3T & Set 1T	20	.443	.050

Table 5.24: The correlation between the paired match values for Set 3 and Set 1.

		Mean	SD	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	Set 3J – Set 1J	.0188	.02220	.004963	3.793	19	.001
Pair 2	Set 3T – Set 1T	.0445	.03223	.007207	6.178	19	.000

Table 5.25: Paired samples t-test between **Set 3** and **Set 1** in dialog *J* and dialog *T*. Both differences are highly significant.

of condition *J* the correlation is rather strong (.777 at the sig. level of p=.000), while in the case of dialog *T* the correlation has only minor strength at a lower significance level (.443 with a p=.05).

The paired samples t-test for condition *J* and condition *T* in Table 5.25 confirms what the descriptive statistics suggest: both comparisons are highly significant (t=3.793, p=.001 for dialog *J*; and t=6.178, p=.000 for dialog *T*). The difference in match values between the two sets is significant. The speakers did reach a higher level of mutual convergence – a balance level, which was the result of both participants' behavior.

A separate analysis of the two talent groups shows significant increases between **Set 1** and **Set 3** in both conditions for the *talented* group (t=4.124, p=.003 in dialog *J*; and t=9.216, p=.000 in dialog *T*), but only in the *T* condition for the *less talented* group (t=2.815, p=.02), and only a non-significant increase in the *J* condition (t=1.561, p=.153). The *talented* speakers showed, on average, higher values for the balance level, though the difference is not significant.

Hypothesis A7 states that the speakers reach a higher level of convergence at the end of the dialog than at the beginning. The null hypothesis $A7_0$ that the constant balance level will not differ between an early and a late point of the dialog was *rejected* by the t-tests performed (Table 5.25). This allows us to **accept Hypothesis A7**: the conversational partners reached an overall higher "balance" level towards the end of the experimental task, with the exception of the *less talented* speakers in dialog *J*.

5.3 Summary of dialog results

The results of the dialog data analyses presented in Chapter 5.1 and Chapter 5.2 have allowed us to accept **Hypothesis A1, Hypothesis A2, Hypothesis A6** and **Hypothesis A7**, while forcing a rejection of **Hypothesis A3, Hypothesis A4** and **Hypothesis A5**. The main claims of this thesis that convergence can be observed in native-nonnative conversations, is bi-directional, and is, a.o., contingent on the individual difference of *talent* in the nonnative speakers, could be confirmed (**Hypothesis A1, Hypothesis A2, Hypothesis A6**). Nonnative speakers generally do converge to their native speaking dialog partners, and, at the same time the native speakers converge to them. This behavior is, along with other things, mediated by the factor *talent*: talented speakers do indeed converge more to their partners than less talented ones, who sometimes fail to converge and rather diverge instead. The native speakers, despite being told to maintain their own speaking style as far as possible, also converged.

Hypothesis A3, which stated that female subjects converge more to their conversational partners than male subjects, was **not confirmed**. The differences in the amount of convergence (higher for male subjects) for the factor *gender* did not reach a significant level. **Hypothesis A4**, arguing that talented subjects should show lower self-consistency values (caused by their higher convergence), could **not** be confirmed either. Both groups showed perturbed self-consistency values, indicating that their pronunciation was altered. The between-group differences, however, did not reach statistical significance.

The analysis of the persistence of the convergence effect revealed that the adaptation to the conversational partner did not remain at the same level after the switch from dialog to narrative. This led to the rejection of **Hypothesis A5**.

Hypothesis A7, claiming a mutually achieved balance level, could be **confirmed**. The conversational partners reached an overall higher level of convergence, which balanced towards the end of the experimental task. The *convergence balance* level was higher for the *talented* group in both dialogs, but the difference did not reach significance.

The second results section in Chapter 6 will be concerned with the analyses of the read speech data from both pre- and post-tests and their relation to the dialog data presented here.

Chapter 6

Results: read speech

The following section describes the results of the pre- and post-test sessions in which subjects were reading word lists out loud. As described in detail in Chapter 4, the word lists preceded and followed both dialogs with the native speakers. It is therefore possible to test whether the dialog situation had any influence on the way words were read. In order to obtain a full range of comparisons, the following measurements were taken (see also Figure D.1 in Appendix D):

- match values of the nonnative subjects for lists 1 and 2 and native speaker T[1]
- match values of the nonnative subjects for lists 2 and 3[2] and native speaker J
- self-consistency[3] of the German subjects between lists 1, 2 and 3.

The following hypotheses were formulated regarding the read speech data:

1. **Hypothesis B1**
   Subjects show convergence to the native speaker in the read speech task following the dialog.
2. **Hypothesis B2**
   Convergence in read speech is positively correlated with convergence in dialog speech.
3. **Hypothesis B3**
   Talented and non-talented subjects behave differently in the read speech task.
4. **Hypothesis B4**
   Female and male subjects behave differently in the read speech task.

[1] Both native speakers read the word list only once before the whole experimental session started, since their adaptation or the lack of it were not the main research question. Therefore only the results for the nonnative speakers will be presented in this chapter.

[2] In cases where the subject was first talking to native speaker *J* and then to NS *T*, the order of the comparisons has been reversed accordingly.

[3] i.e., the measurement indicating how true to their own *read* speaking style the subjects stayed in the course of the whole experimental session, measured as the comparison between word lists 1, 2 and 3.

5. **Hypothesis B5**
   Talented subjects will show more perturbed self-consistency values than less talented ones.

Further commentary on the hypotheses and the respective results of the analyses will be given in sections 6.1 and 6.2.

6.1 Nonnative speakers' convergence in the reading task

The following section describes all results obtained for comparisons between the German subjects and the English native speakers *T* and *J*. Hypotheses **B1** to **B4** predict that the nonnative speakers will accommodate in a certain way to their native speaking partners, even after the dialog. Moreover, they put forward the claim that this adaptation is dependent on such factors as:

- magnitude and direction of adaptation in the dialog,
- talent group membership,
- gender

This can be measured by comparing the match values of the subjects and native speakers for the pre- and the post-tests: the word list tasks. The analysis will start by testing **Hypothesis B1**, which provides the most important evidence for convergence in read speech.

6.1.1 Convergence in read speech

Hypothesis B1 states that **subjects show convergence to the native speaker in the read speech task following the dialog**.
If **Hypothesis B1** is true, there should be a significant change in the match value between the subjects' items from the word list preceding and immediately following

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	List 1-T	.7451	20	.02780	.006216
	List 2-T	.7472	20	.02883	.006447
Pair 2	List 2b-J	.7739	20	.02691	.006019
	List 3-J	.7700	20	.02856	.006387

Table 6.1: Mean, standard deviation (SD) and standard error mean for **Pair 1**: the comparison of List 1 and List 2 for matches with native speaker *T*, and **Pair 2**: the comparison of List 2b and List 3 for matches with native speaker *J*. The two conditions in every pair indicate the word lists before and immediately after the dialog with the respective native speaker. All values have been calculated for the complete set of German subjects (20 speakers).

the dialog with the respective NS partner. The word lists preceding and following the dialog with native speaker *T* are coded as *List 1* and *List 2,* the word lists preceding and following native speaker *J* are coded as *List 2b* and *List 3*, respectively. In order to test **Hypothesis B1** a paired samples test was conducted. Results of the analyses are given in Tables 6.1, 6.2 and 6.3.

According to the statistics provided in Table 6.1, the mean match value for the pairing **German subject and native speaker *T*** was .745 in the baseline measurement (List 1) and .747 in the second measurement (List 2). The mean match value for the second pairing **German subject and native speaker *J*** was .774 in the baseline measurement (List 2b) and .770 in the second measurement (List 3). All match results lie within a mid to mid-high range of goodness[4].

Table 6.2 displays the correlation between the paired measurements in relation to native speaker *T* and *J*. The correlation between the baseline measurement and measurement for List 2 for NS *T* equals .773 at a significance level of .000, while the same correlation for NS *J* is even stronger and amounts to .821, also at a significance level of .000.

[4] The match values range from 0 to 1, where 1 indicates a perfect match and 0 no match. Practice showed that pairs of identical words pronounced by different speakers usually do not obtain match values lower than 0.5 on the scale. Thus values around .75 can be considered average- to high-average matches.

		N	Correlation	Significance
Pair 1	baseline & after T (List 1-2)	20	.773	.000
Pair 2	baseline & after J (List 2b-3)	20	.821	.000

Table 6.2: The correlation between the paired match values for the baseline list and the word list following the dialog for both native speakers.

		Mean	SD	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	baseline & after T (List 1-2)	-.0021	.01911	.004273	-.494	19	.627
Pair 2	baseline & after J (List 2b-3)	.0038	.01669	.003732	1.027	19	.317

Table 6.3: Paired samples t-test for the word list before and after the dialog with *T* (Pair 1), and before and after the dialog with *J* (Pair 2). The significance level of the changes is given in the last column.

Table 6.3 shows the details of the paired samples t-test for the relevant read speech lists. The changes in match values between the baseline (List 1) and the second list in the case of native speaker *T* do not reach significance (sig. 2-tailed = .627). The same is true for the changes in match values between the baseline (here, List 2b) and the following measurement (List 3) for native speaker *J* – significance here was not reached either (sig. 2-tailed = .317). Therefore **Hypothesis B1**, that subjects show convergence to the native speaker in the read speech task following the dialog, is **rejected**.

Figure 6.1 and Figure 6.2 show a graph comparing the change in match values between all German subjects and the respective native speakers. The German subjects are displayed in alphabetical order, without any grouping according to talent and/or gender. It is clear that the occurring changes are small in magnitude, which is supported by the comparison of means given in Table 6.1. The changes do also occur bi-directionally – indicating movements towards both better and worse matches between the pronunciation of the word lists of NNS and NS compared

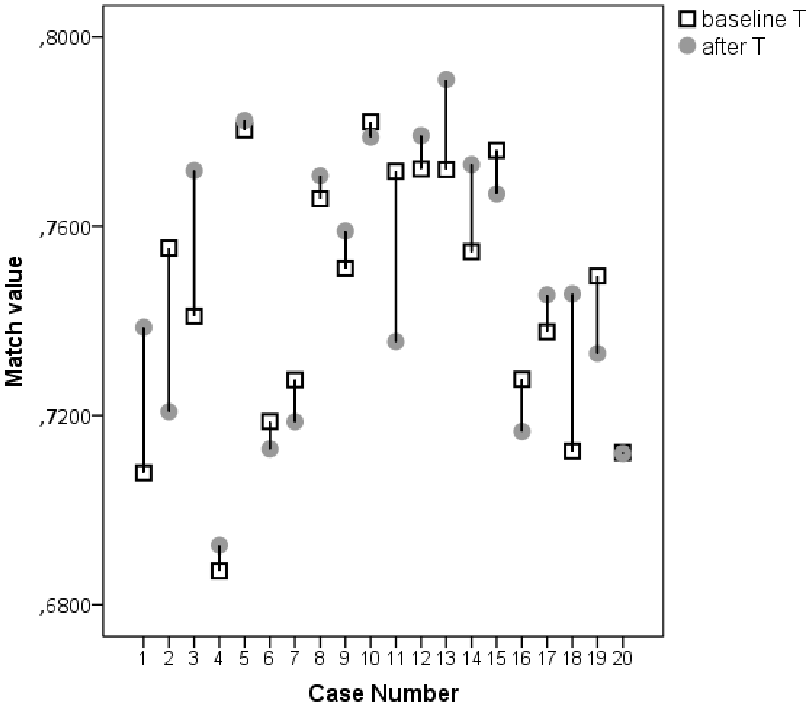


Figure 6.1: The magnitude of change in the match values of all 20 German subjects compared to native speaker *T* as measured in the word lists before (Baseline) and after the dialog with NS *T*.

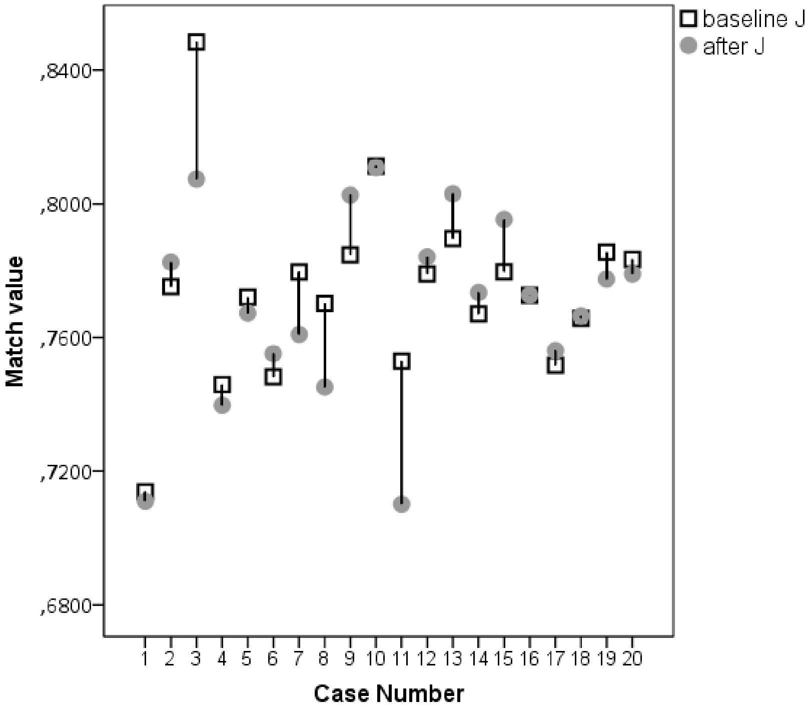


Figure 6.2: The magnitude of change in the match values of all 20 German subjects compared to native speaker *J* as measured in the word lists before (Baseline) and after the dialog with NS *J*.

to the baseline before the dialog. Those slight drifts, however, are not significant overall, as shown by the t-test reported in Table 6.3.

6.1.2 Correlation between dialog and read speech convergence

It seems plausible to assume that the degree of convergence observed during the native-nonnative interactions is correlated to the convergence in read speech, measured as the difference before and after the respective dialog. **Hypothesis B2** thus states that convergence in read speech is positively correlated to convergence in dialog speech.

Table 6.4 displays the values for the Pearson correlation between the dialog and read speech convergence. None of the conditions shows a correlation between the degree of convergence in the dialog to the results in the read speech task. The correlation values are very low, .141 for the *J* condition and .106 for the *T* condition, with significance levels of above .50 for both. Therefore **Hypothesis B2** can be rejected – even though the subjects showed significant convergence beforehand, their performance in the read speech task is **not correlated** to their performance in the dialog. Figure 6.3 and 6.4 are a graphical representation of the relation between the convergence measured in the dialogs and the read task, separately for both conditions – *J* and *T*. The results in Table 6.4 show that the values do not line up, i.e., they do not show a positive or negative correlation.

		N	**Correlation**	**Significance**
Pair 1	convergence XJ & convergence read XJ	20	.141	.553
Pair 2	convergence XT & convergence read XT	20	.132	.579

Table 6.4: The Pearson correlation between convergence in the dialog and in the read speech task of the nonnative toward the native speakers. Convergence XJ and XT are calculated as the difference between measurement *point 1* and *point 2* (compare Chapter 5), convergence read XJ and XT are the difference between, respectively word list 3 and 2b, and word list 2 and 1.

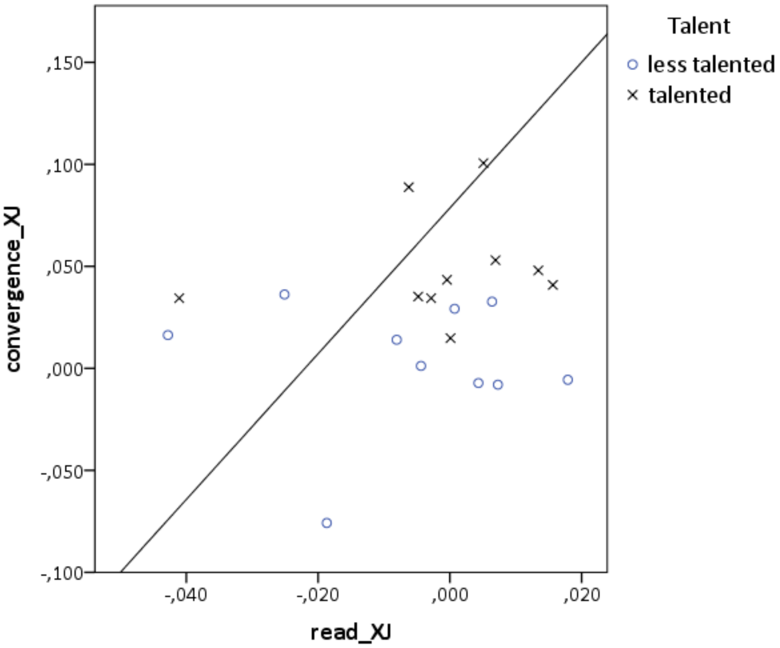


Figure 6.3: A scatterplot of the convergence values in the *J* condition, with the dialog values displayed on the Y axis and the read values on the X axis.

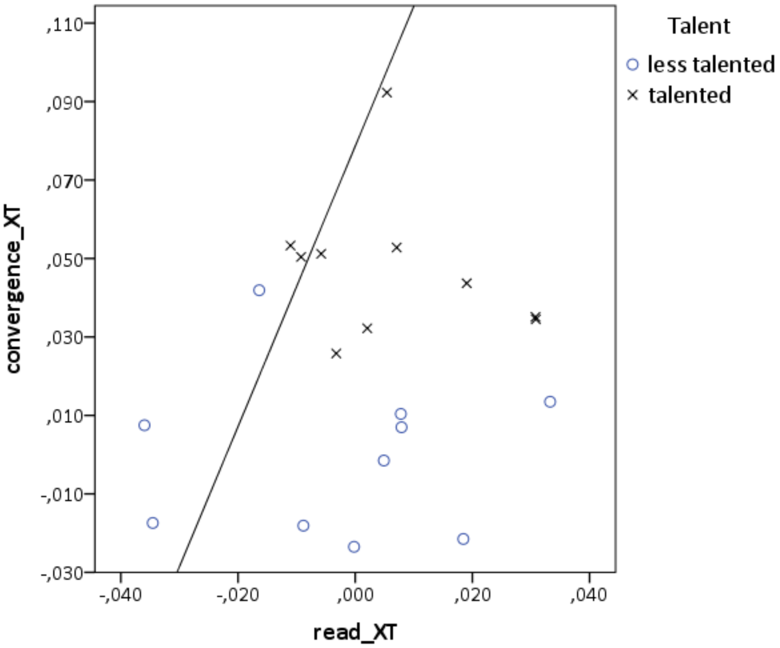


Figure 6.4: A scatterplot of the convergence values in the *T* condition, with the dialog values displayed on the Y axis and the read values on the X axis.

6.1.3 Talent in read speech

Going into more detail concerning the read speech data, two additional group comparisons were conducted: one for the factor *talent* (covering Hypothesis **B3**) and one for the factor *gender* (covering Hypothesis **B4**). Hypothesis **B3** states that talented and non-talented subjects behave differently in the read speech task.

In order to get a first overview, a separate analysis for the two subsets *talent* vs. less talented was conducted, that is analogous to the analyses for the whole data set presented in Table 6.1, Table 6.2 and Table 6.3. If **Hypothesis B3** is true, the following ANOVA should show significant differences between the two talent groups for the read speech data. Then the talented and less talented speakers should display differing patterns of convergence (or the lack thereof) in the read speech task. In the case that the ANOVA does not show significant differences between the two groups, **Hypothesis B3** must be rejected and the factor *talent* discarded as bearing any influence on the outcome of the read speech task.

The following analyses were performed on the two subsets *less talented* and *talented* of the whole experimental group. Tables 6.5, 6.6 and 6.7 show the results for the former subset, and Tables 6.8, 6.9 and 6.10 for the latter.

As visible from Table 6.5, the subset of *less talented* subjects showed very similar means for the two crucial measurements before and after the dialogs. The cor-

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	List 1–T	.7438	10	.020766	.006567
	List 2-T	.7414	10	.021384	.006763
Pair 2	List 2b–J	.7716	10	.012403	.003922
	List 3-J	.7654	10	.025055	.007923

Table 6.5: Mean, standard deviation (SD) and standard error mean for **Pair 1**: the comparison of List 1 and List 2 for matches with native speaker *T*, and **Pair 2**: the comparison of List 2b and List 3 for matches with native speaker *J*. The two conditions in every pair indicate the word lists before and immediately after the dialog with the respective native speaker. All values are calculated for the subset *less talented* only, N=10.

		N	Correlation	Significance
Pair 1	baseline & after T (List 1-2)	10	.453	.188
Pair 2	baseline & after J (List 2b-3)	10	.732	.016

Table 6.6: The correlation between the paired match values for the baseline list and the word list following the dialog for both native speakers – for the subset *less talented* only.

		Mean	SD	Std. Error Mean	t	df	Sig. (2-tailed)
Pair 1	baseline & after T (List 1–2)	.0024	.022048	.006972	.339	9	.742
Pair 2	baseline & after J (List 2b–3)	.0062	.018072	.005715	1.092	9	.303

Table 6.7: Paired samples t-test for the word list before and after the dialog with *T* (Pair 1) and before and after the dialog with *J* (Pair 2) for the subset *less talented*. The significance level of the changes is given in the last column.

relation between the two sets of values is not significant in the *T* condition and amounts to .732 in the *J* condition, at a significance level of slightly above .01 (see Table 6.6).

A direct comparison of the two measurements in a paired samples t-test for both conditions (*T* and *J*) does not indicate any significant changes in the match values for the *less talented* subset. The significance of the changes lies at .742 in the *T* condition, and at .303 in the *J* condition.

The exact same analyses were also carried out for the subset of *talented* subjects. A slightly bigger change of means in the convergence direction can be observed in the *T* condition, while the change in the *J* condition for this subset remains marginal (see Table 6.8). The correlation between the data points here is very strong and highly significant – around .90 at a significance of .000 in both conditions (Table 6.9). Although the increase in match values in the *T* condition for the *talented* subset seemed to be higher than for the *less talented* group, results for both conditions are

		Mean	**N**	**SD**	**Std. Error Mean**
Pair 1	List 1-T	.7469	10	.03458	.010936
	List 2-T	.7761	10	.03496	.011054
Pair 2	List 2b-J	.7716	10	.03694	.011680
	List 3-J	.7747	10	.03236	.010232

Table 6.8: Mean, standard deviation (SD) and standard error mean for **Pair 1**: the comparison of List 1 and List 2 for matches with native speaker *T*, and **Pair 2**: the comparison of List 2b and List 3 for matches with native speaker *J*. The two conditions in every pair indicate the word lists before and immediately after the dialog with the respective native speaker. All values are calculated for the subset *talented* only, N=10.

		N	**Correlation**	**Significance**
Pair 1	baseline & after T (List 1–2)	10	.901	.000
Pair 2	baseline & after J (List 2b–3)	10	.905	.000

Table 6.9: The correlation between the paired match values for the baseline list and the word list following the dialog for both native speakers – for the subset *talented* only.

		Mean	**SD**	**Std. Error Mean**	**t**	**df**	**Sig. (2-tailed)**
Pair 1	baseline & after T (List 1–2)	-.0066	.015495	.004899	-1.344	9	.212
Pair 2	baseline & after J (List 2b–3)	.0014	.015762	.012700	.286	9	.782

Table 6.10: Paired samples t-test for the word list before and after the dialog with *T* (Pair 1) and before and after the dialog with *J* (Pair 2) for the subset *talented*. The significance level of the changes is given in the last column.

also non-significant here (see Table 6.10.).

By taking the general talent *TalentG* as the group factor, a one-way ANOVA was calculated for the changes that occurred between the pre- and the post-test in both conditions, *T* and *J*. The new variable *conv_T* is the difference in match values of

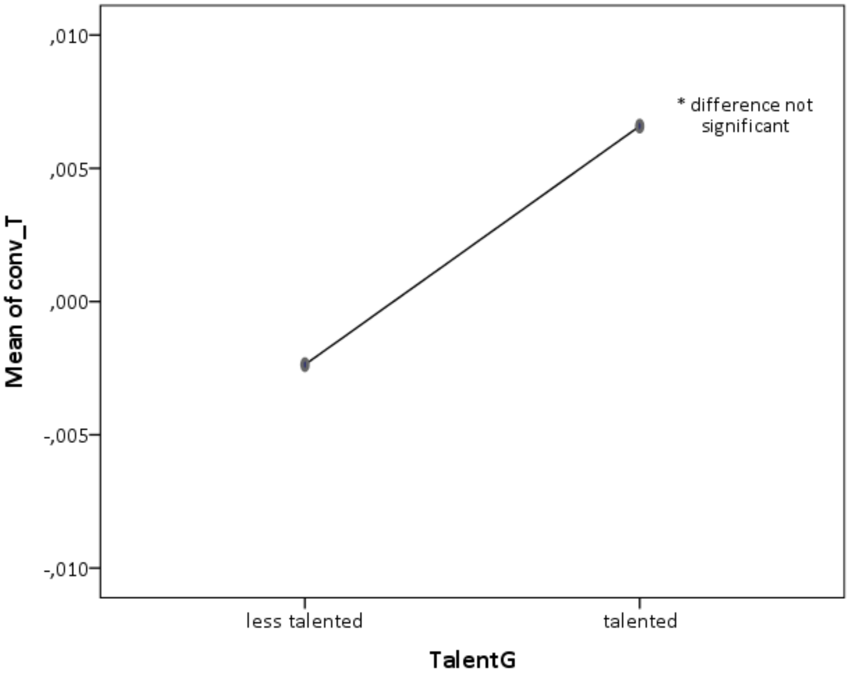


Figure 6.5: The magnitude of change in the match values of the 10 talented and 10 less talented German subjects compared to native speaker *T* (*conv_T*).

List 2 and *List 1*, while *conv_J* was calculated as the difference in match values of *List 3* and *List 2b*. Figure 6.5 and 6.6 show the plots of the mean values for the two groups *talented* and *less talented*.

Condition *T* is shown in Figure 6.5, with the two talent[5] groups displayed on the X-axis and the mean magnitude of change in match values – mean of *conv_T* – on the Y-axis. Convergence to *T* in the *less talented* group is below zero, indicating that the match after the dialog was even less strong than before, and if anything, we observe a minimal divergence. However, as is evident from Table 6.7, the magnitude of change was not significant in this group. The *talented* subset, on the other hand, shows a slight positive shift here, though it is not significant either (see Table 6.10). The difference between the two groups in the *T* condition, as evidenced by the one-way ANOVA (see Table 6.12), does not reach significance.

Condition *J* is shown in Figure 6.6, with the mean of *conv_J* displayed on the

[5] *TalentG* stands for "Talent Group".

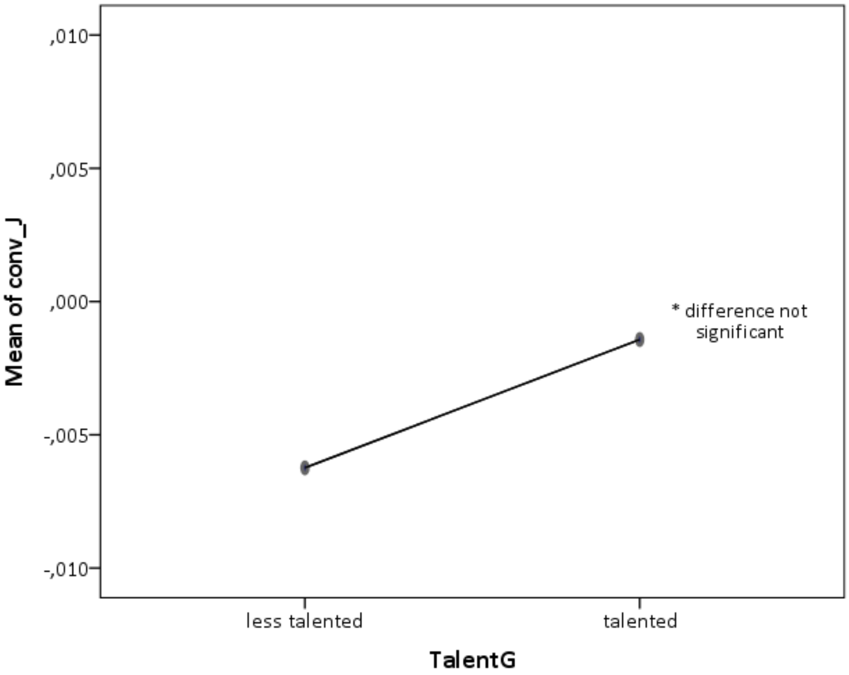


Figure 6.6: The magnitude of change in the match values of the 10 talented and 10 less talented German subjects compared to native speaker *J* (*conv_J*).

	Levene Statistic	df1	df2	Sig.
conv_T	1.094	1	18	.310
conv_J	.664	1	18	.426

Table 6.11: Results of the Levene Test for Homogeneity of Variance. Both results stay above the crucial significance level, indicating that the variance of the variables is indeed comparable and calculating an ANOVA is therefore legitimate.

Y-axis. The value of *conv_J* for both groups here is slightly negative, indicating that the match for the second list after the dialog was worse than before the dialog. As already described (see Table 6.7 and Table 6.10 for details), the shifts here are also not significant. Just as with the *T* condition, the one-way ANOVA did not show a significant difference for the group factor *talent* in this condition either (Table 6.12).

The Levene statistic presented in Table 6.11 shows homogeneity of variance for our two conditions (significance twice above .05), so an ANOVA for the group com-

		Sum of Squares	df	Mean Square	F	Sig.
conv_T	Between Groups	.000	1	.000	1.103	.307
	Within Groups	.007	18	.000		
	Total	.007	19			
conv_J	Between Groups	.000	1	.000	.403	.533
	Within Groups	.005	18	.000		
	Total	.005	19			

Table 6.12: The results of the one-way ANOVA analysis for the group factor *Talent*, separately for condition *T* and condition *J*.

parison is possible.

The results of the one-way ANOVA with the group factor **talent** did not show any significant differences between the two groups (between group sig. .307 for the *T* condition and .533 for the *J* condition). Therefore, **Hypothesis B3**: talented and non-talented subjects behave differently in the read speech task, can be *rejected*. **Neither** talent group shows any significant increases or decreases of the match values for the two word lists in any of the conditions. Thus talent is not a factor for convergence in read speech.

6.1.4 Gender in read speech

Another possible influencing factor in read speech is the gender of the participants. It might be, as **Hypothesis B4** states, that **female and male subjects behave differently in the read speech task**. In order to test this claim, a one-way ANOVA was calculated for the two subsets of female and male participants.

	Levene Statistic	df1	df2	Sig.
conv_T	.961	1	18	.340
conv_J	.176	1	18	.679

Table 6.13: The results of the Levene Test for Homogeneity of Variance. Both results are above the crucial significance level, indicating that the variance of the variables is indeed comparable and calculating an ANOVA is therefore legitimate.

		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
conv_T	Between Groups	.000	1	.000	1.150	.298
	Within Groups	.007	18	.000		
	Total	.007	19			
conv_J	Between Groups	.000	1	.000	.000	.990
	Within Groups	.005	18	.000		
	Total	.005	19			

Table 6.14: The results of the one-way ANOVA analysis for the group factor *Gender*, separately for condition *T* and condition *J*.

The Levene statistic (see Table 6.13) shows homogeneity of variance, so the requirement for calculating an ANOVA is met. As the data in Table 6.14 state, the differences in the *T* and the *J* condition between the groups are not significant. The difference of values for *conv_J* between women and men in the experimental group is close to zero, as the significance in this case approaches 1 (between-group comparison, significance =.990).

To illustrate the results of the one-way ANOVA, additional plots of the mean values for *conv_T* and *conv_J* are shown in Figure 6.7 and Figure 6.8.

The boxplot in Figure 6.7 shows a higher and positive *conv_T* for female speakers, indicating that their post-test match values were indeed higher than the pre-test values before the dialog. The male speakers show a small negative value instead. As the ANOVA showed, the difference between the two groups is not significant. Figure 6.8 shows the same boxplot for condition *J*. Here, the ANOVA results in Table 6.14 state a significance of close to 1, which is illustrated by the almost equal distribution of the *conv_J* values for female and male speakers. The mean values here are almost identical (-.0039 and -.0038). Two subjects, one female from the talented group and one male from the less talented group, are outliers from the main tendency. The *conv_J* values of -.04 each point towards a slight tendency to diverge. When taken into account separately from the remaining 18 subjects, a paired samples t-test shows an almost significant effect (sig.= .013) for divergence in the pre-/post-test comparison for these two subjects.

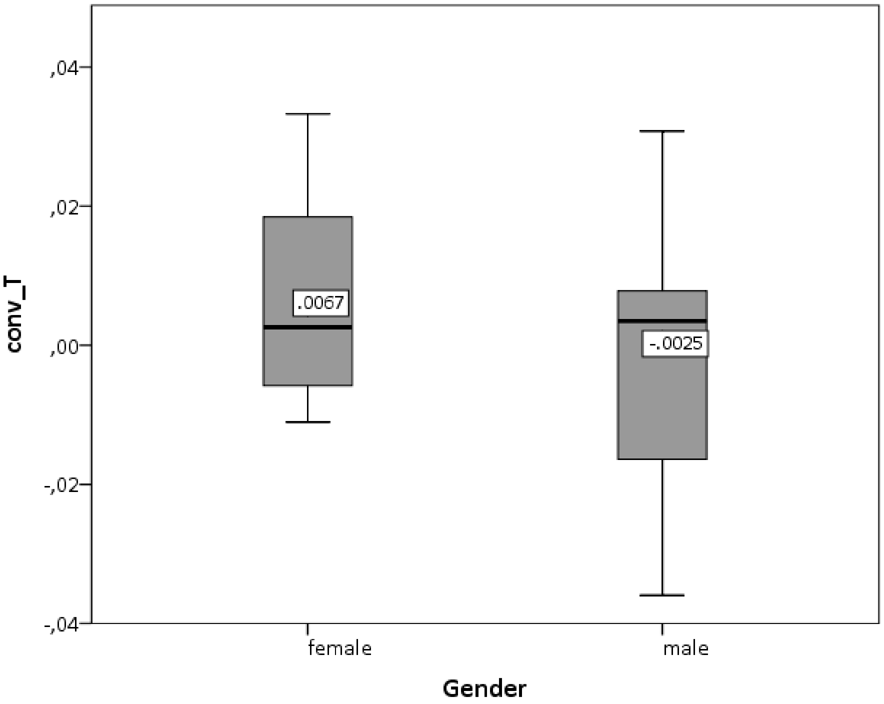


Figure 6.7: Mean change of *conv_T* and the SD of the match values of the 10 male and 10 female subjects compared to native speaker *T*. The numbers on the boxplots indicate the mean value.

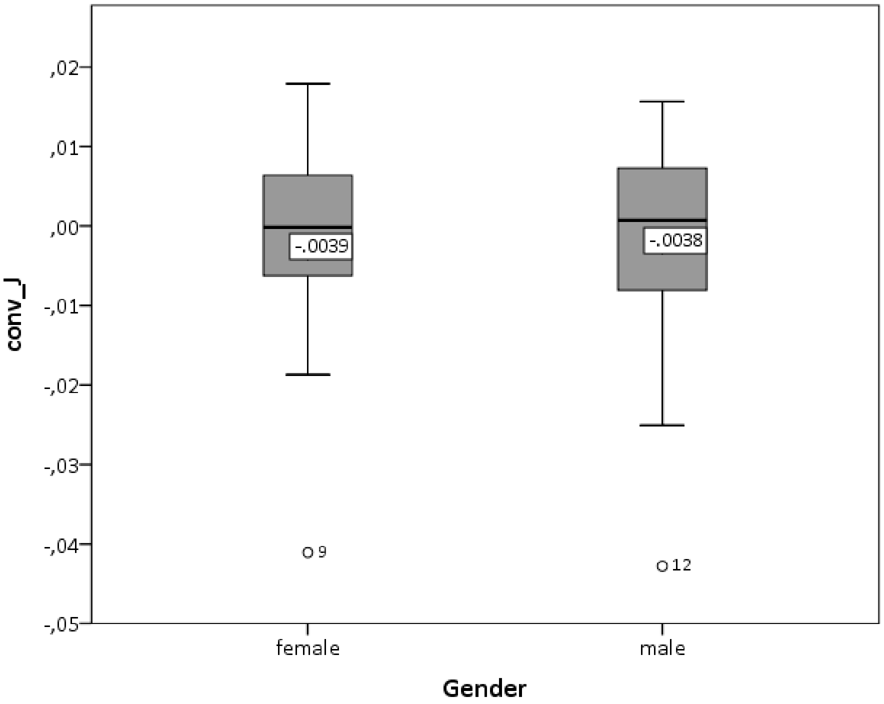


Figure 6.8: Mean change of *conv_J* and the SD of the 10 male and 10 female subjects compared to native speaker *J*. The numbers on the boxplots indicate the mean value.

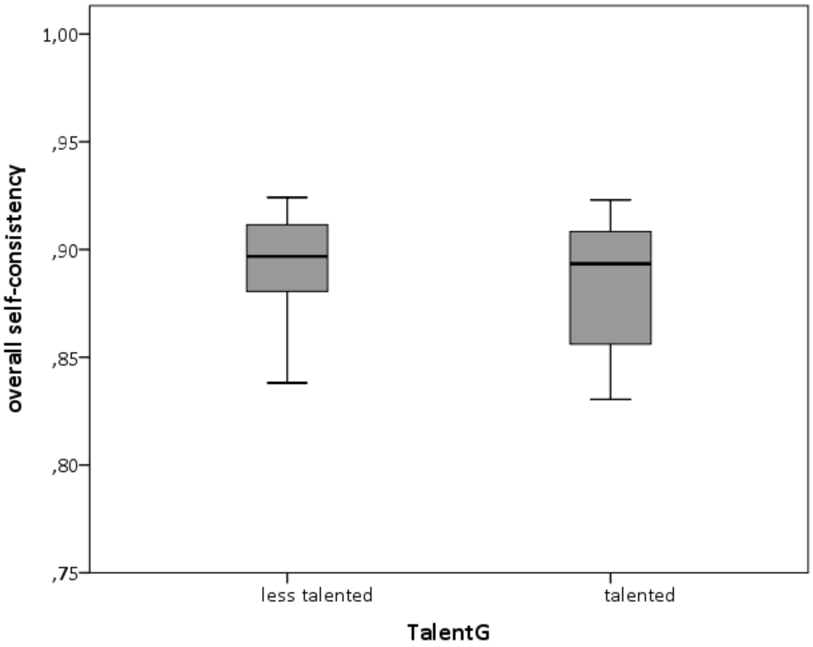


Figure 6.9: The mean self-consistency values from all three list comparisons for the 10 talented and 10 less talented subjects.

After considering the results of the one-way ANOVA and the descriptive statistics, **Hypothesis B4**: **female and male subjects behave differently in the read speech task**, can be **rejected** as well. Gender was not a factor in convergence within the read speech task.

6.2 Nonnative speakers' self-consistency

The comparison of match values for the pre- and post-test of the NNS and NS have served to answer questions about the *convergence* of the NNS *towards* the native speakers in the read speech task. If we instead compare the nonnative speakers' amplitude envelopes from the three word lists to one another, we can find out how self-consistent the pronunciation of the word list contents actually were. Using the definition given in Chapter 5.1.2, the self-consistency measure allows us to track how variable one's own pronunciation is, without relating the match to another speaker. The variance, or *perturbation*, in one's own speech can but does not have to be due to the dialog partner nor need it go in the direction of the dialog partner. However, it is supposed that *talented* speakers show *lower* self-consistency values in the read speech task, since they converged more to their dialog partners and this might carry over in *perturbed* read speech afterwards.

Therefore **Hypothesis B5** states that **talented subjects will show more perturbed self-consistency values than less talented ones**. Following this line of thought, subjects from the talented group will show less good matches of their own read speech when comparing the three renderings of the word lists.

Table 6.15 presents the mean, standard deviation (SD) and range of the self-consistency values for the comparisons between:

- all three lists compared against each other – the overall self-consistency value
- List 1 & List 2 – the self-consistency value for before/after the *T* dialog
- List 2[6] & List 3 – the self-consistency value for before/after the *J* dialog
- List 1 & List 3 – the self-consistency value for the comparison of the values before the dialog session & after the last dialog.

The overall self-consistency for the *less talented* speakers amounts to .895, while the same value for the *talented* group is .885. The *talented* subjects show a slightly smaller intra-speaker consistency, but at the same time also a higher SD and a greater range (.093 for the talents vs. .086 for the less talented speakers), which is also displayed on the boxplots in Figure 6.9. However, an ANOVA[7] with the between-subject factor of *Talent* did not show any significant difference for the overall self-consistency of the two groups (sig. of the variance test .456). The same result was obtained for all other self-consistency measures: none of them proved to be significantly different considering the group factor of *Talent*. The significances in the ANOVA (between-group factor *talent*) were as follows:

[6] i.e. List 2b used for calculating convergence in read speech in the previous chapters.

[7] Homogeneity of variance was given in all following cases.

Talent		overall self-consistency	self-consistency 1-2	self-consistency 2-3	self-consistency 1-3
less talented	Mean	.8945	.8947	.9013	.8884
	SD	.0249	.0258	.0215	.0329
	Range	.086	.089	.068	.109
talented	Mean	.8848	.8889	.8899	.8765
	SD	.0316	.0322	.0340	.0340
	Range	.093	.092	.114	.104
total	Mean	.8896	.8918	.8956	.8825
	SD	.0281	.0286	.0283	.0331
	Range	.094	.094	.121	.109

Table 6.15: Mean and range of the self-consistency values for all three word list comparisons and the overall self-consistency value, separately for the two talent groups and in total.

- List 1-2 sig.= .665
- List 1-3 sig.= .435
- List 2-3 sig.= .383

If we take a look at the self-consistency for all three list comparisons in Figure 6.10, both groups display exactly the same directions of changes, with *less talented* subjects showing slightly higher values overall. Self-consistency drops the most for the comparison of List 2 and 3. That these changes are not significant becomes evident from Figure 6.11, where the scale has been zoomed out to cover the range of match values from 0.5 to 1.0. The differences between the groups are of a very small magnitude, hence the insignificant results. A paired samples t-test for the differences between the self-consistency values of all participants showed one almost significant result (sig. 2-tailed= .023) for the pair *List 1-2 vs. List 1-3*, and one significant result (sig. 2-tailed= .003) for the pair *List 2-3 vs. List 1-3*.

To conclude, **Hypothesis B5**, which stated that talented subjects should show more perturbed self-consistency values than less talented ones, is **rejected** due to

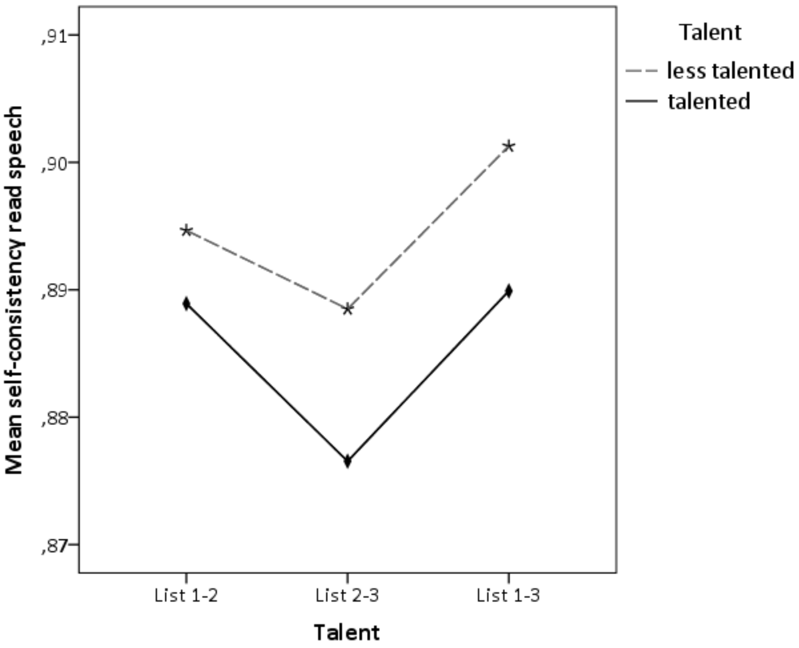


Figure 6.10: Mean self-consistency values from all list comparisons for the two talent groups and the direction of change in the match values. To emphasize the direction of changes, the scale has been zoomed into (compare Figure 6.11).

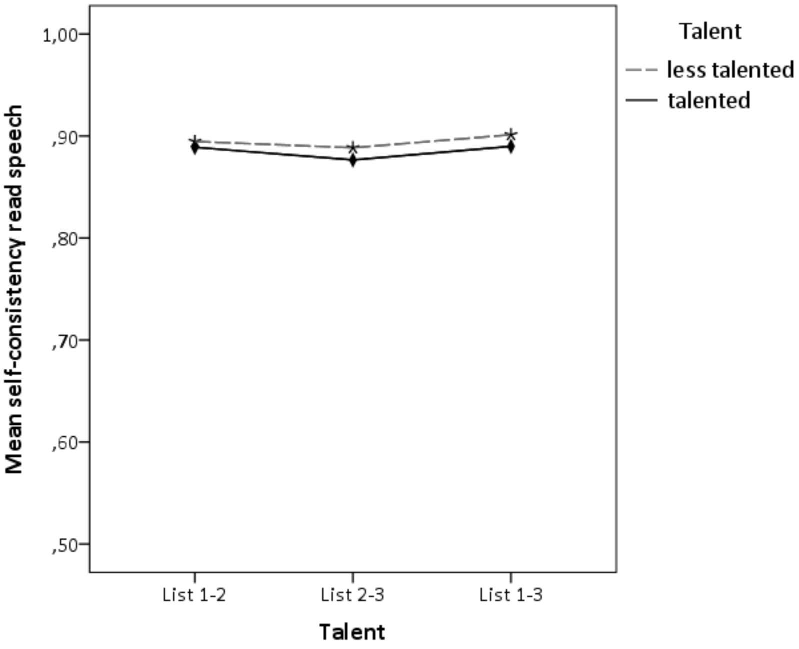


Figure 6.11: The same mean self-consistency values as in Figure 6.10, but on a different scale – the minimum here is 0.5 and the maximum 1.0.

the lack of evidence for any significant group differences in the self-consistency values of the nonnative speakers. Instead, both groups show slight tendencies going in the same direction.

6.3 Summary of read speech results

The following five hypotheses were tested regarding the read speech data:

1. **Hypothesis B1**
   Subjects show convergence to the native speaker in the read speech task following the dialog.

2. **Hypothesis B2**
   Convergence in read speech is positively correlated with convergence in dialog speech.

3. **Hypothesis B3**
   Talented and non-talented subjects behave differently in the read speech task.

4. **Hypothesis B4**
   Female and male subjects behave differently in the read speech task.

5. **Hypothesis B5**
   Talented subjects will show more perturbed self-consistency values than less talented ones.

Chapter 6.1 dealt with the analyses of convergence between the nonnative speakers and the English native speakers. Hypotheses **B1**, **B2**, **B3** and **B4** were all rejected. The analysis of the read speech pre- and post-test data suggests that subjects showed no convergence toward the respective native speakers in the read speech task (Hypothesis **B1**). The shifts in match values that occurred were small in magnitude and were not correlated to the subjects' convergent behavior in the dialogs

(Hypothesis **B2**). The subsequent detailed analyses of group effects showed no significant differences between the performance of *talented* and *less talented* speakers or *female* and *male* speakers in the pre- and post-test (Hypotheses **B3** and **B4**).

The self-consistency data reported in Chapter 6.2 also pointed to no significant differences between the two talent groups. Neither of the groups displayed significantly more or less consistent match values. On the contrary, both showed similarly directed patterns, which led to the rejection of Hypothesis **B5**.

Chapter 7

Discussion of results

The following chapter contains a critical discussion of both the dialog and read speech results of the studies performed. The results will be compared to similar as well as differing results and backed up with possible theoretical explanations. The first part is concerned with the main experimental data from the dialog task (Chapter 7.1), while the remainder of this chapter will elaborate on the analysis results for the read speech pre- and post-test (Chapter 7.2).

7.1 Discussion of dialog results

This section begins with several general remarks and a critical review of the dialog experiment methodology (Chapter 7.1.1), which will then be followed by arguments in favor of a hybrid model of convergence (Chapter 7.1.2) and a commentary on the apparent speech style encapsulation of convergence (Chapter 7.1.3). The last part will deal with the convergence results from the angle of a usage-based account of language, with ideas for the incorporation of attentional selection processes in convergence into an exemplar theoretic framework (Chapter 7.1.4).

7.1.1 General remarks on the methodology

The first general statement that can be made, is the confirmation of **convergence** effects for native-nonnative interactions (as stated in **Hypothesis A1** in Chapter 5.1). The mean values for the comparison of early vs. late measurements revealed positive shifts in pronunciation of the nonnative toward the native speakers, as expected.

The applied elicitation technique has positive as well as negative sides, but with advantages still outweighing the drawbacks. Eliciting data with the Diapix task (Chapter 4.1.3), on the one hand, allows for the collection of rich speech in a communicative setting, which is an undeniable prerequisite for investigating naturally-occurring convergence. On the other hand, though, deciding for a quasi-spontaneous setting, takes away the possibility of controlling the amount of data at hand, which is possible in more controlled settings[1]. It is obvious that every speaker uttered our lexical target items a varying number of times and it was strictly not our intention to force a certain number of repetitions. This led to a *variable amount* and *type* of data for each pair of speakers[2]. By using quasi-spontaneous or fully spontaneous designs, we cannot guarantee a high number of neatly produced repetitions, but we are certainly able to capture the core of accommodation processes which naturally occur in a communicative situation. These processes were not forced or mechanical due to highly controlled laboratory conditions, which rather leads to *imitation* (compare Chapter 3.2 and 3.3.2). The need for a real communicative setting instead of task-oriented settings with only one person being *actively* involved, becomes even more warranted when considering the implications of **Hypothesis A5**, which states that shifts in speech style even within the experimental task strongly impact convergence (more in Chapter 7.1.3).

A consequence of dealing with variable amounts of data in general is the variability of the objects of comparison in the measurement **sets** (see Chapter 5). When speaking of the comparison of, for instance, **Set 1** and **Set 2** for all nonnative speakers, we must bear in mind that the composition[3] of those sets differs from speaker to speaker as well as from set to set. This might be a possible explanation for the lack of a direct relationship between the achieved balance level at a late point in the dialog (Chapter 5.2.2) and the amount of convergence of both the native and nonnative speakers, as tested in **Hypothesis A7**. The sets for determining the native and nonnative speakers' convergence contained items occurring both *early* **and** *late* in the dialogs, while the measurement of the mutually achieved balance level contained **only** items occurring *late* in the dialog. These *late* items in **Set 3** might therefore have been different target words than

[1] as, e.g., in word repetition studies described in Chapter 3.2.

[2] One speaker, for instance, might have used the target word "door" six times, another speaker only once but used "window" ten times instead.

[3] i.e., the type of target words included and their number.

those measured in **Set 2**, therefore producing slightly differing match values. It is very possible that some target words were accommodated to a stronger degree, depending a.o. on their frequency, which was not controlled for due to the limited amounts of data suitable for comparison[4] produced by the subjects. Given the important role frequency of occurrence plays in usage-based accounts of speech (compare Chapter 2.1.3), it cannot be ruled out that it influenced the degree of convergence for individual words in the present study. As suggested in Hintzman's MINERVA2 model, high-frequency words might be less likely to produce good "imitations" due to the naturally higher number of *echoes* or activated exemplars [Hin86, Gol98] (compare Chapter 2.1.3). This might be a possible reason for the lack of a direct correlation between our convergence and balance measurements (compare Chapter 5.2.2), since different underlying sets of target words might also have produced different match values, in some cases probably not elligible for a direct comparison. This fact, paired with the rather small magnitude of effects we usually find for convergence (compare Cappella and Planalp [CP81] and Nielsen [Nie07]), could explain the non-significant findings for the relation of the *balance level* and the amounts of convergence of the NNS and NS in the dialogs, as well as the lack of a clear inverse relationship of the natives' and nonnatives' convergence to one another. There, only a *tendency* for compensating the lack of convergence on the other speaker's part was found, but no significant negative correlation could be confirmed (see **Hypothesis A6** in Chapter 5.2). A clear proof for a *state of balance* dialog partners achieve, as Gregory and Webster [GW96] argued for, can therefore not be reported.

As far as the measurement method is concerned, it allows us to gain a more global and objective evaluation of acoustic properties[5], and not only a subjective perceptual evaluation. Using amplitude envelopes, as argued for elaborately in

[4] Meaning the target words, which were content words only and, as explained previously, appeared in varying numbers within the dialogs.

[5] which was a.o. proposed by Nielsen [Nie07] to get more robust results. However, what she implied was solely a global *perceptual* evaluation.

Chapter 4.2.2, allows us to obtain a measure of similarity between two stretches of speech (as our target words) without the necessity of relying on individual features. As described in Chapter 4.2.2, amplitude envelopes are a measure of the *spectral* similarity of two signals. When the curves of two acoustic waves, decomposed into their four amplitude envelope signals, fit onto each other very well, they obtain a high match[6]. However, this is only true if the two signals were uttered with the same speed. If the speed, however, varies, and the timing of the spectral events is misaligned, the match value calculated for the amplitude envelopes will be automatically lower. Considering Wade and colleagues' [WDS$^+$10] argument that memory takes a spectro-*temporal* form, a temporal match between two sequences should be as important as the spectral similarity. As the current method captures only the spectral part of convergence, we cannot exclude the possibility that the achieved match values would be slightly altered by the inclusion of a suitable measurement of timing differences. On the other hand, if a temporal match is as important as a spectral match and speakers and listeners encode this dimension, they should also have access to it to the same extent as the spectral dimension and be able to converge to it, too. Ideally, what we then measured with amplitude envelopes already comprised the convergent effect for both dimensions simultaneously. Nevertheless, having an additional component accounting for timing variability would allow us to compare how well both dimensions of the speech signals are accommodated by the speakers and determine if one might be harder to grasp than the other. Chapter 8.3 discusses possible solutions for such an incorporation of temporal information into the measurement of convergence.

Another essential property of exemplar models is the acknowledgement of the importance of context [Haw03, WDS$^+$10] in storage and retrieval. Our choice of extracting *words* as the unit for comparison goes in line with Hawkins' view of stored exemplars, but runs partially against the context sequence model of Wade and colleagues, who argued for a full contextual storage of linguistic material. Goldinger [Gol98], whose 1998 study based on Hintzman's model [Hin86], also speaks of

[6] In our case, a high match signifies a value close to 1.

"episodic traces" rather than of words. They, however, only argue against the existence of "word prototypes", not of a multitude of stored word exemplars (compare Chapter 2.1.2). We left aside the contextual embeddedness of the relevant target words on purpose, since controlling for context and pairing only those items coming from the same environment would have required a much richer data set than dialogs of only 10-15 minutes length.

The time frames defined for dividing the dialogs into an *early* and *late* part were set to respectively $\frac{1}{3}$ starting from the beginning of the dialog, and $\frac{2}{3}$ until the end of the dialogic interaction part (before the summary started). A perceptual evaluation of those boundaries could confirm fairly well that little or no adaptation occurred during the first third of the dialog, while the accommodation was audible in most cases after the two-third mark. Obviously, there were a few exceptions in both directions. One talented speaker[7] started particularly early to accommodate to a British accent, with the first utterance that sounded British appearing only 90 seconds after the beginning of the task[8]. For others (mostly less talented speakers), it was sometimes hard to identify any clear boundaries due to the lack of subjectively-detectable convergence, which was later on confirmed in the measurements for the *less talented* group presented in **Hypothesis A2** in Chapter 5.1. It is still possible that some part of the convergent behavior was not accurately encompassed by our measurements, for instance, in cases with a "ceiling effect", where convergence started immediately after the beginning of the task (referred to also by Delvaux and Soquet [DS07] and Willemyns and colleagues [WGCP97], compare Chapter 3.3.2). Nonnative speaker number *1* partially displayed this ceiling effect, with a very fast switch from an American to a British accent. A more elaborate discussion of early selection mechanisms connected to this phenomenon will be given in Chapter 7.1.4.

[7] Speaker number *1*.

[8] Which was still well before the one third boundary in this dialog.

7.1.2 A hybrid model for convergence

Confirming overall convergence for the twenty nonnative subjects for an early-late comparison in both dialog conditions[9], despite the members of the *less talented* group showing mostly little or no positive shifts toward their native speaking partners (see **Hypothesis A1** and **Hypothesis A2** in Chapter 5.1.1), implies that the *talented* group members converged fairly strongly. This is in line with our expectations based on the assumption of fluid social identities (e.g. Park [Par07], see Chapter 1.1) and a status inequality between the conversational partners, due to their native vs. nonnative speaker identities (Chapter 1.2).

The nonnative speaker faces a situation in which the following questions must be answered (though not necessarily fully consciously):

- Who am I...
- Who do I have to be...
- Who do I want to be... in this dialog?

The starting point for the nonnative speakers is a forced identity change from who they normally are: *native speakers of German*, to who they have to be in the experimental task: *nonnative speakers of English*. Imposing a new identity onto the German subjects goes hand in hand with a shift in dominance. At least when mastering the experimental language English, the subjects have to submit to a lower status in the dyad and face the dominant status of the linguistic experts – the native speakers of English. At this point a negotiation of identities starts in the dialog situation (see Chapter 1.1.1 and 1.1.3), in which, as we predicted, the nonnative speakers would try to prove their competence in EFL[10] and put a considerable effort in coming across as good or even excellent speakers of English (the "desired self-image", Ting-Toomey [TT99]), and "live up to the expectations"

[9] For the dialogs with the SSBE speaker *J* and the GA speaker *T*.

[10] English as a foreign language.

the situation imposes on them[11]. The negotiation of status usually proceeds as a simultaneous interplay of reflective positioning (self-presentation) and interactive positioning (through the interacting partner) [BP01] (compare Chapter 1.1.3). However, the negotiation process in the current study was manipulated by the experimenter. By informing the native speakers about the purpose of the study and asking them to maintain their own speaking style and **not** to converge to the nonnative subjects, we interfered with the factor of *interactive positioning*. We will now focus on the outcome of this interference and on the results of **Hypotheses 2, 3** and **6**, which are directly related to the questions of identity negotiation and the degree of control in accommodation.

In spite of sharing the same starting point for the negotiation of a new identity, **Hypothesis 2** showed that there are significant differences in phonetic convergence between the two **talent** groups. This is to say that, although all twenty speakers shared a high proficiency level in English (compare Chapter 4.1.1), their phonetic talent was a decisive factor for the amount of phonetic convergence they displayed toward their native speaking partners. The *less talented* group showed, on average, much lower match values, ranging on average from divergence to maintenance and small degrees of convergence. The *talented* group showed considerably stronger convergence in both native speaker conditions. The two NS functioned as model speakers, whom the nonnative speakers should (ideally) approach by "being responsive and accommodative" [Dav03]. The accommodative part of this equation was clearly mediated by the *talent* factor of the nonnative speakers, showing that *individual differences* have an influence on convergence in the L2 and that the process itself by no means runs off completely automatically, as suggested by Pickering and Garrod [PG04b, PG04a, PG05, PG06], since the necessary phonetic skills are a prerequisite, without which convergence toward the target is not possible. Moreover, the performance of the speakers in the two dialog conditions was correlated,

[11] compare [Bab09, BG77, Gil73, GW96, GG02, Par06, WvVed] for the influence of status and dominance. See Chapter 3.3.1.

meaning that a talented speaker showing a lot of convergence in one dialog also showed a considerable positive shift in the second dialog, and vice versa (see Figure 5.1 in Chapter 5.1.1). The correlation of the two measurements points, once again, to between-subject differences at the individual level. The existence of a core automatic *component*[12] in convergence, however, cannot be totally denied when bearing in mind **Hypothesis A6**, which we were able to confirm. The accommodation of the native speakers toward the NNS proceeded *despite* their explicit knowledge about the purpose of the study and the request for the suppression of any positive shifts in pronunciation.

Although *phonetic talent* proved to be a very good predictor for the amount of convergence displayed by the nonnative speakers, the between-subject variance in both groups entails the presence of other explanatory factors. These might be the personality or psychological features of the subjects, for instance the need for social approval, ratings of mutual attractiveness and/or liking of the conversational partners, as have been found in many previous studies (compare Chapter 1.3 and 3.3.1, [ACGSY11, LMG77, Nat75b, Nat75a, Nat76, PG08]. A further factor that could give more insight into the between-subject variance is a finer distinction of the *talent* of the nonnative subjects. The rather coarse grouping into only two subsets is based on a preliminary analysis of the tests described in Jilka [Jil09a] (see Chapter 4.1.1). A re-analysis of the current study to include multiple groups arranged according to the subjects' phonetic talent might account for even more variation than the current two-group classification.

In contrast to talent, *gender* did not prove to be a decisive factor for the amount of convergence of the nonnative speakers (**Hypothesis A3**). This finding supports the majority of convergence studies, which also have not reported any gender-related differences. The tendency, though not reaching a significant level, goes into the direction of male speakers showing more convergence than female speakers. This goes against the results reported in Namy [NNS02], but is in line with the tendency presented by Pardo [Par06] (compare Chapter 3.3.1). What

[12] and thus, a component operating subconsciously.

Pardo suggests in the face of rather inconclusive results as to gender differences, is the greater impact of *attentional* mechanisms on accommodation rather than any strictly gender-specific mechanism (as was suggested by Namy). This points, again, to the strong influence of individual differences on accommodative behavior in dialog. The predictory value of gender for convergence thus remains as weak as it has been found for general L2 performance by Piske [PMF01]. The influence of attentional mechanisms on convergence will be revisited in Chapter 7.1.4.

As was mentioned in the introductory part to this chapter, the experimenter manipulated the conditions for the identity negotiation process in this study. The native speakers were explicitly asked not to engage in pronunciation convergence, in the hope of eliciting more convergence from the nonnative speakers. However, the statement in **Hypothesis A6** about the convergence of the **native speakers** toward the nonnative subjects was accepted. This points to the occurrence of accommodation where it was not expected, especially given the request for suppressing convergent tendencies. Given that the native speakers tried not to converge and were mostly positive about having succeeded in doing so[13] (yet they still converged), we probably witnessed automatic and consciously uncontrollable convergence, described by Pickering and Garrod as *alignment* [PG04b]. The possibility that speakers lack a conscious awareness of the influence they exert on each other has also been proposed by Cappella and Planalp [CP81].

Although a tendency toward more convergence in dialogs with the *less talented* group was found, the difference was not significant for either of the two native speakers. The lack of a clear inverse correlation between NNS and NS accommodation, which has been proposed by Gregory and Webster [GW96], could be due to the measurement limitations[14] described earlier, and also to the introduced *manipulation*. This is the more likely of the two possibilities. The native speakers might have been able to suppress a part of the accommodation they would have normally

[13] When asked after the dialogs, both native speakers mostly stated that they tried to speak with their usual English accent and they did not adapt to the nonnative speakers.

[14] different target words measured for the specific early-late comparisons, compare Chapter 7.1.1.

displayed toward nonnative speakers if they had not been asked for maintenance, so that only the automatic, subconsciously controlled part was left and surfaced in the dialogs. Since there was no significant difference between the accommodation toward the two talent groups, i.e., no explicit pattern was detectable in the native speakers' behavior, the two following explanations seem conceivable:

1. The NS were able to consciously turn off the overcompensating part which would have led to especially strong convergence toward the phonetically less talented subjects.

2. All nonnative speakers are automatically categorized as one group of linguistically less proficient speakers (compared to the "expert" status the NS has), therefore receiving on average the same amounts of convergence.

Personality as well as psychological and social factors might, of course, play an additional role in explaining the amount of convergence in the above cases. The observed convergence of the NS in the dialog did not happen consciously; this, however, does not mean that it proceeded in an uncontrolled manner. It could have been a *subconsciously controlled* mechanism serving to enhance communication and build up an acceptable balance level (compare Styles [Sty06] and Neuman [Neu84] in Chapter 2.2.3 on the automaticity and control). Subconsciously controlled processes probably facilitate communication and are maybe also grounded in a basic biological need for synchrony[15], but they do not meet with the strictly socially motivated explanation provided by CAT followers [GS79, GCC91b, GCC91a] (see Chapter 1.3.1).

The changes in communicative style described by CAT imply that the speaker at some point has a choice, which in its nature must be consciously accessible. If, however, a conscious decision is made against accommodation and positive shifts *are* still observed, it would seem as though this part of adaptation eludes an *active* negotiation of social distance. Unless, of course, the need for reducing social dis-

[15] or *entrainment/coordination dynamics*, as suggested a.o. by Kelso and colleagues [Kel97, KE06, Kel09, OdGJ+06, OdGJ+08].

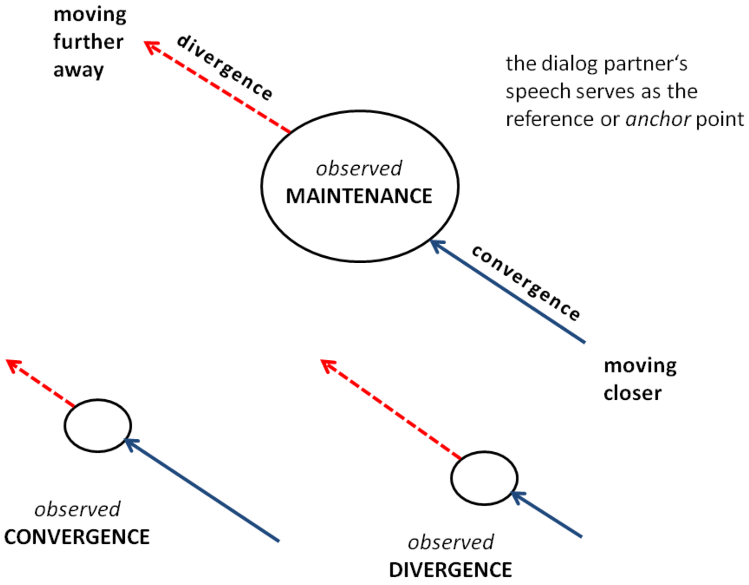


Figure 7.1: Illustration of a dynamic interplay between convergence and divergence with three possible observed outcomes: maintenance, convergence and divergence. The dialog partner's speech is the anchor point from which the speaker can move away or come closer in her or his pronunciation.

tance is reflected precisely in the *automatic part* of convergence we are looking at in the native speaker data. But then it would still not be a conscious decision on the part of the speaker. Speculating even further, if this automatic part of convergence were biologically founded and, assuming that the linguistic prerequisites were met (the necessary proficiency in the language or dialect), maintenance and divergence might in fact be composed of *convergence* **and** *divergence* happening simultaneously, as has, for instance, been suggested by Wedel and van Volkinburg [WvVed]. The observed measurable outcome would then depend on the dynamic interplay and the proportions of both trends, as presented in Figure 7.1.

Summarizing the above **hypotheses** leads to the conclusion that convergence most likely requires a **hybrid model** to account for the *individual differences* influencing the mechanism of pronunciation adaptation. Such a hybrid model would involve features as, for instance **talent**, which is not subject to *conscious* control, and also personality and psychological mechanisms, which form the frame for the

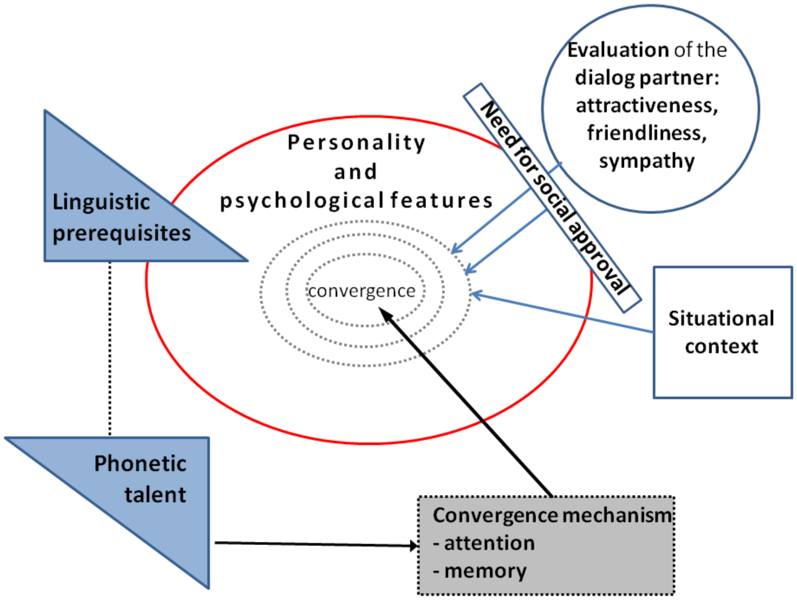


Figure 7.2: A hybrid model of convergence: a network of factors influencing the measurable amount of convergence present in a dialog. On the left side: the necessary prerequisite of a linguistic proficiency level allowing for convergence (concerning not only languages but also dialects) and the variable factor of phonetic talent, which bear a direct influence on the mechanism of convergence, with possible underlying attentional and memory components. The amount of convergence is delimited by the framework of individual differences in personality and psychological features. Further impacting factors are the ratings/evaluations of the dialog partner and the given situational context. All three factors determine the current need for social approval, which also impacts the final amount of convergence/divergence surfacing in the dialogic interaction, shown by the dotted line circles in the center of the picture.

amount of convergence displayed. *Social goals* are also relevant[16] (see Figure 7.2). Talent for pronunciation might have special underlying *attentional, memory and control* components[17] which influence the mechanisms of exemplar storage and selection. As shown in Figure 7.2, talent has a direct link to the convergence mechanism. The fact that some parts of convergence seem to proceed without our conscious knowledge could be attributed to a procedural memory component (PM) playing a major role at this stage (compare Ullman's declarative/procedural model

[16] Such a hybrid model has been argued for by, for instance, Krauss and Pardo [KP04].
[17] further discussed in Chapter 7.1.4.

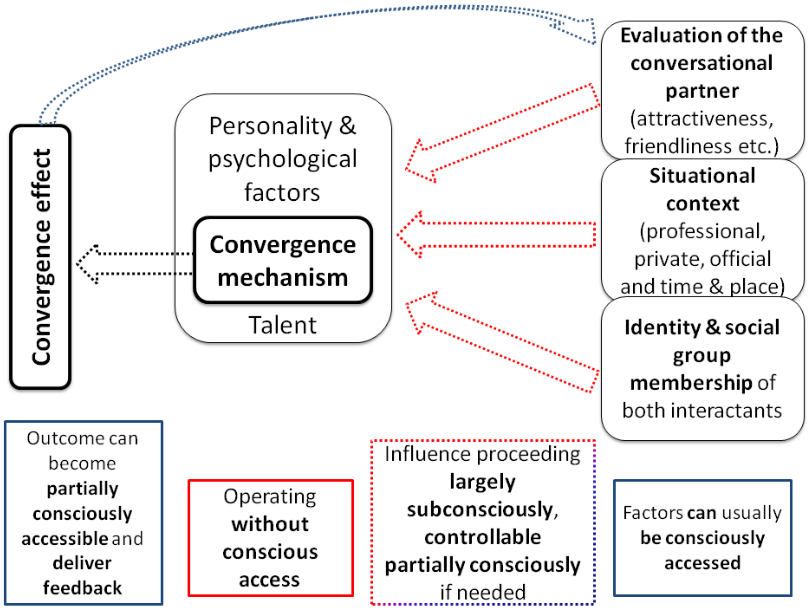

Convergence effect
Personality & psychological factors
Convergence mechanism
Talent
Evaluation of the conversational partner (attractiveness, friendliness etc.)
Situational context (professional, private, official and time & place)
Identity & social group membership of both interactants
Outcome can become partially consciously accessible and deliver feedback
Operating without conscious access
Influence proceeding largely subconsciously, controllable partially consciously if needed
Factors can usually be consciously accessed

Figure 7.3: A hybrid model of convergence: the supposed influence paths of those factors which can be consciously accessed and those which remain largely below a conscious level. Although the presence of some situationally relevant factors may be overtly accessible, their influence on the convergence mechanism still remains covert. Becoming aware might be part of the end product – the convergence effects – which might in turn provide feedback for a re-evaluation of the situational factors and introduce a fine-tuning to the manipulable part of the process.

[Ull01b].). Ullman described this part of his memory system as neither overtly accessible nor consciously controllable. A speaker could, however, become aware of the outcome of the convergence she or he displayed if it is sufficiently strong in degree or concerns particularly prominent features. This could be caused by a comparison of produced exemplars with stored exemplars in declarative memory (DM), in a type of self-monitoring loop. Such an assumed dual-route reliance on memory resources might explain the aforementioned back-and-forth movement of convergence[18] of nonnative speaker number *1* toward native speaker *J* and her British English accent (compare Figure 7.3). The convergence process itself was not con-

[18] The reported fluctuation has been assessed perceptually and the observed dynamic pattern corresponds fairly well to the amplitude envelope measurements within the early and late portions of the dialog.

sciously accessible and, given the favorable situational context and the high phonetic talent, the speaker consequently found himself producing strongly accommodated lexical items. This he probably noticed[19] and therefore consciously tried for a brief moment to switch back to his usual accent, American English. Eventually, the convergence mechanism kicked back in and the situation was repeated. Figure 7.3 illustrates the proposed interactions of consciously aware and unaware processes in convergence.

Although accommodation can certainly be influenced and altered by many endo- and exogenous variables, the existence of *basic alignment*, beyond our conscious control seems to be very probable. So the question is not, whether there *is* accommodation, but rather *how much* of it surfaces.

7.1.3 Speech style exclusivity

The persistence of convergence effects beyond the main experimental task reported in some studies (Pardo [Par06], Delvaux & Soquet [DS07]) could not be confirmed. Instead, as analyzed in **Hypothesis 5** in Chapter 5.1.3, the match values of the nonnative subjects paired with the native subjects **decreased significantly** once the summary part of the experiment started. Noteworthy here is that the conversational partner remained in the sound attenuated booth while the nonnative speaker summarized the task findings and the summary followed the experimental task immediately, with no break in between. The only thing that changed in the setting was the switch from a dialogic interaction to a first person narrative, without the second speaker actively participating (see Chapter 4.1.2). The comparison between match values at a late point in the dialog and the summary revealed that any convergence achieved early on fell back to a level comparable to the beginning state, as measured by the early-early comparison in **Set 1** (see Figure 5.7 and 5.8) between the subjects.

[19] Some speakers reported after the experiment that they were at times surprised about their own pronunciation, indicating that they were aware of at least some of the changes.

The results may at first seem surprising, but if we take into account that style is highly context-dependent and dynamically changing (see Chapter 1.1.2), and that we took away from the subjects the basic prerequisite for convergence happening in a communicative situation – namely the **dialog partner**, the nonnative speakers' behavior becomes justified. As described in Chapter 1.1.2, dialog and monologue show a multitude of distinct features and make different demands on the speakers. Once the summary starts, the speaker is no longer confronted with an interacting partner, but merely with a not actively participating audience – the native speaker and possibly also the experimenter, who asked the NNS to deliver the summary. Neither the native speaker nor the experimenter were involved in an interaction with the speaker, but rather allowed him to present his findings as an uninterrupted monologue. In any case, the previously forced "status inequality" between the native and nonnative speaker which caused an asymmetrical alignment disappeared [Par07]. It can be viewed as a situation where the nonnative speaker is suddenly left without her or his "anchor point", using the term from Figure 7.1. Without the necessary reference person for pronunciation being an active participant in the situation, phonetic accommodation toward this person becomes unnecessary. This crucial change in the setting probably inhibits both the automatic and the consciously influenced part of convergence. Consequently, the whole convergence process described in Figure 7.2, with all its subconscious mechanisms and conscious considerations, is probably not even initiated.

Given that the observed drop in match values for the summary is supported fairly well by sociolinguistic factors arising through a change of the situational setting, the more interesting question here is how the speakers were *able to change* their pronunciation according to the speech style they are currently using. Hawkins stated in her *Polysp model* that experienced listeners use information about style and accent instead of just relying on a hypothetical "canonical" pool of exemplars in speech recognition (see Chapter 2.2.2, Hawkins [Haw03] and also Pierrehumbert [Pie06]). If they utilize these speech style-dependent pools for recognition, it seems straightforward that they use them in speech production as well. Convergence to another speaker might be connected not only to actively choosing exemplars fitting the current context but also to the **suppression** of their own exemplars. A shift in speech style away from the necessity to converge[20] allows them to return to a non-convergent way of speaking, which probably comes more naturally for the speakers. The two talent groups did not differ significantly in this measurement of speech style change, indicating that *talent* stops playing a role once convergence toward the native speaker of English is not necessary anymore. Here, all speakers simply returned to a more "self-centered" way of speaking, in which aptitude for phonetic parameters no longer influences the selection process. A rich indexing of exemplars, as proposed by usage-based theories of language, including a.o. indexes for speech style, would allow for such an immediate switch to a narrative style, which is not directly influenced by dialog exemplars. In assuming this, we do not claim that the convergence displayed during the dialog has magically disappeared; it is simply not surfacing in the subsequent monologue tasks (both the summary and the post-test involving read speech), because the new situational context demands the selection of other, better fitting exemplars. If the experimental setting had required the two speakers to continue their dialog after the summary had been completed, the match values would with high probability have been on the increase again.

A more detailed account of how the storage and selection of exemplars might proceed in such a case is presented in Chapter 7.2, together with the discussion of the pre- and post-test, where nonnative speakers were found to display exactly the same pattern, i.e. missing carry-over effects for the read speech task.

[20] As reported in Chapter 5.1.1, the less talented group showed on average more divergence and maintenance. Nevertheless, their equally perturbed self-consistency values showed that their way of speaking did change, meaning that they might have tried to converge but simply failed. See Chapter 7.1.4 for more details.

7.1.4 Attentional Selection in Convergence

Phonetic talent has proved to be a decisive factor for accommodation in dialog, with talented speakers displaying strong *convergence*, and less talented speakers on average more *divergence*. A closer look at the self-consistency values for both groups, however, revealed that they were equally perturbed (see **Hypothesis A4** in Chapter 5.1.2). Lower values were expected on the part of the talented NNS, since they also displayed higher convergence, so they should have theoretically “given up” more of their own speaking style. The less talented NNS were expected to show less altered self-consistency values since they converged on average significantly less within the dialogs. This did not turn out to be the case, though.

If the less phonetically-talented subjects did not converge but still showed the same pattern of perturbed self-consistency, their speech must have passed through a similar process as the talented speakers’ did, but with a different outcome. Considering the possibility presented in Figure 7.1 in Chapter 7.1.2 that all accommodation, be it positive or negative, might in fact consist of a mixture of divergence and convergence, we might say that the less talented NNS simply showed little convergence and a lot of divergence caused by other personality, psychological and social factors. The assumption, however, that such a mixture of other individual and situational factors accounted for the behavior of *all* ten nonnative subjects in this group is fairly unlikely. The reasons behind the pattern of negative convergence paired with lower self-consistency values might thus be found within the process sequence of the convergence mechanism, where talent seems to play an important role.

As Goldinger suggests, perception proceeds by activating all memory traces similar to the currently experienced one, depending on their degree of resemblance [Gol98] (see Chapter 2.1.2). A first possible limiting factor for this claim was presented in the two models of attention redirection in Chapter 2.2.6 on the status of talent in exemplar-based models (see Figure 2.7 and Figure 2.8). The problem for some (less talented) speakers might start as early as at the stage of recognition and storage, where not enough exemplars (or not sufficiently richly indexed ones) are stored in memory, and therefore both the *quantity* and *quality* of the activated exemplars might be decidedly different in the two talent groups. As Treisman noted, without attention only the presence of bits and pieces of information is recorded, but not where they came from or how they belong together [Tre99][21]. A deficit at this stage could correspond to the aptitude complex *phonemic coding ability*, proposed by Skehan [Ske03] and described in more detail in Chapter 1.4.2. The *less talented* speakers might thus have faced a situation where their activated exemplar pools did not provide them with the necessary acoustic information for convergence. Since the exemplar pools themselves *were* activated, the amount of exemplar choices the speaker had increased; and this could have led to a considerable *overspecification* of exemplars. Finding her- or himself in such an exemplar "overload" situation might have contributed to selecting less suitable exemplars[22]. The difference in talent could thus be traced back to deficits in **exemplar memory** caused by insufficient acoustic indexing during **storage**.

A second location for the direct influence of *talent* on the convergence mechanism is better access to the most similar exemplar pools. People pay attention to linguistic input in differing degrees. They are said to thereby focus on events which seem to be "most informative", as Pierrehumbert put it [Pie06]. A difference in this attentional directing toward the essential (acoustic) features in the signal in the *talented* group might bring forward fast access to relevant exemplar sets, while hindering the same process in the group of *less talented* speakers. If a speaker recognizes the incoming signal as **British English** (and this recognition must by no means be fully conscious), it should start a top-down modulation of the speech signal. As laid out in the Adaptive Resonance Theory (ART, see Chapter 2.2.2), Grossberg proposed that bottom-up processes activate a top-down

[21] See Chapter 2.2.1.

[22] This is comparable to the "mixed" output situation lowering the chances of a good match described by Hintzman [Hin86], see Chapter 2.1.3.

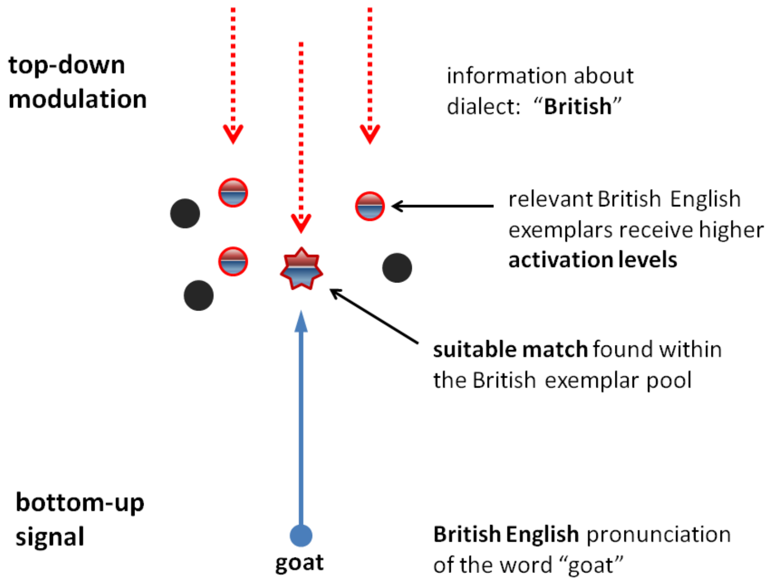


Figure 7.4: Scenario no.1 of the possible interplay of bottom-up information from the speech signal and top-down modulation for the label "accent/dialect type", leading to the activation of the relevant exemplar pool(s) in the dialog and facilitating the selection of a suitable exemplar. Model based partially on the ART by Grossberg [Gro03].

modulation, with the two together forming a resonant state [Gro03]. Ideally, items labelled as "British" should therefore set in motion top-down modulatory processes that alter the reaction threshold of cells with this label. In this case, this means the exemplar pool with British English accented tokens (see Figure 7.4). The unfavorable disturbed scenario would involve bottom-up information meeting no top-down feedback, which would in consequence inhibit the firing of cells and the desired British English exemplars would not be assigned higher activation levels. This mismatch of bottom-up information and top-down response could occur in speakers with low phonetic aptitude (Figure 7.5). Directly linked to the presented attentional process might be the *central executive*, a working memory component proposed by Baddeley and Gathercole [GB01, Bad03] (see Chapter 2.1.2). The central executive has the role of a managing device, directing attention to either a stimulus and its relevant features, or to stored information in

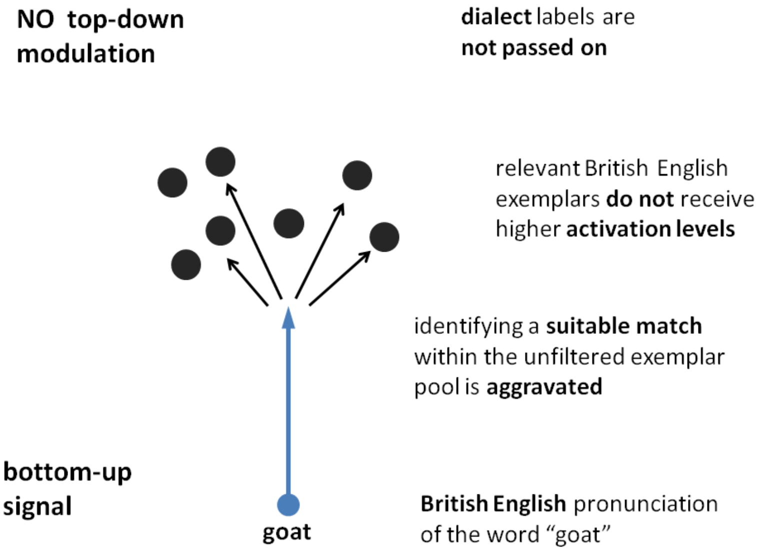

NO top-down modulation
dialect labels are not passed on
relevant British English exemplars do not receive higher activation levels
identifying a suitable match within the unfiltered exemplar pool is aggravated
bottom-up signal
goat
British English pronunciation of the word "goat"

Figure 7.5: Scenario no.2. Here, the bottom-up information from the speech signal is not met with top-down modulation for the label "accent/dialect type", preventing the activation of the relevant exemplar pool(s) in the dialog and thereby hampering the selection of a suitable exemplar. Model based partially on the ART by Grossberg [Gro03].

long-term memory (LTM). Individual differences in the functioning of the central executive might therefore cause problems either in the *detection* of those essential features necessary for the proper activation of similar memory traces, or, given that the features received attention and were accurately recognized, in the *activation process* of similar exemplars in LTM (and in holding them in this active state). Another theory assigns the crucial influence at this stage to the hippocampus and the MTLC[23], the former of which is said to perform direct recall decisions, and the latter supposedly being involved in "familiarity judgments", which might correspond to the activation of all similar exemplars in memory[24] (Goldinger and Norman & O'Reilly [Gol07, NO03], see Chapter 2.2.7). Episodic memory is also assumed to perform a delimiting function in the interpretation of incoming

[23] medial temporal lobe cortex.

[24] e.g., all American English exemplars of a certain word.

signals [Gol07], which might be a potential source of further complications for less talented subjects. Talented speakers proved to have no difficulties in accessing exemplar pools of appropriate English words for both dialects of English (British and American). Blackledge stated that pronunciation “is fluid and skillfully deployed by individual speakers” [BP01, 244]. We might add that it is skillfully deployed by those who possess full access to the appropriate exemplar pools – and this goes probably hand in hand with *talent*. For speakers who do not show a high phonetic aptitude, several exemplar clouds might be activated, thereby increasing the total number of items to choose from. This process, however, might have been inaccurate in that the wrong clouds have received higher activation levels, or such clouds that may have only partially met the conditions set by the incoming speech of the conversational partner. In such a case, again, *less talented* subjects would choose the wrong exemplars for the given context and show no (or only small) convergence. The crucial difference between phonetically talented and less talented speakers would therefore lie within the **processing component** of working memory which controls the information flow from perception to memory access.

A third possible spot for talent to come into play lies within the **selection process** of suitable exemplars. Assuming that the first two conditions are satisfied, i.e. the storage of exemplars proceeded correctly and situationally-appropriate exemplar pools have been activated after perceiving the dialog partner’s speech, difficulties could still arise within the retrieval process of those exemplars. Talented speakers can with great probability handle the selection of suitable exemplars simultaneously with other skilled tasks[25] without needing to allocate huge amounts of attention to the process. In Meyer and Kieras’ EPIC model[26] this means the selection process does not demand control processes [MK97]. In less talented speakers, though, the retrieval process might draw on such central processing capacities and require the

[25] as concentrating on the meaning of the incoming speech signal and formulating own responses.

[26] executive process interactive control, see Chapter 2.2.3.

allocation of attention to the task.

Norman and Shallice [NS86, 2] have listed situations in which deliberate attentional control might be needed (see Chapter 2.2.3). In this list we find, a.o. tasks that have a planning or decision-making component, those that have not been learned properly are new or difficult, or those that demand suppressing a strong habitual answer. The first task type might correspond to the decision the speakers make about the extent to which they want to sound American or British, given that they have the linguistic skills and talent to make such a choice. As reported earlier, some speakers appeared to have noticed that they converge fairly strongly to the accent they usually do not speak themselves, and tried to stop that tendency. This points to deliberate attentional control in order to contain the displayed convergence – in other words to influence how the speakers did *not* want to sound. Given that convergence is the default behavior, it seems reasonable to have attention activated as an "emergency brake" in cases where the convergence process is assessed as being too strong[27].

Norman and Shallice's second scenario [NS86] corresponds to storage difficulties, which were described earlier in this chapter. If an exemplar has not been properly learned, has been encountered for the first time or is especially difficult[28], its retrieval necessarily demands more attention. If there are no additional attentional resources available, the selection process is automatically disturbed. The last type of task involves the suppression of a strong habitual answer, which also requires a considerable amount of attention. This last possibility seems to be a potential influencing factor for the *less talented* subjects' behavior, as well. Whereas talented speakers do rather not appear to have any difficulties in suppressing the production of their usually-used "personal" exemplars and convergence can proceed undisturbed, speakers with less phonetic aptitude might automatically fall

[27] The usage of attentional processes to make those "higher order" choices is probably tied to a high talent level, since it demands a high degree of control that *less talented* speakers might not be endowed with.

[28] 'Difficult' here could mean, for instance, that the speaker does not have good access to the properties of an American accent, and she or he must put a lot of effort in finding the right American English exemplars.

back to choosing their own most frequent exemplars. These are not situationally colored accordingly and therefore do not constitute the *best choice* at the given moment.

The explanations presented here might shed more light onto the role **talent** plays as a modulating factor in convergence. The argument to split **attention** into *storage, processing* and *selection/retrieval* has been made for reasons of greater clarity in the description. However, it is straightforward that these processes overlap to a great extent and are mutually contingent. It seems feasible that the individual difference of *talent* is located within this larger network of attention and memory, where it comes into play at various processing stages. Attention seems particularly promising for holding the key to answering the remaining questions about phonetic talent in second language acquisition and usage.

The following section will be concerned with the discussion of the read speech results from the pre- and post-test.

7.2 Discussion of read speech results

In contrast to previous findings, such as those in Delvaux and Soquet's or Pardo's studies [DS07, Par06] (see Chapter 3.3.2), in this study no carry-over effects of convergence outside the dialog could be confirmed, neither for the summary part after the dialog, nor for the read speech pre- and post-test reported here. As reported in Chapter 6.1.1, **Hypothesis B1** was not verified, meaning that no significant accommodation was found in the comparison of the pre- and post-tests after the dialog with the respective native speaker. That is to say that, although our German speakers converged in the course of the dialog, their read speech remained to a large extent unaffected by the previous adaptation. This is confirmed by the decrease in convergence we reported after the switch from dialog to narrative style (Chapter 5.1.3 and Chapter 7.1.3). There was no correlation between the speakers' convergent behavior during the dialog session and the subsequent word list task

(**Hypothesis B2**). This is also in sharp contrast to Nielsen's findings [Nie07, Nie08] and the study of Abrego-Collier and colleagues [ACGSY11], who found convergence for read word repetition after the subjects had been exposed to a first-person narrative.

The mean match values for the crucial comparisons between the nonnative speakers and native speaker *T* in T1–T2[29] and native speaker *J* in J2b–J3[30] ranged from .74 to .77 and are, on average, slightly higher than matches within the dialog situations. Despite these generally high values for the matches between NNS and NS in read speech, no significant increase was found in the post-test. The mid to mid-high values can be explained by the usually much clearer and more careful speaking style in read speech, where the speaking rate is also more stable than in conversational speech and coarticulation at the word level is not an issue. The match value of two read speech items should therefore in general be higher than that of two dialog items from a conversation.

When the match values between the NNS and NS are compared to the self-consistency values of the NNS obtained for read speech (see Chapter 6.2), the self-consistency values prove to be higher, ranging from .83 up to .92. This, again, supports the claim that the consistency effect for a speaker usually is the stronger effect, as advocated by Cappella and Planalp [CP81] (compare Chapter 3.3.2). With regard to group differences, neither the list comparisons nor the self-consistency values yielded any significant effects for talent or gender (**Hypotheses B3, B4 and B5**). In the case of the self-consistency analysis, the two talent groups even showed remarkably similar patterns for the value changes between the lists.

Given that the self-consistency values of all nonnative speakers changed between the dialog sessions[31] but that the match values did not show significant convergence toward the respective dialog partners, it can be concluded that the changes in read speech were not *convergence* changes[32]. The measured changes go at least partially

[29] List 1 and List 2.

[30] List 2b and List 3.

[31] see Chapter 6.2 for the detailed results.

[32] Meaning that they cannot be classified as a positive shift toward the native speakers.

beyond simple random deviations from an assumed unperturbed or uninfluenced repetition of the word list, as it probably would have been had the subjects been simply asked to read the same list twice or three times sitting in total silence without any dialog in between. Even then the self-consistency values would never go up to reach 1, because two utterances are simply never identical. The values would instead remain in the range of high to very-high matches of close to 1.

Although it cannot be denied that the dialogs must have exerted some influence on the read speech of the nonnative speakers, judging from the results obtained in Chapter 6, this influence could rather be classified as *noise*. Assuming in exemplar theoretic terms that read speech will be labelled as such in our exemplar memory (see Chapter 2.1.2), those read speech exemplars will also bear different situational and contextual functions. A word in read speech will have completely different labels than the same word encountered in dialog speech, lacking a.o. such obvious labels as *speaker identity*, *accent* or *variety*[33]. Read speech exemplars are probably also much less often "updated" than dialog exemplars because in fact we have to introduce these "updates" ourselves, "manually" so to speak, by simply reading something. It is not stored away as automatically or immediately as it is in dialog speech. This will be discussed later in this chapter.

Considering that read speech lacks another crucial component – namely the dialogic function – our read speech exemplars "are" generally under no external "social pressure"[34] to directly accommodate to some target. When reading, the social motivation for convergence is simply not given. Read speech is also clearer and usually slower than dialog speech because it is not necessary to follow any conversational flow or adapt to a partner's pace. These features summed up lead to a greater de-

[33] Neglecting the rather rare occasions in which a text is being read out loud to an audience, where speaker identity and accent information would of course be given and stored. On the other hand, note that this speaking style also falls under the category of Bell's audience design ([Bel01, Bel07] in Chapter 1.1.2 and Chapter 1.1.3.), which is governed by a distinct set of rules and accordingly "labelled".

[34] Again, some situations may of course introduce that kind of pressure, as e.g. a foreign language exam where the candidate's task is to read a text out loud. These, however, are not standard usages of read speech and, in addition, there is no immediate "dialog partner" toward whom the convergence should be directed.

gree of *control* over our read speech.

As to why read speech might also not underlie an automatic alignment process[35] (to such an extent or of the same type) as dialog speech, the structure of exemplar memory needs to be considered. If we ask the question how many times a day we hear texts that are read out loud lacking any signs of spontaneity and without any audience-design character at all, and how often, by contrast, we hear elements of genuine dialogs or are involved in them ourselves, dialog speech clearly wins out. In adulthood, most read speech items are probably produced by us ourselves, while reading in silence. The dialog items by far outnumber the read speech exemplars, and they are stored away with the essential labels that are necessary from a re-usage point of view (compare Chapter 2.1 and 2.1.1). The more exemplars are encountered, the faster this update process is[36]. Read speech, however, cannot benefit from such a large number of new incoming items. Assuming such a dynamic usage-based account with exemplars featuring different labels also accordingly grouped together, read speech and dialog speech would not be mixed during the storage process in memory. This does not imply a radical and impermeable dissociation of both groups, but highlights that those groups of exemplars are accessible only by style-compliant exemplars when arriving from the *outside*. Incoming exemplars would then naturally flow into the *clouds* with maximum shared similarity, with speech style labels taken into account, as well (compare [Pie01]).

Read speech exemplars or exemplar groups might therefore be updated either *directly* by new incoming *read speech items* or, indirectly as a two-stage process in which conversational speech exemplars function as a *relais box*. Exemplars from the neighboring dialog speech area might percolate to the read speech area and slightly “stir up” the read speech contents, thereby introducing temporary instability to the cloud, which could affect both storage and retrieval processes (see Figure 7.6 for a graphical representation).

[35] as proposed by Pickering and Garrod [PG04b, PG05, PG09].

[36] Frequency effects, see Chapter 2.1.3.

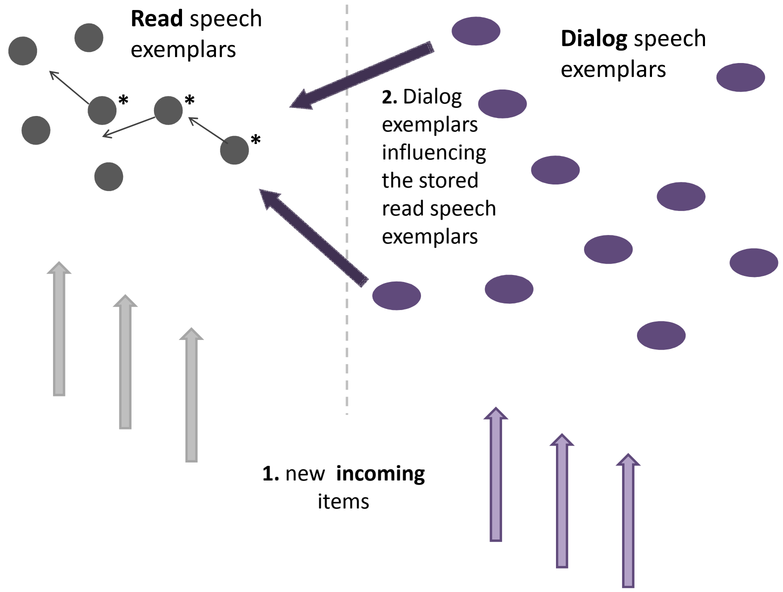

Read speech exemplars
*
*
*
Dialog speech exemplars
2. Dialog exemplars influencing the stored read speech exemplars
1. new **incoming** items

Figure 7.6: A graphical model of the interaction between dialog and read speech exemplars in memory. (1) Incoming exemplars are fitted into the appropriate areas/categories, according to their labels (read, conversational, male, female, etc.). (2) More frequent and recent dialog exemplars might influence the neighboring read exemplars and introduce shifts in the resting activation levels, marked by the asterisks and arrows.

This might be one of the explanations for the more perturbed self-consistency values the nonnative speakers showed in the pre- and post-test. The preceding dialog led to an update of dialog exemplars and could therefore have introduced perturbations in the neighboring read speech exemplar area/category and shifted the existing activation levels (compare [Joh97, Pie01] in Chapter 2.1), thereby causing more variability during the subsequent exemplar retrieval.

Since read speech is much more controlled than dialog speech and is to a large degree independent of external factors, the specification process of how we eventually pronounce a word is probably more dependent on our own choices, since we also have more time to make these choices consciously (compare Chapter 2.2.3 on the automaticity in language). It might thus also be conceivable that we choose to update our read speech exemplars by using exemplars from our dialog speech repository and integrating them into our read speech pool (see Figure 7.6). This process, however, supposedly proceeds over a longer period of time, comparable to lenition or deletion processes in language[37]. It apparently does not happen in real-time, as it is the case in dialog speech.

The supposed special character of read speech elaborated on above, and the differences between *imitation* on the one hand, and *convergence*[38] on the other hand, might provide the answer as to why some studies found “convergence” effects for read speech. In fact, both Nielsen [Nie07, Nie08] and Abrego-Collier and colleagues [ACGSY11] tested only isolated words throughout their experimental sessions, with the difference that the latter study used a first-person-narrative during the training phase. Nevertheless, both studies clearly exposed the subjects only to read speech. At no point was there conversational speech involved. Following the predictions of the exemplar theoretic model laid out above (Figure 7.6), it is very possible that in those two cases newly-encountered read speech exemplars from the training sessions entered the exemplar memory and were subsequently re-used. This was not the case in the present work because of the mismatch between speech styles.

Since there was no direct communicative interaction but only more controlled read speech involved in both Nielsen’s and Abrego-Collier’s study, it is advisable to exercise caution in assigning the same meaning or assuming the same underlying mechanisms to such data and to data obtained from real interactive dialogs. The present results on dialog and read speech offer support for considering differing speech styles as well as testing modes with varying degrees of control (isolated word repetition vs. (quasi) spontaneous dialog) as probably having divergent storage and recall mechanisms. These might in the future turn out to belong under different headings, and not under the single heading *convergence*.

[37] which make their way into language dependent, o.a., on the frequency of usage (compare, e.g. Bybee [BH01] and Chapter 2.1.3).

[38] compare Chapter 2.2 and Chapter 3.3.2.

Chapter 8

Conclusion and outlook

8.1 Conclusion

Summarizing the findings about the influence of talent on nonnative convergence leads to the assumption that convergence is the *default tendency* for natural communicative interaction, with talent probably influencing its core mechanisms, causing significant differences to arise between phonetically talented and less talented speakers. Apart from talent, convergence is also very likely influenced by other individual factors, such as personality and psychological features, the need for social approval and other social and contextual factors determining the communicative situation. Gender, on the other hand, was not shown to have any influence on the level of convergence displayed in the dialog task, nor did it modulate the behavior in the read speech tasks. The occurrence of convergence has proved to be strongly tied to the existence of a direct communicative interaction, since its effects could neither be confirmed for the summary part of the main experimental task, nor for the read speech pre- and post-test. Thus, no carry-over to speech styles other than dialog could be found. This implies that studies investigating spontaneous speech phenomena should also take place in spontaneous or at least quasi-spontaneous dialog environments rather than in carefully constructed (and therefore rather unnatural[1]) word repetition or reading tasks. It seems equally important to draw a strict terminological line between *imitation* and *convergence*, the former of which is a fully conscious and controlled process, while the latter is only a partially consciously or largely subconscious process.

It has been proposed that convergence at its core is a biologically founded drive for more synchrony, which, however, is delimited by many endo- and exogenous factors inherent to every dialogic interaction. The outcome is a dynamic interplay of factors enhancing and limiting it, thereby introducing divergence. Both automatic and controlled mechanisms play a role and this presupposes a *hybrid model* of convergence. Although divergence is more likely to be based on conscious decisions, the speaker might also become aware of the *outcome* of the rather subconscious

[1] i.e., unsuitable for capturing *naturally* occurring convergence.

convergence processes[2] and deliberately decide to influence them.

The following two sections will focus on presenting possible practical applications for the above findings and will discuss future research directions.

8.2 Practical applications

Phonetic convergence occurs in natural speech dialogs and is tied to the communicative demands[3] inherent in such a speech style and in the social component present – namely the orientation toward the conversational partner. This orientation toward the interacting speaker seems to be an essential part of communicative success. It bears an emotional component as well, since the lack of even minimal or partial[4] adaptation (which is the natural and automatic tendency) might invoke negative feelings, such as unfriendliness, a lack of interest from the interactant, a feeling of not being treated seriously, or even the impression of being rejected. The best case scenario would be an intangible feeling of something being “strange” in the conversation. This phenomenon that happens naturally in everyday life is still a missing feature in computer-generated speech, and this is the reason for the certain uncomfortableness with which we react to dialog systems imitating human-human interaction.

The incorporation of a convergence mechanism into synthetic speech might therefore make all computer **dialog systems** much more comfortable and user-friendly. The possible applications here range from telephone dialog systems to navigation systems for cars or pedestrians and could reduce the negative reactions and feelings of anxiety usually related to the handling of such devices[5]. Oviatt, Darves

[2] Also, social factors and, for instance, the ratings of attractiveness and friendliness exert their influence subconsciously, see Figure 7.3 in Chapter 7.1.2.

[3] i.e., the need to understand and to be understood correctly, as expressed, e.g., in more careful or clear speech or the decision to use simplified vocabulary.

[4] Partial = ocurring for some features at least.

[5] Listening, for instance, to the same monotonous voice of your car navigation system for several hours in a row, usually starts to be perceived as a source of mild to severe annoyance before it is eventually shut down. This problem could certainly be changed by a dynamically-adaptive component.

& Coulston [ODC04] were able to show that a responsive human-computer interface which accommodates some features of the interacting person's speech[6] goes beyond simple user-friendliness and even leads to enhanced learning effects. This has implications for the whole market of **learning software** with a speech output component, which might altogether achieve better effects if an adaptive component reacting to individual differences of a speaker were to be built in. **Language learning** and **pronunciation training software** are amongst the most affected applications by the current lack of speaker-accommodating systems. Additional convergence mechanisms could help reduce the impersonal robotic aura of such learning systems and instead create a more natural, encouraging and also rewarding learning environment. An environment with such qualities is likely to increase learning effects for the user.

Another strand of research leads us to the **second language classroom** and the question of the specific talent for pronunciation. The identification of the exact mechanisms – for instance in *attention* and *memory* – underlying or connected to **phonetic talent** would provide a chance to explicitly address these problem areas while learning. This is far from suggesting that talent itself can be "practiced", but it might be possible to explicitly train certain attentional mechanisms for noticing prominent acoustic features and memory strategies for the storage and retrieval of phonetic information (after these features are identified). Well-directed training of such small subcomponents of talent, naturally only after the network of interactions and their roles are satisfactorily identified, might help students make small but nevertheless very valuable improvements to their pronunciation in a second language, thereby allowing them to enhance their communicative skills. A trivial-sounding but nonetheless essential prerequisite for applying such teaching methods in second language pronunciation practice would be an increased focus on the phonetic component in curricula. This implies first of all more time assigned to practice sessions for this skill, as well as a smaller teacher to student ratio. Just as convergence mechanisms require a person-to-person setting to achieve an individual fine-tuning

[6] The subjects were children interacting with a computer animated persona.

of acoustic parameters, phonetically less gifted second language learners certainly need more individual practice and targeted training, and this is conceivable only in smaller learning groups.

8.3 Future directions

Even though phonetic convergence has been investigated to a fairly great extent, with rising interest especially in recent years, there still remain many open questions regarding its exact mechanisms. While it has become rather obvious that convergence occurs in every dialogic interaction[7], we still need to gain much more insight into the detailed neuro- and psycholinguistic mechanisms governing convergence and all the possible factors influencing it, since it is surely not a purely automatic nor a totally uncontrolled process.

Our knowledge about the components or dimensions that are actually being accommodated is still very sparse. For instance, there is no satisfactory insight into the role of timing information. This study focused solely on amplitude envelopes, which are a measure of spectral similarity[8]. Incorporating timing information, e.g. based on a DTW analysis[9], as a second measurement might complete the picture of phonetic convergence in native-nonnative interactions we have so far. Since it is suggested in usage-based accounts that memory for linguistic events is of a spectro-*temporal* nature (compare Wade et al. [WDS$^+$10] in Chapter 2.1.2) rather than just including spectral features alone, temporal features are just as crucial a component in the storage mechanism of linguistic input. A joined analysis of both dimensions might thus clarify whether one of them is easier or faster converged to than the other, which in turn might also form a new basis for more sophisticated models

[7] Be it positive or negative or uni- vs. bilateral.

[8] The final *match values* also reflect timing mismatches in envelope alignment, so the timing dimension is partially included in the measurement, albeit not as a separable value to be further analyzed.

[9] *Dynamic Time Warping*, a method designed to compute the distances between well matching spectral points of two waves (sequences), is used mainly in speech recognition.

of the mechanisms underlying exemplar storage and retrieval. Controlling for frequency of occurrence of the target words and including this as an influencing factor in convergence experiments is another important future goal. A partial re-analysis of the current study – for subjects for which a sufficient number of low and high frequency words can be extracted, so that a comparison is possible – is also conceivable for this purpose.

Considering especially nonnative convergence, we still need to obtain a more complete picture of prosodic accommodation in such encounters. An analysis of more fine-grained features, such as the convergence of the types and timing of pitch accents[10], stress placement, or the adaptation of F0 range and mean, could provide a clearer picture of those elements that are readily re-used in convergence, on the one hand, and those which pose difficulty, on the other hand.

Another essential direction in convergence which still needs more in-depth investigation is the influence of personality and psychological features, as they might deliver the explanation for the remaining between-subject variance in convergence data, not only with respect to communication in the L2 but mostly also in the L1.

For oral communication in the L2, for instance, an extraverted nature seems to have advantages[11]. Hu and Reiterer [HR09] have measured moderate correlations of phonetic aptitude with conscientiousness and agreableness, but not with extraversion, which they explained with task design differences. In a more communicatively oriented task (such as a spontaneous dialog, for instance), the relation of pronunciation and extraversion could be assumed to show more significant results, hence possibly also for alignment in such a dialog. Language anxiety is another individual factor which might activate compensation by increased effort on the learner side, as Hu and Reiterer point out. The cause and amount of this compensation is attributed to another personality construct, the *behavioral withdrawal-approach system* (BIS/BAS) proposed by Gray [Gra72, HR09]. This system deals with a per-

[10] As, for instance, with ToBI labels and PaintE parameters [MC98].

[11] However, as Skehan admits [Ske03], introverts can benefit in other L2 related tasks, which could partially account for the conflicting evidence in this field.

son's behavior inhibition and activation systems and their interplay. Yet another possible influential trait might be the cognitive styles of learners – field dependence (FD) and field independence (FI) (see Dörnyei and Skehan [DS03]). If we speculate about this dichotomy in relation to pronunciation accommodation, we might assume that field independents will have an advantage due to their analytic skills, which are probably crucial for directing attention to the detailed acoustic information in the speaking partner's message.

There also seems to be support for a correlation between empathy and pronunciation. Hu and Reiterer's study (based on E-Scale[12]) points to a significant correlation of the factor 'readiness for empathy' and pronunciation talent. Empathy could intuitively also hold some explanatory power for the magnitude of accommodation in an L2-dialog, since it certainly facilitates the establishment of a "common ground", an idea put forward by Pickering and Garrod [PG04b] (see Chapter 1.3.2). One could imagine that the suggested impact of readiness for empathy on pronunciation might translate as well into a greater approximation of phonetic features between two dialog partners.

Taking into account the multitude of aspects and probable outcomes, it might be best to look for combined effects of personality traits and other individual differences variables, since a single individual trait might not have the explanatory power to account for all individual differences in convergence. A combination of certain traits, however, might give us more answers as to the communicative mechanisms and strategies of speakers – and many questions as to how accommodation works, would be answered as well.

A very intriguing and so far neglected experimental method (probably due to the technical complexity of the set-up) is the combination of acoustic measurements with measurements of neural activity. Building an experimental set-up for a spontaneous or quasi-spontaneous dialog involving, for instance, an *EEG*, would certainly

[12] A German empathy assessment questionnaire with the two major dimensions: readiness for empathy and social concern [HR09, 118].

pose some technical and also analysis-related difficulties[13]. It would, nevertheless, allow for a far more precise look into the timing of convergence processes. The application of neuroimaging techniques, such as the fMRI[14], would allow us to answer the question of which areas are involved in the storage and immediate re-usage of linguistic material. It would also provide answers as to the possible activation of areas assigned to various memory types or attentional components (see Chapter 2.1.2).

In order to find out more about how "perturbed" read speech (=speech with rather low self-consistency values) is processed in the brain, an experiment with the involvement of the fMRI technique could be designed. The current study where no convergence but the aforementioned decreased self-consistency values were found, could be repeated on a smaller scale. The activation patterns in the brain could be tracked while performing the pre- and post-test with read speech presented here, leaving out the measurement of the dialog phase. Comparing the obtained activation patterns in different areas with their strength might shed light on what exactly is happening in a person's brain after they have had a "perturbing" conversation and their speech repertoire is somewhat out of its normal order. Using the same simplified design with ERP responses could provide insight into any timing or response type variability between the reactions in the pre- and post-test.

Having found that dialog speech does not directly influence read speech, it would be worthwhile to test whether the opposite is true – namely, can read speech influence dialog speech? If not, then this might suggest that the storage and retrieval processes for those speech styles indeed run separately and the linguistic input is accordingly labelled. If read speech, on the other hand, can cause convergence in a dialog, then the processes probably proceed in a one-way manner, with no bi-lateral links in between.

[13] Regarding the time-locking of the ERP (event related potentials) responses to the relevant events, so that they can be attributed to specific instances of convergence that occurred within the dialog.

[14] Functional Magnetic Resonance Imaging.

A last potential direction in the investigation of phonetic convergence with a focus on nonnative interactions, would lead us towards a more elaborate analysis of the role of attentional processes in accommodation and their relation to the individual difference of talent. This could help us answer how big an impact attention to form as opposed to attention to meaning has in the storage of specific exemplars, and also in the processing and retrieval stages. Moreover, it could give us a lead on whether the key to pronunciation talent really lies in an inherent attentional component.

Bibliography

[ABB$^+$01] A. Anderson, M. Bader, E. Bard, E. Boyle, M. G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, S. H. Thompson, and R. Weinert. The hcrc map task corpus. *Language and Speech*, 34(4):351–366, 2001.

[ACGSY11] C. Abrego-Collier, J. Grove, M. Sonderegger, and A. C. L. Yu. Effects of speaker evaluation on phonetic convergence. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 192–195. 2011.

[AS68] R. C. Atkinson and R. M. Shiffrin. Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence, editors, *The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 2, pages 89–195. Academic Press, New York, 1968.

[AST06] K. Abbot-Smith and M. Tomasello. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23:275–290, 2006.

[Bab09] M. E. Babel. *Phonetic and Social Selectivity in Speech Accommodation*. PhD thesis, University of California, Berkeley, 2009.

[Bad86] A. Baddeley. *Working memory*. Oxford Univ. Press, Oxford, 1986.

[Bad03] A. Baddeley. Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003.

[Bar88] R. Bartsch. *Norms of language: Theoretical and practical aspects*. Longman, London, 1. publ. edition, 1988.

[BAS$^+$00] E. G. Bard, A. H. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon, and A. Newlands. Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42:1–22, 2000.

[BBC$^+$07] A. R. Bradlow, R. E. Baker, A. Choi, M. Kim, and K. J. van Engen. The wildcat corpus of native- and foreign-accented english. *Journal of the Acoustical Society of America*, 121(5):3072, 2007.

[BC96] S. E. Brennan and H. H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, (22):1482–1493, 1996.

[Bel99] A. Bell. Styling the other to define the self: A study in new zealand identity making. *Journal of Sociolinguistics*, 3(4):523–541, 1999.

[Bel01] A. Bell. Back in style: reworking audience design. In P. Eckert and J. R. Rickford, editors, *Style and sociolinguistic variation*, pages 139–169. Cambridge Univ. Press, Cambridge, 2001.

[Bel07] A. Bell. Style in dialogue: Bakhtin and sociolinguistic theory. In R. Bayley and C. Lucas, editors, *Sociolinguistic variation: Theories, methods, and applications*, pages 90–109. Cambride Univ. Press, Cambridge, 2007.

[BG77] R. Y. Bourhis and H. Giles. The language of intergroup distinctiveness. In H. Giles, editor, *Language, Ethnicity and Intergroup Relations*, pages 119–135. Academic Press, London, 1977.

[BGLT79] R. Y. Bourhis, H. Giles, J. P. Leyens, and H. Tajfel. Psycholinguistic distinctiveness: Language divergence in belgium. In H. Giles and R. N. St Clair, editors, *Language and Social Psychology*, pages 158–185. Blackwell, Oxford, 1979.

[BH81] M. M. Bakhtin and M. Holquist. *The dialogic imagination: 4 essays*. Univ. of Texas Pr., Austin, 1981.

[BH01] J. Bybee and P. Hopper, editors. *Frequency and the emergence of linguistic structure*, volume 45 of *Typological studies in language*. Benjamins, Amsterdam, 2001.

[Bir06] D. Birdsong. Age and second language acquisition and processing: A selective overview. *Language Learning*, 56:9–49, 2006.

[BL10] G. Bailly and A. Lelong. Speech dominoes and phonetic convergence. In *Proceedings of Interspeech*, pages 1153–1156, Tokio (Japan), 2010. ISCA.

[Bla49] J. W. Black. Loudness of speaking: the effect of heard stimuli on spoken responses. *Journal of Experimental Psychology*, 39(3):311–315, 1949.

[Blo07] D. Block. *Second language identities*. Continuum, London, 2007.

[BMH10] S. Brouwer, H. Mitterer, and F. Huettig. Shadowing reduced speech and alignment. *The Journal of the Acoustical Society of America*, 128(1):EL32–EL37, 2010.

[Bod06] R. Bod. Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, 23:291–320, 2006.

[Bou91] R. Y. Bourhis. Organizational communication and accommodation: Toward some conceptual and empirical links. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*, Studies in emotion and social interaction, pages 270–304. Cambridge Univ. Press, Cambridge, 1991.

[BP01] A. Blackledge and A. Pavlenko. Negotiation of identities in multilingual contexts. *International Journal of Bilingualism*, 5(3):243–257, 2001.

[BPC00] H. P. Branigan, M. J. Pickering, and A. A. Cleland. Syntactic coordination in dialogue. *Cognition*, 75:B13–B25, 2000.

[BPS95] T. Bongaerts, B. Planken, and E. Schils. Can late starters attain a native accent in foreign language? a test of the critical period hypothesis. In D. Singleton and Z. Lengyel, editors, *The age factor in second language acquisition*, Multilingual matters, pages 30–50. Multilingual Matters, Clevedon, 1995.

[Bro58] D. E. Broadbent. *Perception and communication*. Pergamon Press, London, 1958.

[Bro84] D. E. Broadbent. The maltese cross: A new simplistic model for memory. *Behavioral and Brain Sciences*, 7:55–68, 1984.

[But30] S. Butterworth. On the theory of filter amplifiers. *Experimental Wireless and the wireless engineer*, (7):536–541, 1930.

[Byb02] J. Bybee. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14:261–290, 2002.

[Byb06] J. Bybee. From usage to grammar: the mind's response to repetition. *Language*, 82(4):711–733, 2006.

[Car81] J. B. Carroll. Twenty-five years of research on foreign language aptitude. In K. C. Diller, editor, *Individual differences and universals in language learning aptitude*, pages 83–118. Newbury House, Rowley, Mass., 1981.

[CES$^+$05] N. Cowan, E. M. Elliot, J. S. Saults, C. C. Morey, S. Mattox, A. Hismjatullina, and A. R. A. Conway. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1):42–100, 2005.

[CJ97] N. Coupland and A. Jaworski, editors. *Sociolinguistics: A reader and coursebook*. Modern linguistics series. Macmillan [u.a.], Basingstoke, 1997.

[Cla07] K. Claßen. *Prosodische und dysprosodische Variation linguistischer und paralinguistischer Funktionen im spontansprachlichen Dialog*. PhD thesis, IMS, Univ. Stuttgart, Stuttgart, 2007.

[CM85] J. Cheesman and P. M. Merikle. Word recognition and consciousness. In T. G. Waller and G. E. MacKinnon, editors, *Reading research: advances in theory and practice*. Academic Press, Orlando, 1985.

[Cou01] N. Coupland. Language, situation, and the realtional self: theorizing dialect-style in sociolinguistics. In P. Eckert and J. R. Rickford, editors, *Style and sociolinguistic variation*, pages 185–210. Cambridge Univ. Press, Cambridge, 2001.

[Cou07] N. Coupland. *Style: Language variation and identity*. Cambridge Univ. Press, Cambridge, 2007.

[Cow88] N. Cowan. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, 104(2):163–191, 1988.

[CP81] J. N. Cappella and S. Planalp. Talk and silence sequences in informal conversations iii: Interspeaker influence. *Human Communication Research*, 7(2):117–132, 1981.

[CS59] J. B. Carroll and S. M. Sapon. *Modern language aptitude test (MLAT)*. Psychological Corporation, San Antonio, 1959.

[CWG86] H. H. Clark and D. Wilkes-Gibbs. Referring as a colaborative process. *Cognition*, 22:1–39, 1986.

[DÏ0] Z. Dörnyei. The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro, editor, *Continuum companion to second language acquisition*, pages 247–267. Continuum, London, 2010.

[Dav03] A. Davies. *The native speaker: Myth and reality*. Multilingual Matters, Clevedon, 2003.

[DD63] J. A. Deutsch and D. Deutsch. Attention, some theoretical considerations. *Psychological Review*, 70:80–90, 1963.

[DH90] B. Davies and R. Harré. Positioning: The discursive production of selves. *Journal for the Theory of Social Behavior*, (20):43–63, 1990.

[DR09] G. Dogil and S. M. Reiterer, editors. *Language Talent and Brain Activity*, volume 1 of *Trends in Applied Linguistics*. De Gruyter, Berlin, 1. edition, 2009.

[DS03] Z. Dörnyei and P. Skehan. Individual differences in second language learning. In C. Doughty and M. H. Long, editors, *The handbook of second language acquisition*, volume 14 of *Blackwell handbooks in linguistics*, pages 589–630. Blackwell, Malden, Mass., 2003.

[DS07] V. Delvaux and A. Soquet. Inducing imitative phonetic variation in the laboratory. In *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 369–372, Saarbrücken, 2007.

[Ell85] R. Ellis. *Understanding second language acquisition*. Oxford Univ. Press, Oxford u. a., 1985.

[Ell03] D. Ellis. Dynamic time warp (dtw) in matlab, 2003.

[Far94] M. Farah. Visual perception and visual awareness: A tutorial review. In C. Umiltà and M. Moscovitch, editors, *Attention and performance XV*, A Bradford book, pages 37–76. MIT Press, Cambridge, Mass., London, 1994.

[FH87] C. Fowler and J. Housum. Talkers' signaling ”new” and ”old” words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26:489–504, 1987.

[Fle95] J. E. Flege. Second language speech learning: theory, findings and problems. In W. Strange, editor, *Speech perception and linguistic experience: theoretical and methodological issues*, pages 229–273. York Press, Timonium, 1995.

[FTD$^+$89] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16(03):477, 1989.

[FYKL99] J. E. Flege, G. Yeni-Komshian, and H. Liu. Age constraints on second language acquisition. *Journal of Memory and Language*, 41:78–104, 1999.

[GA87] S. C. Garrod and A. Anderson. Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*, 27:181–218, 1987.

[GAES80] A. Guiora, W. Acton, R. Erard, and F. Strickland. The effects of benzodiazepine (valium) on permeability of language ego boundaries. *Language Learning*, 30(2):351–363, 1980.

[GB82] H. Giles and J. Byrne. An intergroup approach to second language acquisition. *Journal of Multilingual and Multicultural Development*, 3(1):17–41, 1982.

[GB01] S. E. Gathercole and A. Baddeley. *Working memory and language*. Essays in cognitive psychology. Psychology Press, Hove, repr. in paperback. edition, 2001.

[GBC97] S. Grossberg, I. Boardman, and M. A. Cohen. Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 23:481–503, 1997.

[GBD72] A. Guiora, R. Brannon, and C. Dull. Empathy and second language learning. *Language Learning*, 22:111–130, 1972.

[GBHB$^+$72] A. Guiora, B. Beit-Hallahmi, R. Brannon, C. Dull, and T. Scovel. The effects of experimentally induced changes in ego states on pronunciation ability in a second language: An exploratory study. *Comprehensive Psychiatry*, 13(5):421–428, 1972.

[GCC91a] H. Giles, J. Coupland, and N. Coupland, editors. *Contexts of accommodation: Developments in applied sociolinguistics*. Studies in emotion and social interaction. Cambridge Univ. Press, Cambridge, 1991.

[GCC91b] H. Giles, N. Coupland, and J. Coupland. Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*, Studies in emotion and social interaction, pages 1–68. Cambridge Univ. Press, Cambridge, 1991.

[GDM92] R. C. Gardner, J. B. Day, and P. D. MacIntyre. Integrative motivation, induced anxiety, and language learning in a controlled environment. *Studies in Second Language Acquisition*, 14:197–214, 1992.

[GG02] S. W. Gregory and T. J. Gallagher. Spectral analysis of candidates' nonverbal communication: Predicting u.s. presidential election outcomes. *Social Psychology Quaterly*, 65(3):298–308, 2002.

[GGJ+95] C. Gallois, H. Giles, E. Jones, A. Cargile, and H. Ota. Accommodating intercultural encounters: Elaborations and extensions. In R. L. Wiseman, editor, *Intercultural communication theory*, volume 19 of *International & intercultural communication annual series*, pages 115–147. Sage, Thousand Oaks, 1995.

[Gil73] H. Giles. Accent mobility: A model and some data. *Anthropological Linguistics*, 15:87–105, 1973.

[Gil01] H. Giles. Couplandia and beyond. In P. Eckert and J. R. Rickford, editors, *Style and sociolinguistic variation*, pages 211–219. Cambridge Univ. Press, Cambridge, 2001.

[GL72] R. C. Gardner and W. E. Lambert. *Attitudes and motivation in second language learning*. Newbury House, Rowley, Mass., 1972.

[GM91] R. C. Gardner and P. D. MacIntyre. An instrumental motivation language study: Who says it isn't effective? *Studies in Second Language Acquisition*, 13:57–72, 1991.

[GM00] S. Grossberg and C. W. Myers. The resonant dynamics of speech perception: Interword integration and duration-dependent backwards effects. *Psychological Review*, 4:735–767, 2000.

[GO06] H. Giles and T. Ogay. Communication accommodation theory. In B. B. Whaley and W. Samter, editors, *Explaining Communication: Contemporary theories and exemplars*, pages 293–310. Lawrence Erlbaum Assosiates, Mahwah, 2006.

[Gol96] S. Goldinger. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5):1166–1183, 1996.

[Gol98] S. Goldinger. Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105(2):251–279, 1998.

[Gol07] S. Goldinger. A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain and W.J. Barry, editors, *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 49–54. Saarbrücken, 2007.

[GP75] H. Giles and P. F. Powesland. *Speech style and social evaluation*, volume 7 of *European monographs in social psychology*. Acad. Press, London, 1975.

[GP97] H. Giles and F. P. Powesland. Accommodation theory. In N. Coupland and A. Jaworski, editors, *Sociolinguistics: A reader and coursebook*, Modern linguistics series, pages 232–239. Macmillan [u.a.], Basingstoke, 1997.

[GP07] S. Guion and E. Pederson. Investigating the role of attention in phonetic learning. In O.-S. Bohn and M. J. Munro, editors, *Language experience in second language speech learning: In honor of James Emil Flege*, Language learning and language teaching, pages 57–77. Benjamins, Amsterdam, 2007.

[Gra72] J. A. Gray. The psychophysiological basis of introversion-extraversion: A modification of eysenck's theory. In V. D. Nebylicyn and J. A. Gray, editors, *Biological bases of individual behavior*, pages 182–288. Acad. Press, New York, NY, 1972.

[Gre83] S. W. Gregory. A quantitative analysis of temporal symmetry in microsocial relations. *American Sociological Review*, 48:129–135, 1983.

[Gre86] S. W. Gregory. Social psychological implications of voice frequency correlations: Analyzing conversation partner adaption by computer. *Social Psychology Quaterly*, 49(3):237–246, 1986.

[Gre90] S. W. Gregory. Analysis of fundamental frequency reveals covariation in interview partners' speech. *Journal of Nonverbal Behavior*, 14(4):237–251, 1990.

[Gro03] S. Grossberg. Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31(3-4):423–445, 2003.

[Gro05] S. Grossberg. Linking attention to learning, expectation, competition, and consciousness. In L. Itti, G. Rees, and J. K. Tsotsos, editors, *Neurobiology of attention*, pages 652–662. Elsevier Academic Press, Amsterdam, Boston, 2005.

[GS79] H. Giles and P. M. Smith. Accommodation theory: Optimal levels of convergence. *Language and Social Psychology*, pages 45–65, 1979.

[Gui90] A. Guiora. A psychological theory of second language pronunciation. *Toegepaste Taalwetenschap in Artikelen*, 37:15–23, 1990.

[Gum82] J. J. Gumperz, editor. *Language and social identity*, volume 2 of *Studies in interactional sociolinguistics*. Cambridge Univ. Press, Cambridge, 1982.

[GW96] S. W. Gregory and S. Webster. A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology*, 70(6):1231–1240, 1996.

[Hal96] M. A. K. Halliday. Linguistic function and literary style: An inquiry into the language of william golding's ”the inheritors”. In J. J. Weber, editor, *The Stylistics Reader: From Roman Jakobson to the Present.*, pages 56–91. Arnold, London, 1996.

[Haw03] S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3-4):373–405, 2003.

[Haw10] S. Hawkins. Phonological features, auditory objects, and illusions. *Journal of Phonetics*, 38:60–89, 2010.

[Hel92] M. Heller. The politics of codeswitching and language choice. *Journal of Multilingual and Multicultural Development*, 13(1/2):123–142, 1992.

[Hel95] M. Heller. Language choice, social institutions, and symbolic domination. *Language in Society*, 24:373–405, 1995.

[HH83] J. C. Hansen and S. A. Hillyard. Effects of stimulation rate and attribute cueing on event related potentials during selective auditory attention. *Psychophysiology*, 21:394–405, 1983.

[Hin86] D. L. Hintzman. Schema abstraction in a multiple-trace memory model. *Psychological Review*, 93:411–428, 1986.

[Hol92] J. Holmes. *An Introduction to Sociolinguistics*. Learning about Language. Longman, London a.o., 2nd impr. edition, 1992.

[HR09] X. Hu and M. S. Reiterer. Personality and pronunciation talent in second language acquisition. In G. Dogil and S. M. Reiterer, editors, *Language Talent and Brain Activity*, Trends in Applied Linguistics, pages 97–129. de Gruyter, Berlin, 2009.

[Hym74] D. Hymes. Ways of speaking. In R. Bauman and J. Scherzer, editors, *Explorations in the ethnography of speaking*, pages 433–451. Cambridge Univ. Pr., London, 1974.

[Inc11] SPSS Inc. Spss 19 for windows, 2011.

[JCM03] D. Jones, B. Collins, and I. M. Mees. *Unpublished writings and correspondence*, volume 8. London, 2003.

[Jil09a] M. Jilka. Assessment of phonetic ability. In G. Dogil and S. M. Reiterer, editors, *Language Talent and Brain Activity*, volume 1 of *Trends in Applied Linguistics*, pages 17–66. de Gruyter, Berlin, 2009.

[Jil09b] M. Jilka. Talent and proficiency in language. In Grzegorz Dogil and Maria Susanne Reiterer, editors, *Language Talent and Brain Activity*, volume 1 of *Trends in Applied Linguistics*, pages 1–16. de Gruyter, Berlin, 2009.

[Joh97] K. Johnson. Speech perception without speaker normalization: An exemplar model. In K. Johnson and J. W. Mullennix, editors, *Talker variability in speech processing*, pages 145–165. Academic Press, San Diego, 1997.

[Joh06] K. Johnson. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499, 2006.

[Kah73] D. Kahneman. *Attention and effort*. Prentice Hall, Englewood Cliffs, NJ, 1973.

[KE06] S. J. A. Kelso and D. A. Engstrøm. *The complementary nature*. MIT Press, Cambridge, Mass, 2006.

[Kel97] S. J. A. Kelso. *Dynamic patterns: The self-organization of brain and behavior*. A Bradford book. MIT Press, Cambridge, Mass., paperback ed. edition, 1997.

[Kel09] S. J. A. Kelso. Coordination dynamics. In *Encyclopedia of complexity and systems science*, Springer-11651 /Dig. Serial], pages 1537–1565. Springer, New York, NY, 2009.

[KHB11] M. Kim, W. S. Horton, and A. R. Bradlow. Phonetic convergence in spontaneous conversations as a function of interlocutor language distance: Laboratory phonology. *Laboratory Phonology*, 2(1):125–156, 2011.

[KP04] R. M. Krauss and J. S. Pardo. Is alignment always the result of automatic priming? *Behavioral and Brain Sciences*, 27(2):203–204, 2004.

[Kra81] S. Krashen. *Second Language Acquisition and Second Language Learning*. Pergamon Press, 1981.

[LaB73] D. LaBerge. Attention and the measurement of perceptual learning. *Memory and Cognition*, 1:268–276, 1973.

[Lab01] W. Labov. The anatomy of style shifting. In P. Eckert and J. R. Rickford, editors, *Style and sociolinguistic variation*, pages 85–108. Cambridge Univ. Press, Cambridge, 2001.

[Lac97] F. Lacerda. Distributed memory representations generate the perceptual-magnet effect. *Journal of the Acoustical Society of America*, 1997.

[LCHD97] S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42, 1997.

[LDT99] P. C. Loizou, M. Dorman, and Z. Tu. On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, 106(4):2097–2103, 1999.

[Lev89] W. J. M. Levelt. *Speaking: From intention to articulation.* MIT Press, 1989.

[LHdFV04] N. Lavie, A. Hirst, J. W. de Fockert, and E. Viding. Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, 133:339–354, 2004.

[Lie67] P. Lieberman. *Intonation, perception and language*. Cambidge, MA: MIT Press, 1967.

[LMG77] K. Larsen, H. J. Martin, and H. Giles. Anticipated social cost and interpersonal accommodation. *Human Communication Research*, (3):303–308, 1977.

[LORC11] C. de Looze, C. Oertel, S. Rauzy, and N. Campbell. Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 1294–1297. 2011.

[LV02] S. J. Luck and S. P. Vecera. Attention. In H. Pashler and S. S. Stevens, editors, *Stevens' handbook of experimental psychology*, volume 1, pages 235–286. Wiley, New York, NY, 2002.

[Maj01] R. C. Major. *Foreign accent: The ontogeny and phylogeny of second language phonology*. Second language acquisition research. Theoretical and methodological issues. L. Erlbaum, Mahwah, NJ, 2001.

[Mar80] A. J. Marcel. Conscious and preconscious recognition of polysemous words: Locating the selective effects of prior verbal context. In R. S. Nickerson, editor, *Attention and performance VIII*, volume 8 of *Attention and performance*, pages 435–458. Erlbaum, Hillsdale, NJ, 1980.

[Mat09] M. W. Matlin. *Cognitive psychology*. Wiley, Hoboken, NJ, international student version, 7th edition, 2009.

[Mat11] Mathworks. Mathworks product documentation: Matlab r2011a documentation online, 2011.

[MC98] G. Möhler and A. Conkie. Parametric modeling of intonation using vector quantization. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 311–316, 1998.

[McL95] B. McLaughlin. Aptitude from an information-processing perspective. *Language Testing*, 12:370–387, 1995.

[McL07] C. T. McLennan. Challenges facing a complementary-systems approach to abstract and episodic speech perception. In J. Trouvain and W.J. Barry, editors, *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 67–70. Saarbrücken, 2007.

[MF98] A. Miyake and N. F. Friedman. Individual differences in second language proficiency: working memory as ”language aptitude”. In A. F. Healy and L. E. Bourne, editors, *Foreign language learning*, pages 339–364. L. Erlbaum, Mahwah, N.J., 1998.

[MK97] D. Meyer and D. Kieras. A computational theory of executive cognitive processes and multiple-task performance: Part 1. basic mechanisms. *Psychological Review*, 104:3–65, 1997.

[ML05] C. T. McLennan and P. A. Luce. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):306–321, 2005.

[MMC09] P. D. MacIntyre, S. P. MacKinnon, and R. Clément. The baby, the bathwater, and the future of language learning motivation research. In Z. Dörnyei and E. Ushioda, editors, *Motivation, language identity and the L2 self*, Second language acquisition, pages 43–65. Multilingual Matters, Buffalo, 2009.

[MMO95] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory,. *Psychological Review*, 102(3):419–457, 1995.

[Moy99] A. Moyer. Ultimate attainment in l2 phonology. *Studies in Second Language Acquisition*, 21:81–108, 1999.

[MW67] J. D. Matarazzo and A. N. Wiens. Interviewer influence on durations of interviewee silence. *Experimental Research in Personality*, 2:56–69, 1967.

[MWSW63] J. D. Matarazzo, M. Weitman, G. Saslow, and A. N. Wiens. Interviewer influence on durations of interviewee speech. *Verbal Learning and Verbal Behavior*, 1:451–458, 1963.

[Nat75a] M. Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804, 1975.

[Nat75b] M. Natale. Social desirability as related to convergence of temporal speech patterns. *Perceptual and Motor Skills*, 40:827–830, 1975.

[Nat76] M. Natale. Need for social approval as related to speech interruption in dyadic communication. *Perceptual and Motor Skills*, 42:455–458, 1976.

[NB75] D. A. Norman and D. G. Bobrow. On data-limited and resource-limited processes. *Cognitive Psychology*, 7:44–64, 1975.

[Neu79] G. Neufeld. Towards a theory of language learning aptitude. *Language Learning*, 29:227–241, 1979.

[Neu84] O. Neuman. Automatic processing: A review of recent findings and a plea for an old theory. In W. Prinz and A. Sanders, editors, *Cognition and motor processes*, pages 225–293. Springer, Berlin, 1984.

[Nie07] K. Nielsen. Implicit phonetic imitation is constrained by phonemic contrast. In J. Trouvain and W. J. Barry, editors, *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 1961–1964. Saarbrücken, 2007.

[Nie08] K. Nielsen. *Word-level and feature-level effects in phonetic imitation*. PhD thesis, University of California, Los Angeles, 2008.

[NNS02] L. L. Namy, C. L. Nygaard, and D. Sauerteig. Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21:422–432, 2002.

[NO03] K. A. Norman and R. C. O'Reilly. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4):611–646, 2003.

[Nor68] D. A. Norman. Towards a theory of memory and attention. *Psychological Review*, 75:522–536, 1968.

[NS86] D. A. Norman and T. Shallice. Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, and D. Shapiro, editors, *Consciouness and self-regulation: Advances in research and theory*, pages 1–18. Plenum, New York, 1986.

[Obe02] K. Oberauer. Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):411–421, 2002.

[ODC04] S. Oviatt, C. Darves, and R. Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Trans. Comput.-Hum. Interact.*, 11(3):300–328, 2004.

[OdGJ$^+$06] O. Oullier, G. C. de Guzman, K. J. Jantzen, J. Lagarde, and S. J. A. Kelso. Spontaneous synchronization and social memory in interpersonal coordination dynamics. In *Proceedings of Enactive 06*. 2006.

[OdGJ$^+$08] O. Oullier, G. C. de Guzman, K. J. Jantzen, J. Lagarde, and S. J. A. Kelso. Social coordination dynamics: Measuring human bonding. *Social Neuroscience*, 3(2):178–192, 2008.

[Par06] J. S. Pardo. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119:2382–2393, 2006.

[Par07] J.-E. Park. Co-construction of nonnative speaker identity in cross-cultural interaction. *Applied Linguistics*, 28(3):339–360, 2007.

[PB04a] A. Pavlenko and A. Blackledge. Introduction: New theoretical approaches to the study of negotiation of identities in multilingual contexts. In A. Pavlenko and A. Blackledge, editors, *Negotiation of identities in multilingual contexts*, volume 45 of *Bilingual education and bilingualism*, pages 1–33. Multilingual Matters, Clevedon, 2004.

[PB04b] A. Pavlenko and A. Blackledge. *Negotiation of identities in multilingual contexts*, volume 45 of *Bilingual education and bilingualism*. Multilingual Matters, Clevedon, 2004.

[PG04a] M. J. Pickering and S. Garrod. Authors' response: The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–219, 2004.

[PG04b] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004.

[PG05] M. J. Pickering and S. Garrod. Automaticity of language production in monologue and dialogue. In A. Meyer, L. Wheeldon, and A. Krott, editors, *Automaticity and control in language processing*. Psychology Press, Hove, 2005.

[PG06] M. J. Pickering and S. Garrod. Alignment as the basis for successful communication. *Research on Language and Computation*, (4):203–228, 2006.

[PG08] M. J. Pitts and H. Giles. Social psychology and personal relationships:: Accommodation and relational influence across time and contexts. In

G. Antos, E. Ventola, T. Weber, and K. Knapp, editors, *Handbook of interpersonal communication*, volume Vol. 2 of *Handbooks of applied linguistics*, pages 15–31. Mouton de Gruyter, Berlin, 2008.

[PG09] M. J. Pickering and S. Garrod. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2):292–304, 2009.

[PGA10] E. Pederson and S. Guion-Anderson. Orienting attention during phonetic training facilitates learning. *Journal of the Acoustical Society of America*, 127(2), 2010.

[Pie01] J. B. Pierrehumbert. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper, editors, *Frequency and the emergence of linguistic structure*, volume 45 of *Typological studies in language*, pages 137–157. Benjamins, Amsterdam, 2001.

[Pie06] J. B. Pierrehumbert. The next toolkit. *Journal of Phonetics*, 34(4):516–530, 2006.

[Pim66] P. Pimsleur. *Pimsleur Language Aptitude Battery*. H. B. Jovanovich, New York, 1966.

[PL00] A. Pavlenko and J. P. Lantolf. Second language learning as participation and the (re)construction of selves. In J. P. Lantolf, editor, *Sociocultural theory and second language learning*, pages 155–177. Oxford University Press, Oxford, 2000.

[PMF01] T. Piske, I. R. MacKay, and J. E. Flege. Factors affecting degree of foreign accent in an l2: a review. *Journal of Phonetics*, 29:191–215, 2001.

[PS75] M. I. Posner and C. R. R. Snyder. Attention and cognitive control. In R. L. Solso, editor, *Information Processing and cognition: The Loyola symposium*, pages 55–85. Lawrence Erlbaum Associates, Hillsdale, NJ, 1975.

[PS80] E. T. Purcell and R. W. Suter. Predictors of pronunciation accuracy: A reexamination. *Language Learning*, 30:271–287, 1980.

[RAC10] T. J. Ricker, A. M. AuBuchon, and N. Cowan. Working memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(4):573–585, 2010.

[Rev93] W. Revelle. Individual differences in personality and motivation: ”non-cognitive” determinants of cognitive performance. In A. Baddeley, L. Weiskrantz, and D. E. Broadbent, editors, *Attention*. Oxford Univ. Press, Oxford, New York, 1993.

[Rob03] P. Robinson. Attention and memory during sla. In C. J. S. Doughty and M. H. Long, editors, *The handbook of second language acquisition*, volume 14 of *Blackwell handbooks in linguistics*, pages 631–678. Blackwell, Malden, Mass., 2003.

[Rob05] P. Robinson. Aptitude and second language acquisition. *Annual Review of Applied Linguistics*, 25:46–73, 2005.

[SD88] E. Schneiderman and C. Desmarais. A neuropsychological substrate for talent in second language acquisition. In L. K. Obler, editor, *The exceptional brain*, pages 103–126. Guilford Press, New York, NY, 1988.

[SF97] M. L. Sancier and C. A. Fowler. Gestural drift in a bilingual speaker of brazilian portuguese and english. *Journal of Phonetics*, 25:421–436, 1997.

[Sfa98] A. Sfard. On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27(2):4–13, 1998.

[SGKW05] R. Shi, B. Gick, D. Kanwischer, and I. Wilson. Frequency and category factors in the reduction and assimilation of function words: Epg and acoustic measures. *Journal of Psycholinguistic Research*, 34(4):341–364, 2005.

[SGLP01] A. C. Shepard, H. Giles, and A. B. Le Poire. Communication accommodation theory. In P. W. Robinson and H. Giles, editors, *The New handbook of language and social psychology*, pages 33–56. Wiley, Chichester, 2001.

[SHCW78] J. Schumann, J. Holroyd, N. Campbell, and F. Ward. Improvement of foreign language pronunciation under hypnosis: A preliminary study. *Language Learning*, 28(1), 1978.

[Ske86] P. Skehan. Cluster analysis and the identification of learner types. In V. Cook, editor, *Experimental approaches to second language learning*, Language teaching methodology series, pages 81–. Pergamon Inst. of English, Oxford, 1986.

[Ske02] P. Skehan. Theorising and updating aptitude. In P. Robinson, editor, *Individual differences and instructed language learning*, volume 2 of *Language learning and language teaching*, pages 69–93. Benjamins, Amsterdam, 2002.

[Ske03] P. Skehan. *A cognitive approach to language learning*. Oxford applied linguistics. Oxford Univ. Press, Oxford, 2003.

[Smi07] C. L. Smith. Prosodic accommodation by french speakers to a non-native interlocutor. In *Proceedings of the 16th International Conference of Phonetic Sciences*, pages 1081–1084, Saarbrücken, 2007.

[SMM88] D. L. Schacter, M. P. McAndrews, and M. Moscovitch. Access to consciousness: Dissociations between implicit and explicit knowledge in neuropsychological syndromes. In L. Weiskrantz, editor, *Thought without language*, Symposia of the Fyssen Foundation, pages 242–278. Clarendon Pr., Oxford, 1988.

[SSVK83] R. Street, N. J. Street, and A. Van Kleek. Speech convergence among talkative and reticent three year-olds. *Language Sciences*, 5(1):79–96, 1983.

[Sta78] R. C. Stalnaker. Assertion. In P. Cole, editor, *Syntax and Semantics, vol. 9: Pragmatics*, pages 315–332. Academic Press, 1978.

[STG76] L. Simard, D. Taylor, and H. Giles. Attribution processes ad interpersonal accommodation in a bilingual setting. *Language and Speech*, 19(4):374–387, 1976.

[Str84] R. Street. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research*, 11(2):139 – 169, 1984.

[Sty06] E. A Styles. *The psychology of attention*. Psychology Press, Hove, 2. ed. edition, 2006.

[Sut76] R. W. Suter. Predictions of pronunciation accuracy in second language learning. *Language Learning*, 26:233–254, 1976.

[SWWM07] H. Schütze, M. Walsh, T. Wade, and B. Möbius. Accounting for phonetic and syntactic phenomena in a multi-level competitive interaction model. In *ESSLI Workshop on Exemplar Based Models of Language Acquisition and Use*, volume 2007. Dublin, 2007.

[SZK+95] R. V. Shannon, F.-G Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[TG80] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[TGC82] J. N. Thakerar, H. Giles, and J. Cheshire. Psychological and linguistic parameters of speech accommodation theory. In C. Fraser and K. R.

Scherer, editors, *Advances in the Social Psychology of Language*, pages 205–255. Cambride Univ. Press, Cambridge, 1982.

[Tre69] A. Treisman. Strategies and models of selective attention. *Psychological Review*, 76:282–299, 1969.

[Tre93] A. Treisman. The perception of features and objects. In A. Baddeley, L. Weiskrantz, and D. E. Broadbent, editors, *Attention*, pages 5–35. Oxford Univ. Press, Oxford, New York, 1993.

[Tre99] A. Treisman. Feature binding, attention and object perception. In G. W. Humphreys, J. Duncan, and A. Treisman, editors, *Attention, space, and action*, pages 91–111. Oxford Univ. Press, Oxford, 1999.

[TT99] S. Ting-Toomey. *Communicating across cultures*. Guilford Press, New York, 1999.

[UD09] E. Ushioda and Z. Dörnyei. Motivation, language identities and the l2 self: A theoretical overview. In Z. Dörnyei and E. Ushioda, editors, *Motivation, language identity and the L2 self*, Second language acquisition, pages 1–8. Multilingual Matters, Buffalo, 2009.

[Ull01a] M. T. Ullman. The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and cognition*, 4(2):105–122, 2001.

[Ull01b] M. T. Ullman. A neurocognitive perspective on language: the declarative/procedural model. *Nature Reviews Neuroscience*, (2):717–727, 2001.

[Ull04] M. T. Ullman. Contributions of memory circuits to language: the declarative/ procedural model. *Cognition*, 92:231–270, 2004.

[UP05] M. T. Ullman and Pierpont. E. I. Specific language impairment is not specific to language: the procedural deficit hypothesis. *Cortex*, 41(3):399–433, 2005.

[vEBBB$^+$10] K. J. van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow. The wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53(4):510–540, 2010.

[WA01] D. L. Woods and C. Alain. Conjoining three auditory features: An event-related brain potential study. *Journal of Cognitive Neuroscience*, 13(4):492–509, 2001.

[WDS$^+$10] T. Wade, G. Dogil, H. Schütze, M. Walsh, and B. Möbius. Syllable frequency effects in a context-sensitive segment production model. *Journal of Phonetics*, 38:227–239, 2010.

[Weg03] D. M. Wegner. The mind's best trick: How we experience conscious will. *Trends in Cognitive Sciences*, 7:65–69, 2003.

[WGC92] D. Wilkes-Gibbs and H. H. Clark. Coordinating beliefs in conversation. *Journal of Memory and Language*, (31):183–194, 1992.

[WGCP97] M. Willemyns, C. Gallois, V. Callan, and J. Pittam. Accent accommodation in the job interview: The impact of interviewer accent and gender. *Journal of Language and Social Psychology*, 16(1):3–22, 1997.

[WMWS10] M. Walsh, B. Möbius, T. Wade, and H. Schütze. Multilevel exemplar theory. *Cognitive Science*, 34(4):537–582, 2010.

[WN01] A. Williams and J. F. Nussbaum. *Intergenerational Communication across the Lifespan*. Lawrence Erlbaum, Mahwah, NJ, 2001.

[WSMS07] M. Walsh, H. Schütze, B. Möbius, and A. Schweitzer. An exemplar-theoretic account of syllable frequency effects. In J. Trouvain and W. J. Barry, editors, *Proceedings of the 16th International Conference of Phonetic Sciences*. Saarbrücken, 2007.

[WvVed] A. B. Wedel and H. van Volkinburg. Modeling simultaneous convergence and divergence of linguistic features between differently-identifying groups in contact. Submitted.

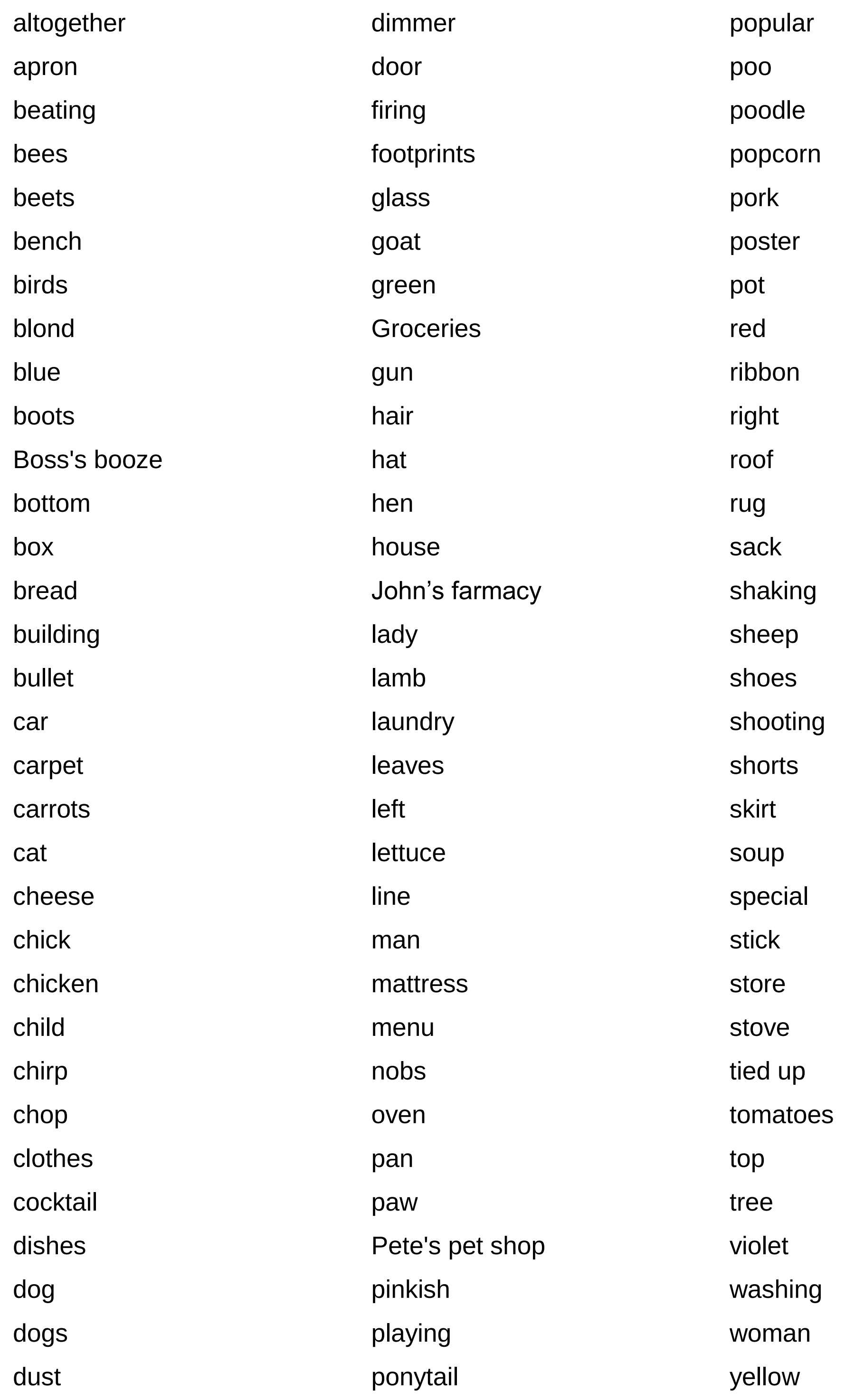
[ZR98] R. A. Zwaan and G. A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, (123):162–185, 1998.

[ZSY97] Y. Zhang, N. Suga, and J. Yan. Corticofugal modulation of frequency processing in bat auditory system. *Nature*, 387:900–903, 1997.

Appendices

Appendix A

Word list



altogether
apron
beating
bees
beets
bench
birds
blond
blue
boots
Boss's booze
bottom
box
bread
building
bullet
car
carpet
carrots
cat
cheese
chick
chicken
child
chirp
chop
clothes
cocktail
dishes
dog
dogs
dust
dimmer
door
firing
footprints
glass
goat
green
Groceries
gun
hair
hat
hen
house
John’s farmacy
lady
lamb
laundry
leaves
left
lettuce
line
man
mattress
menu
nobs
oven
pan
paw
Pete's pet shop
pinkish
playing
ponytail
popular
poo
poodle
popcorn
pork
poster
pot
red
ribbon
right
roof
rug
sack
shaking
sheep
shoes
shooting
shorts
skirt
soup
special
stick
store
stove
tied up
tomatoes
top
tree
violet
washing
woman
yellow

Figure A.1: Word list used for the pre- and post-test recording. The list contains both target and filler words.

Appendix B

Diapix

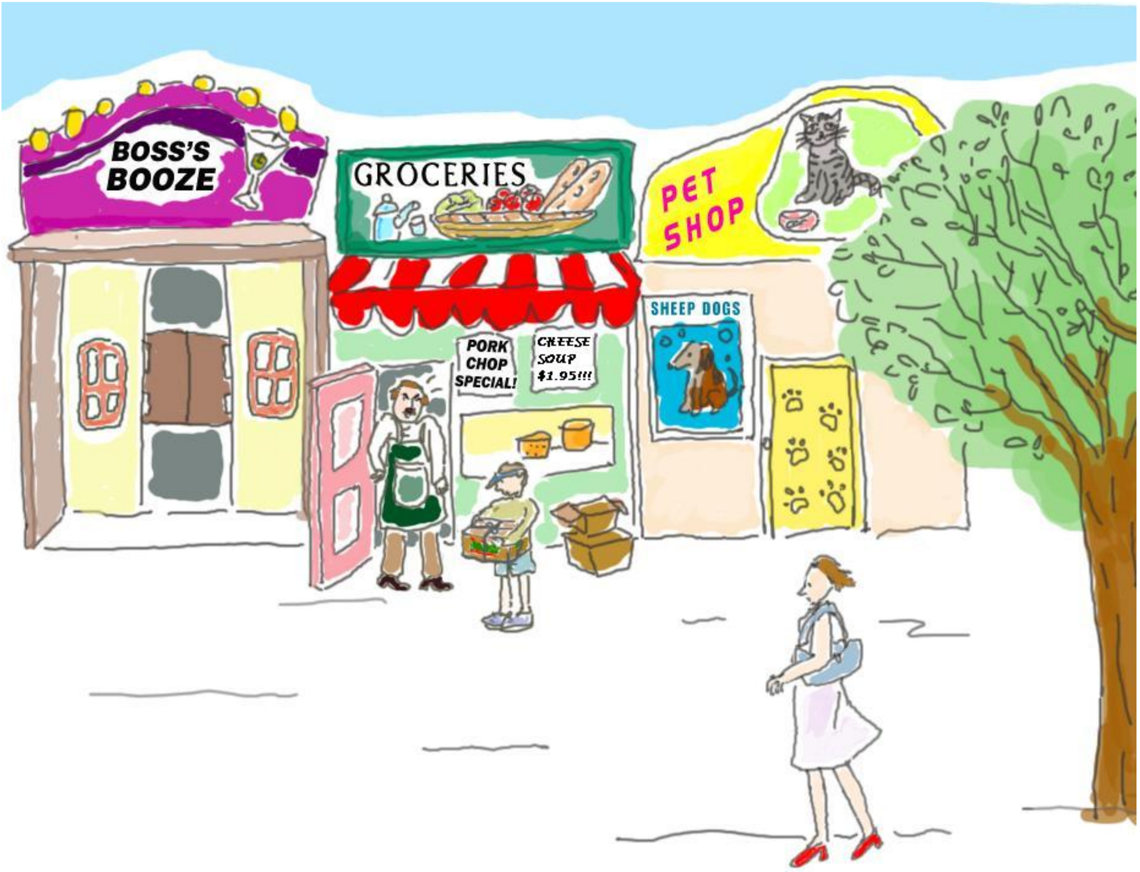


Figure B.1: Diapix picture 1 from the "shop scene", see [BBC+07, vEBBB+10].

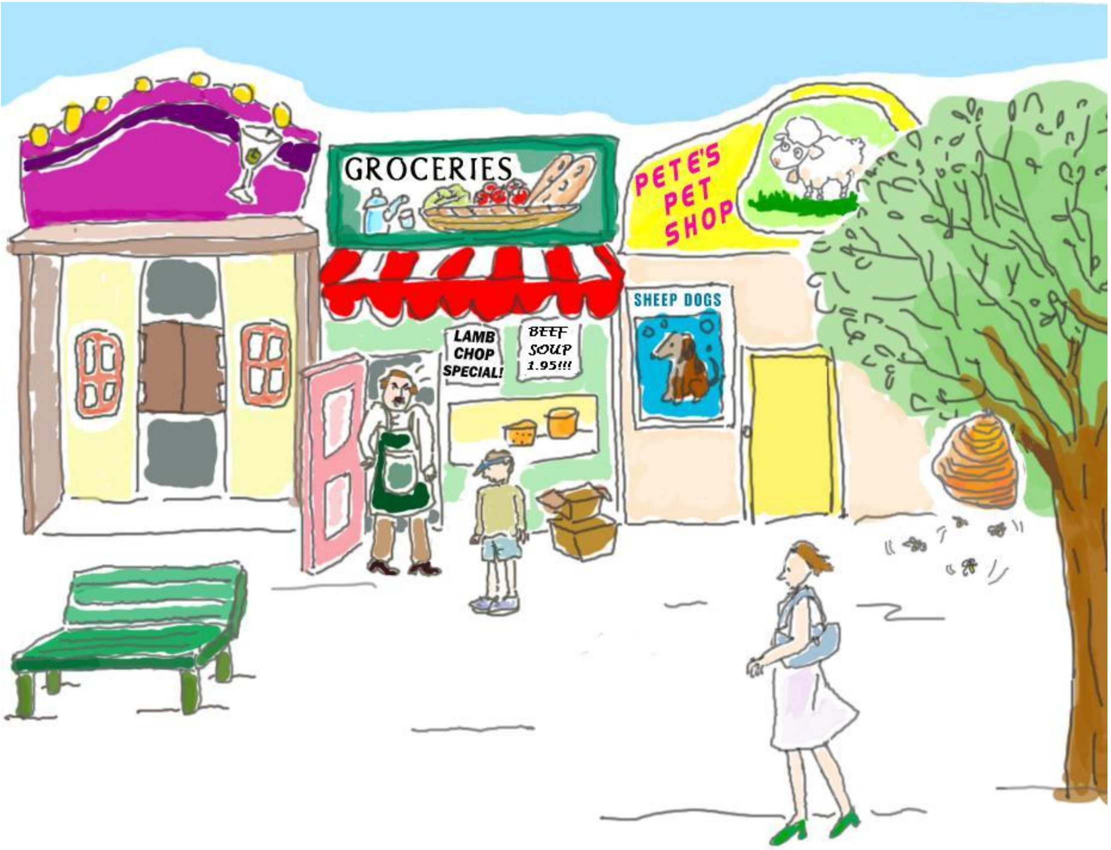


Figure B.2: Diapix picture 2 from the ”shop scene”, see [BBC+07, vEBBB+10].

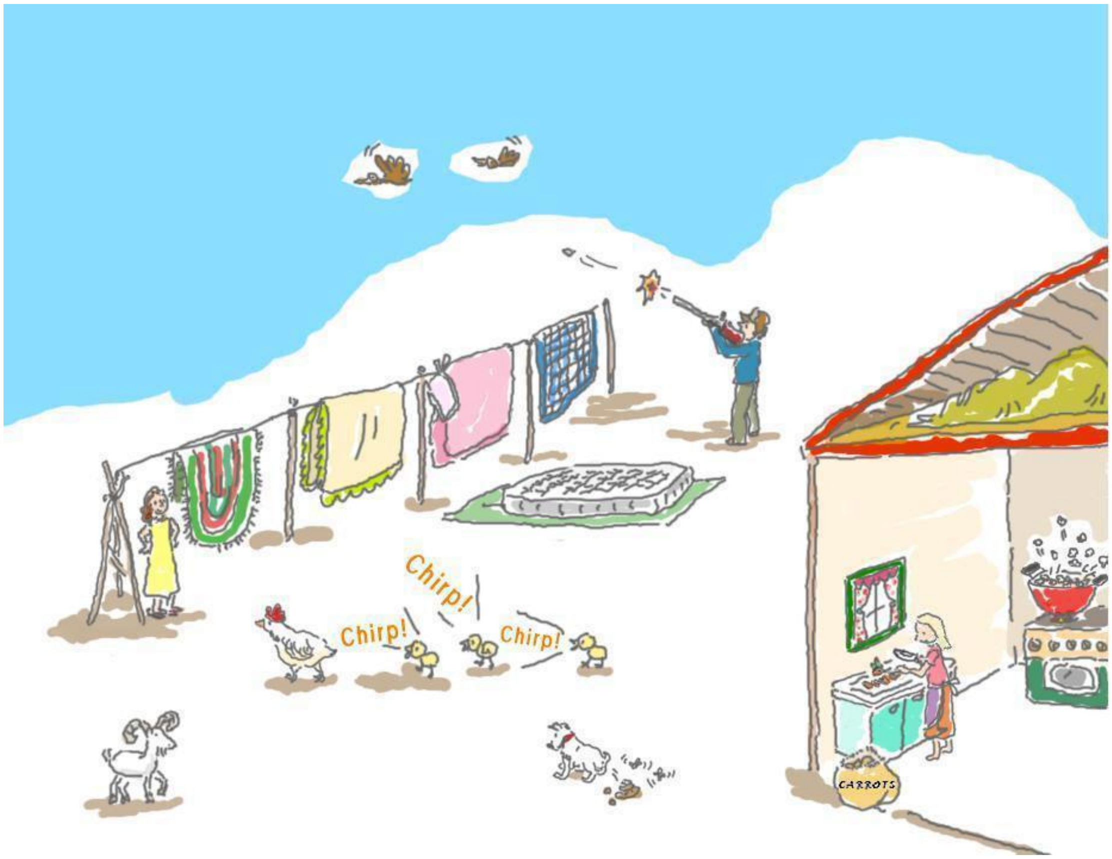


Figure B.3: Diapix picture 1 from the ”farm scene”, see [BBC+07, vEBBB+10].

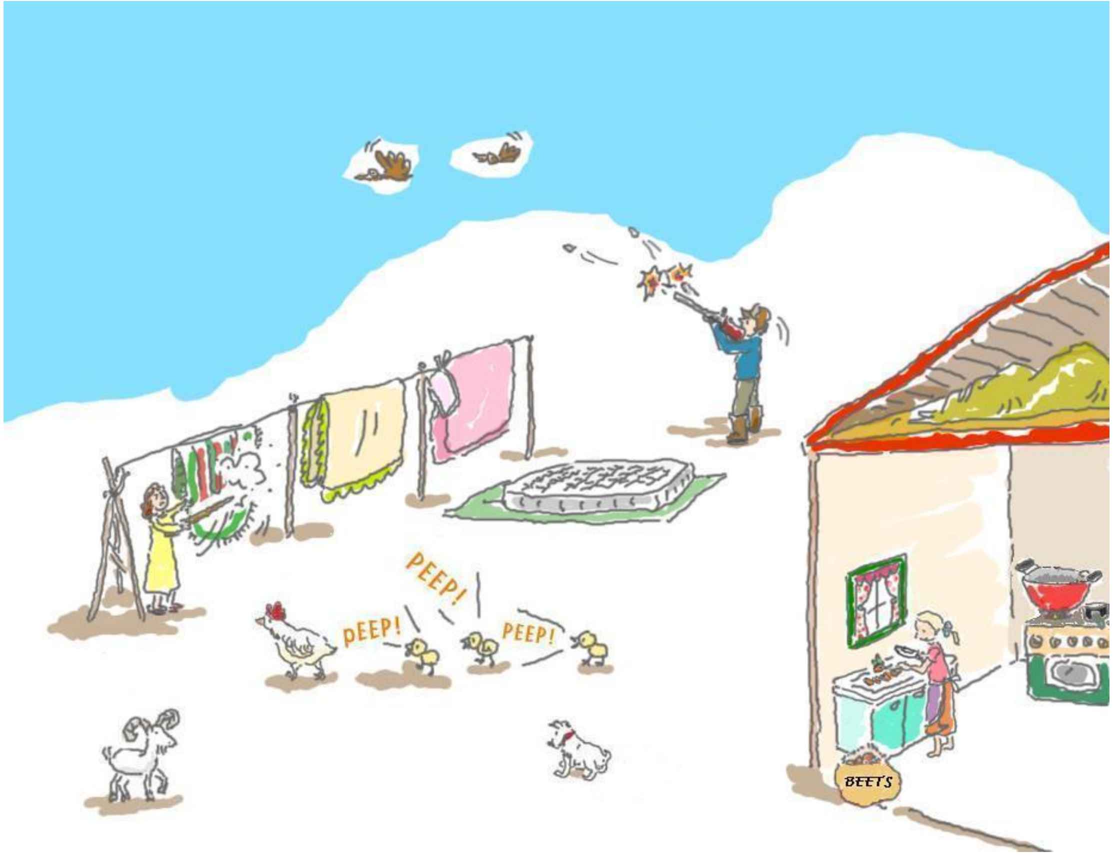


Figure B.4: Diapix picture 2 from the ”farm scene”, see [BBC+07, vEBBB+10].

Appendix C

Amplitude envelope script

C.1 Working loop

```
%% define working path

 dial_filt_flag    = 1; % 1: filter dialogs , 0: no filter
 dialog_list       = {'JNT';'TJN';'JJN';'JNJ'};
 read_list         = {'J';'T';'JN'};

working_path      = [working_path,'/'];

%% read filenames
 files   =   dir([working_path,'*.wav']);

if isempty(files) | length(files)==1
    error(['no␣executable␣files␣found␣in:␣',working_path])
end

Results = {};
Results_dialog = {};
Results_read = {};

num_cmp          = 0;
num_cmp_dialog = 0;
num_cmp_read   = 0;

num_err          = 0;

for i1 = 1:length(files)-1
```

```
sep            =  regexp(files(i1).name,'_');

if numel(sep)==0 %wrong file name
    Files_Not_used{num_err+1} = files(i2).name;
    num_err = num_err+1;
    continue
end

SName1      =  files(i1).name(1:sep(1));

%extract name_tag of file  1
if length(sep)  <= 1
       sep(2)   =  regexp(files(i1).name,'.wav');
end
name_tag1    =   files(i1).name((sep(1)+1):(sep(2)-1));

%output for monitoring only
%     clc
disp(['aktuelles Sample ',...
        '(',num2str(i1),' von ', num2str(length(files)),'): ' ,...
     files(i1).name])

for i2  = i1+1:length(files)
    sep       =  regexp(files(i2).name,'_');

    if numel(sep)==0 %falscher Dateiname
        Files_Not_used{num_err+1} = files(i2).name;
        num_err = num_err+1;
        continue
    end

    SName2 =  files(i2).name(1:sep(1));

    %extract name_tag
    if length(sep)  <= 1
        sep(2)   =  regexp(files(i2).name,'.wav');
    end
    name_tag2    =   files(i2).name((sep(1)+1):(sep(2)-1));

    if (strcmp(SName1, SName2) == 1)

        switch  dial_filt_flag
            case 0 %no filter

```

```
                    num_cmp = num_cmp+1;

                    Results{num_cmp,1}  =  files (i1).name;
                    Results{num_cmp,2}  =  files (i2).name;

                    [Results{num_cmp,3}, Results{num_cmp,4}] = ...
                        sampleMatch_nl([working_path,files(i1).name],[working_path,files(i2).name]);

            case 1 %filter  dialogs
                if  (max(ismember(dialog_list,name_tag1))) && (max(ismember(dialog_list,name_tag2)))

                    disp(['␣dialog␣(',  files (i2).name,')'])

                    num_cmp_dialog = num_cmp_dialog+1;

                    Results_dialog{num_cmp_dialog,1}  =  files (i1).name;
                    Results_dialog{num_cmp_dialog,2}  =  files (i2).name;

                    [Results{num_cmp_dialog,3}, Results{num_cmp_dialog,4}] = ...
                        sampleMatch_nl([working_path,files(i1).name],[working_path,files(i2).name]);

                elseif  (max(ismember(read_list,name_tag1))) && (max(ismember(read_list,name_tag2)))

                    disp(['␣read␣(',  files (i2).name,')'])

                    num_cmp_read = num_cmp_read+1;

                    Results_read{num_cmp_read,1}  =  files (i1).name;
                    Results_read{num_cmp_read,2}  =  files (i2).name;

                    [Results_read{num_cmp_read, 3}, Results_read{num_cmp_read, 4} ] = ...
                        sampleMatch_nl([working_path,files(i1).name],[working_path,files(i2).name]);

                else

                    disp(['␣cross␣(',  files (i2).name,')'])

                    num_cmp = num_cmp+1;

                    Results{num_cmp,1}  =  files (i1).name;
                    Results{num_cmp,2}  =  files (i2).name;

                    [Results{num_cmp,3}, Results{num_cmp,4}] = ...
                        sampleMatch_nl([working_path,files(i1).name],[working_path,files(i2).name]);

                    end
            end

        end
    end

end

save([working_path,'ergebnisse'], 'Results', 'Results_dialog', 'Results_read');
```

C.2 Sample match

```
function [match_val, x] = sampleMatch_nl(dat1, dat2)
%returns envelope match of 2 sample-wavs
%dat1 : Sample 1 (Wave-File)
%dat2 : Sample 2 (Wave-File)
%function call: sampleMatch_nl(dat1,dat2)

KRMS   = 0.03;
KFS    = 16000;
KENVFS = 500;
normRms = @(x)x/sqrt(mean(x.^2))*KRMS;

band_lo = 80;
band_hi = 7800;

%note: I'm hi-emphasizing the sounds to give more weight to the
%(lower-amplitude) high frequency range of the sounds.
hiEmph = @(x, diff1fact)filter([1 -diff1fact], 1, x);

%read in a couple of sounds, normalize amplitudes
wd1 = normRms(hiEmph(wavread(dat1), 0.95));
wd2 = normRms(hiEmph(wavread(dat2), 0.95));

%get the envelopes - see snd2env for explanation
env1 = snd2env(wd1, KFS, [band_lo band_hi], 4, 60, @(x)abs(hilbert(x)), KENVFS);
env2 = snd2env(wd2, KFS, [band_lo band_hi], 4, 60, @(x)abs(hilbert(x)), KENVFS);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% find longer envelope
length_difference = size(env1,1)-size(env2,1);
if(length_difference>0),
    longerEnv = env1;
    shorterEnv = env2;
```

```
else
    longerEnv = env2;
    shorterEnv = env1;
end
% shortest length
sLen = size(shorterEnv(:,1));

% http://labrosa.ee.columbia.edu/matlab/dtw/
% see http://labrosa.ee.columbia.edu/matlab/dtw/simmx.m
%x = shorterEnv' * longerEnv(1:sLen,:) / (sqrt(sum(sum(shorterEnv.^2)))' * sqrt(sum(sum(longerEnv(1:sLen).^2))))

x1 = shorterEnv(:,1)' * longerEnv(1:sLen,1) / (sqrt(sum(sum(shorterEnv(:,1).^2)))' *
                                                   sqrt(sum(sum(longerEnv(1:sLen,1).^2))));
x2 = shorterEnv(:,2)' * longerEnv(1:sLen,2) / (sqrt(sum(sum(shorterEnv(:,2).^2)))' *
                                                   sqrt(sum(sum(longerEnv(1:sLen,2).^2))));
x3 = shorterEnv(:,3)' * longerEnv(1:sLen,3) / (sqrt(sum(sum(shorterEnv(:,3).^2)))' *
                                                   sqrt(sum(sum(longerEnv(1:sLen,3).^2))));
x4 = shorterEnv(:,4)' * longerEnv(1:sLen,4) / (sqrt(sum(sum(shorterEnv(:,4).^2)))' *
                                                   sqrt(sum(sum(longerEnv(1:sLen,4).^2))));
x = [x1,x2,x3,x4];

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%normalize actual envelopes
%normalize to [0 1]
env1 = env1./(sqrt(sum(sum(env1.^2))));
env2 = env2./(sqrt(sum(sum(env2.^2))));

%and estimate the match. see the envelope match function for explanation
match_val = envelopeMatch(env1, env2);
```

C.3 Sound to envelope

```
function [e, CFs, sB] = snd2env(s,...              %original signal
                            iFsOrig ,...           %original sampling rate (Hz)
                        fTotFreqRange,...          %total freq range to consider (lo, hi; hi must be < iFs/2)
                          iNumBands,...            %number of frequency bands within fTotFreqRange
                         fEnvCutOff ,...           %cutoff frequency (Hz) for envelope filter
                        fhEnvMethod,...            %handle of function to extract (raw) envelope
                        iFsNew)                    %new sampling rate (should be >> fEnvCutOff *2)

%function [e, CFs, sB] = snd2env: returns envelope (e), band center
%frequencies (CFs), and band-separated original sound (sB)

```

```
%first get the low and high cutoff frequencies - just log-spaced here:
bandLo = fTotFreqRange(1)*exp(log(fTotFreqRange(2)/fTotFreqRange(1))/iNumBands).^(0:iNumBands-1);
bandHi = fTotFreqRange(1)*exp(log(fTotFreqRange(2)/fTotFreqRange(1))/iNumBands).^(1:iNumBands);

%(also calculate the center frequencies (log scale) of the bands, to return)
CFs = bandLo.*sqrt(bandHi./bandLo);

% initialize the band-separated signal
sB = [];
for band = 1:length(bandLo),

    %filter the signal into the appropriate band, and add it to sB.
    %low order butterworth filter can be reasonable but is not very realistic
    [bBand, aBand] = butter(2, [bandLo(band) bandHi(band)]/(iFsOrig/2), 'bandpass');
    sB = [sB, filtfilt (bBand ,aBand ,s)];
end

%lowpass filter the raw envelopes calculated above.
[bEnv, aEnv] = butter(4, fEnvCutOff/(iFsOrig/2), 'low');
e = filtfilt (bEnv, aEnv, fhEnvMethod(sB));
end
```

C.4 Envelope match

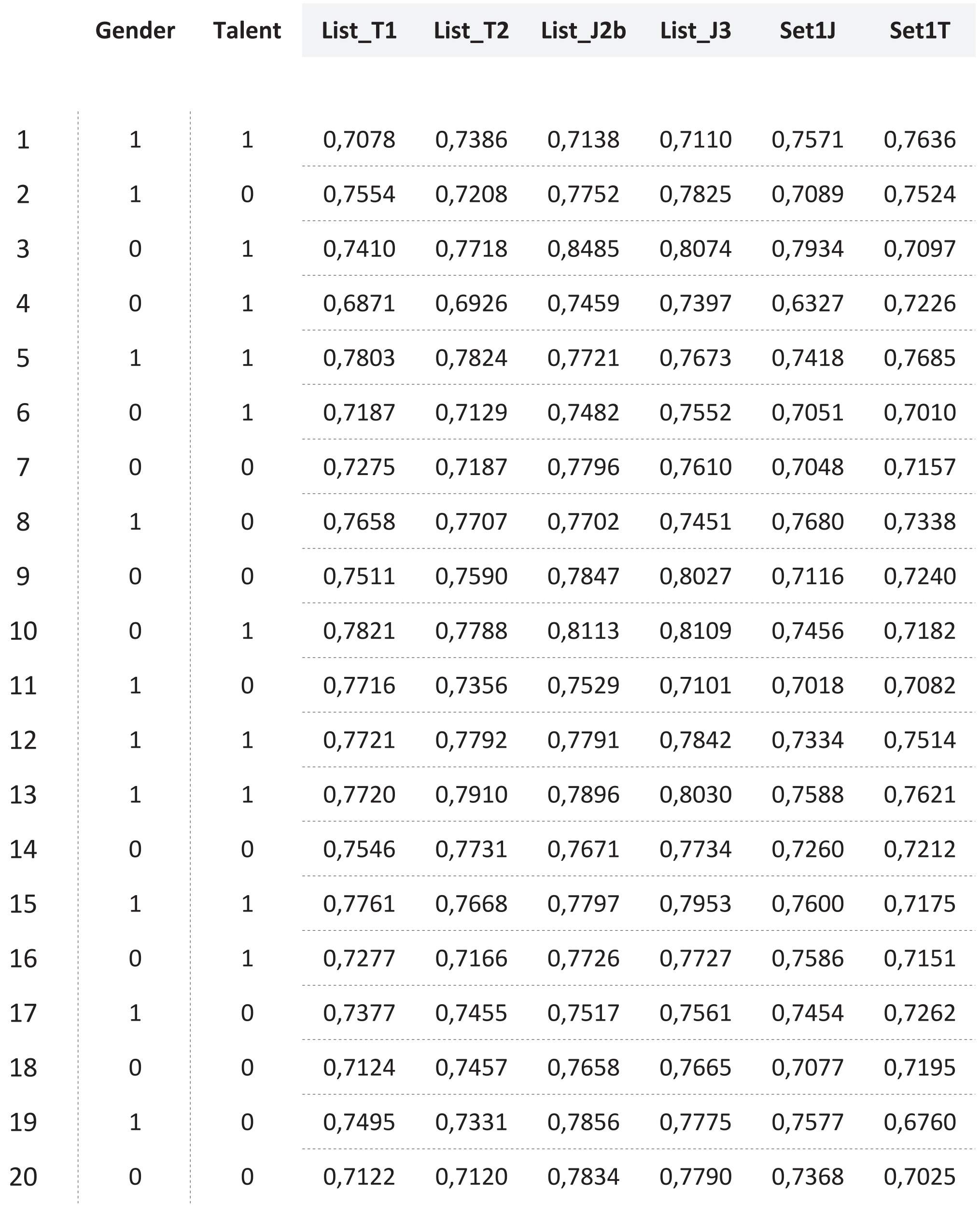
```
function matchVal = envelopeMatch(e1, e2)

length_difference = size(e1,1)-size(e2,1);
if(length_difference>0),
    longerEnv = e1;
    shorterEnv = e2;
    maxLag = length_difference;
else
    longerEnv = e2;
    shorterEnv = e1;
    maxLag = -length_difference;
end

matchSum = zeros(maxLag*2+1,1);
for band = 1:size(e1,2),
    matchSum = matchSum+xcorr(longerEnv(:,band), shorterEnv(:,band), maxLag);
end
matchVal = max(matchSum);
```

C.5 Function evaluation

```
function evaluation (mode)

%% evaluation with a sample of test wavs
switch mode

    case 1
      % 2 identical wavs:
      fname1 = 'popcorn_BSJ_1same.wav';
      fname2 = 'popcorn_BSJ_2same.wav';

    case 2
      % 2 completely distinct wavs
      fname1 = 'carpet_MH_diff2.wav';
      fname2 = 'carpet_MHJ_diff1.wav';

    case 3
      % 3 similar wavs from one speaker
      fname1 = 'apron_MH_1.wav';
      fname2 = 'apron_MH_2.wav';

    case 4
      fname1 = 'apron_MH_1.wav';
      fname2 = 'apron_MH_3.wav';

    case 5
      fname1 = 'apron_MH_2.wav';
      fname2 = 'apron_MH_3.wav';

end

evalpath = '/mount/projekte44/talent-gehirn/natalie_matlab/evaltest/';

[mv, x] = sampleMatch_new_func([evalpath, fname1], [evalpath, fname2])
```

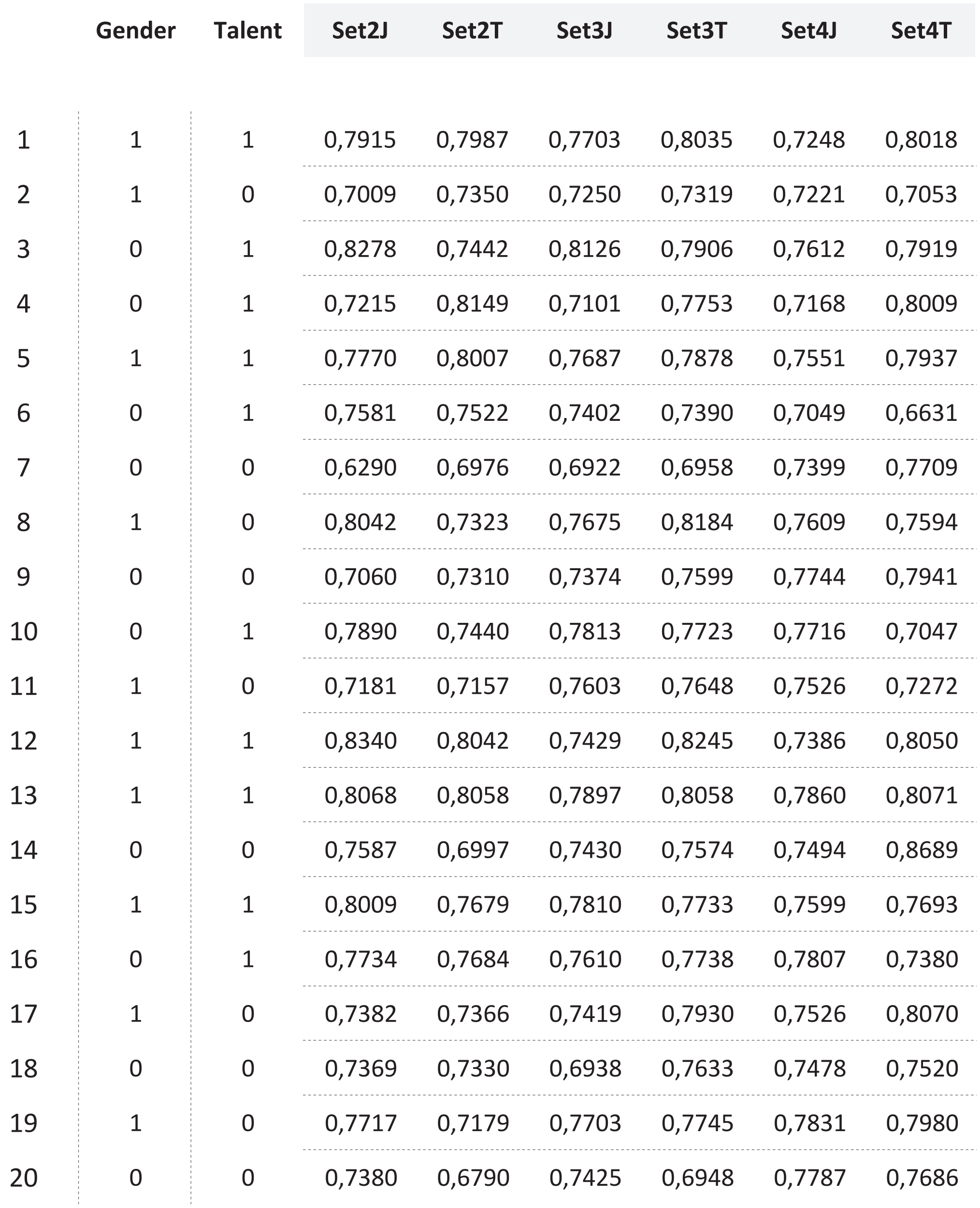
Appendix D

Match value means



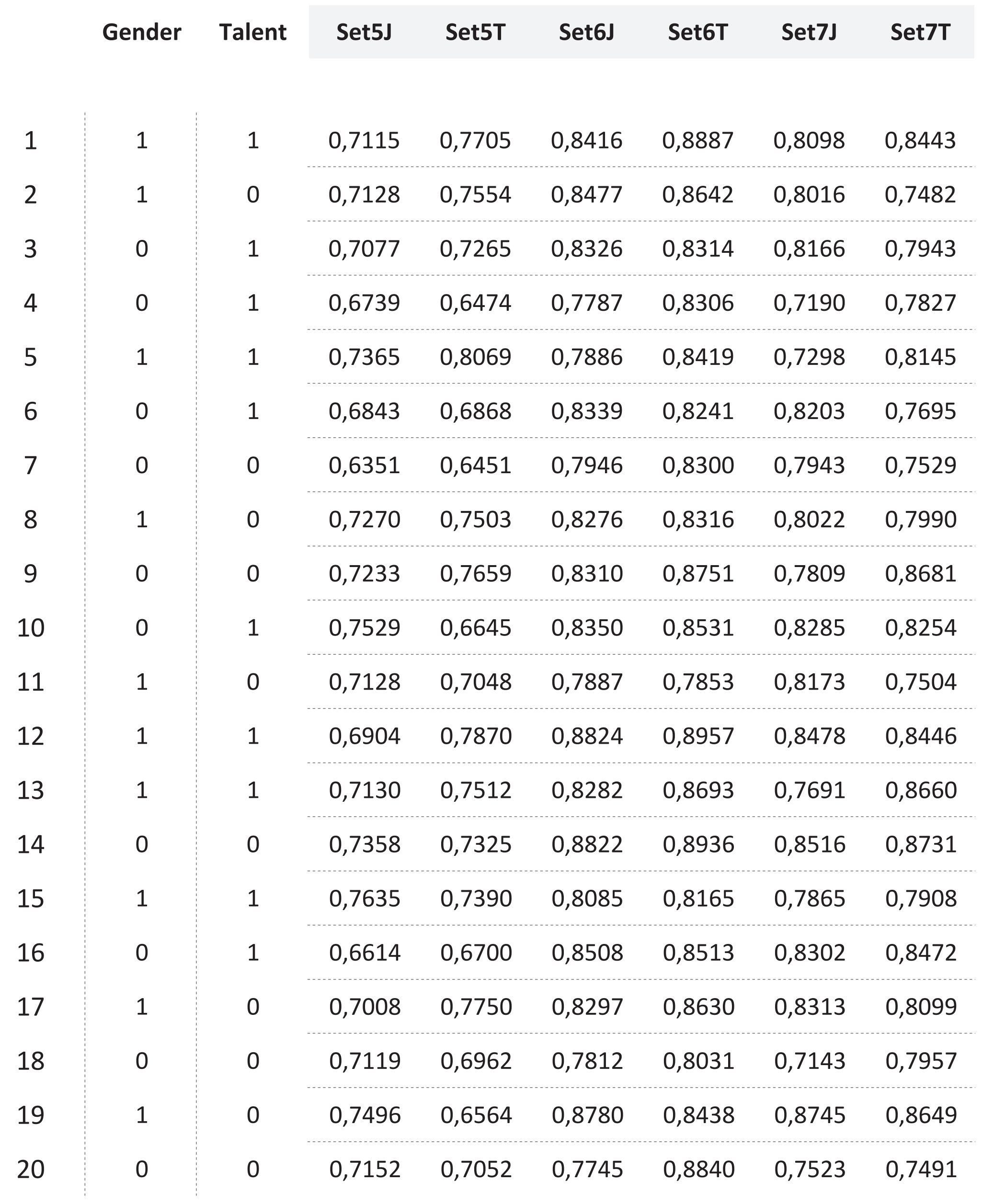
	Gender	Talent	List_T1	List_T2	List_J2b	List_J3	Set1J	Set1T
1	1	1	0,7078	0,7386	0,7138	0,7110	0,7571	0,7636
2	1	0	0,7554	0,7208	0,7752	0,7825	0,7089	0,7524
3	0	1	0,7410	0,7718	0,8485	0,8074	0,7934	0,7097
4	0	1	0,6871	0,6926	0,7459	0,7397	0,6327	0,7226
5	1	1	0,7803	0,7824	0,7721	0,7673	0,7418	0,7685
6	0	1	0,7187	0,7129	0,7482	0,7552	0,7051	0,7010
7	0	0	0,7275	0,7187	0,7796	0,7610	0,7048	0,7157
8	1	0	0,7658	0,7707	0,7702	0,7451	0,7680	0,7338
9	0	0	0,7511	0,7590	0,7847	0,8027	0,7116	0,7240
10	0	1	0,7821	0,7788	0,8113	0,8109	0,7456	0,7182
11	1	0	0,7716	0,7356	0,7529	0,7101	0,7018	0,7082
12	1	1	0,7721	0,7792	0,7791	0,7842	0,7334	0,7514
13	1	1	0,7720	0,7910	0,7896	0,8030	0,7588	0,7621
14	0	0	0,7546	0,7731	0,7671	0,7734	0,7260	0,7212
15	1	1	0,7761	0,7668	0,7797	0,7953	0,7600	0,7175
16	0	1	0,7277	0,7166	0,7726	0,7727	0,7586	0,7151
17	1	0	0,7377	0,7455	0,7517	0,7561	0,7454	0,7262
18	0	0	0,7124	0,7457	0,7658	0,7665	0,7077	0,7195
19	1	0	0,7495	0,7331	0,7856	0,7775	0,7577	0,6760
20	0	0	0,7122	0,7120	0,7834	0,7790	0,7368	0,7025

Figure D.1: The mean match values for all subjects in the read and dialog speech conditions. **List T1, List T2, List J2b** and **List J3** refer to the read pre- and post-test, **Set 1** through **Set 7** refer to the dialog experiment. **Part 1**.



	Gender	Talent	Set2J	Set2T	Set3J	Set3T	Set4J	Set4T
1	1	1	0,7915	0,7987	0,7703	0,8035	0,7248	0,8018
2	1	0	0,7009	0,7350	0,7250	0,7319	0,7221	0,7053
3	0	1	0,8278	0,7442	0,8126	0,7906	0,7612	0,7919
4	0	1	0,7215	0,8149	0,7101	0,7753	0,7168	0,8009
5	1	1	0,7770	0,8007	0,7687	0,7878	0,7551	0,7937
6	0	1	0,7581	0,7522	0,7402	0,7390	0,7049	0,6631
7	0	0	0,6290	0,6976	0,6922	0,6958	0,7399	0,7709
8	1	0	0,8042	0,7323	0,7675	0,8184	0,7609	0,7594
9	0	0	0,7060	0,7310	0,7374	0,7599	0,7744	0,7941
10	0	1	0,7890	0,7440	0,7813	0,7723	0,7716	0,7047
11	1	0	0,7181	0,7157	0,7603	0,7648	0,7526	0,7272
12	1	1	0,8340	0,8042	0,7429	0,8245	0,7386	0,8050
13	1	1	0,8068	0,8058	0,7897	0,8058	0,7860	0,8071
14	0	0	0,7587	0,6997	0,7430	0,7574	0,7494	0,8689
15	1	1	0,8009	0,7679	0,7810	0,7733	0,7599	0,7693
16	0	1	0,7734	0,7684	0,7610	0,7738	0,7807	0,7380
17	1	0	0,7382	0,7366	0,7419	0,7930	0,7526	0,8070
18	0	0	0,7369	0,7330	0,6938	0,7633	0,7478	0,7520
19	1	0	0,7717	0,7179	0,7703	0,7745	0,7831	0,7980
20	0	0	0,7380	0,6790	0,7425	0,6948	0,7787	0,7686

Figure D.2: The mean match values for all subjects in the read and dialog speech conditions. **List T1, List T2, List J2b** and **List J3** refer to the read pre- and post-test, **Set 1** through **Set 7** refer to the dialog experiment. **Part 2**.



	Gender	Talent	Set5J	Set5T	Set6J	Set6T	Set7J	Set7T
1	1	1	0,7115	0,7705	0,8416	0,8887	0,8098	0,8443
2	1	0	0,7128	0,7554	0,8477	0,8642	0,8016	0,7482
3	0	1	0,7077	0,7265	0,8326	0,8314	0,8166	0,7943
4	0	1	0,6739	0,6474	0,7787	0,8306	0,7190	0,7827
5	1	1	0,7365	0,8069	0,7886	0,8419	0,7298	0,8145
6	0	1	0,6843	0,6868	0,8339	0,8241	0,8203	0,7695
7	0	0	0,6351	0,6451	0,7946	0,8300	0,7943	0,7529
8	1	0	0,7270	0,7503	0,8276	0,8316	0,8022	0,7990
9	0	0	0,7233	0,7659	0,8310	0,8751	0,7809	0,8681
10	0	1	0,7529	0,6645	0,8350	0,8531	0,8285	0,8254
11	1	0	0,7128	0,7048	0,7887	0,7853	0,8173	0,7504
12	1	1	0,6904	0,7870	0,8824	0,8957	0,8478	0,8446
13	1	1	0,7130	0,7512	0,8282	0,8693	0,7691	0,8660
14	0	0	0,7358	0,7325	0,8822	0,8936	0,8516	0,8731
15	1	1	0,7635	0,7390	0,8085	0,8165	0,7865	0,7908
16	0	1	0,6614	0,6700	0,8508	0,8513	0,8302	0,8472
17	1	0	0,7008	0,7750	0,8297	0,8630	0,8313	0,8099
18	0	0	0,7119	0,6962	0,7812	0,8031	0,7143	0,7957
19	1	0	0,7496	0,6564	0,8780	0,8438	0,8745	0,8649
20	0	0	0,7152	0,7052	0,7745	0,8840	0,7523	0,7491

Figure D.3: The mean match values for all subjects in the read and dialog speech conditions. **List T1, List T2, List J2b** and **List J3** refer to the read pre- and post-test, **Set 1** through **Set 7** refer to the dialog experiment. **Part 3**.