Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Intrinsic and Extrinsic Component Evaluation in Interactive Multilingual Speech Applications

Anne H. Schneider

Thesis submitted for the Degree of Doctor of Philosophy

Department of Computer Science

University of Dublin

Trinity College

September 2012

„We are stuck with technology when what we really want is just stuff that works. “

Douglas Adams, The Salmon of Doubt.

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Published or unpublished work of others wherever it is included, is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Summary

Due to the steady progress in technology, together with the rapid increase of powerful mobile devices, the use of voice interfaces and other speech enabled technologies has invaded our every day lives. Today people talk to their mobile phones, get directions from a car navigation system or have an electronic book read out to them. One of these emerging language applications is speech-to-speech (S2S) translation which takes on a fundamental problem of the economic and cultural exchange, the language barrier. The three main components of S2S translation, automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS), saw a huge leap forward in the last two decades due to a change to probabilistic methods which caused a boost in S2S translation technology. However, it is still evident that the output of these three components is far from perfect, deeming good evaluation methods essential. It will be observed that it is widespread practice for ASR, MT and TTS, to be evaluated intrinsically, without regard to the context of the application (e.g. the task or the user). In recent years there have been calls for paying more attention to extrinsic evaluation in machine translation in particular and natural language technologies in general. Extrinsic evaluation aims to assess the (often indirect) effect of a module on task- and context-dependent variables such as user performance.

This thesis is set to investigate whether ASR, MT and TTS as main components of S2S translation can benefit from extrinsic evaluation methods. In this context the correlation of intrinsic measures with human judgement will be scrutinised by comparing intrinsic and extrinsic evaluation results. By answering this question this research aims at contributing to the discussion that natural language processing (NLP) and human-computer interaction (HCI) could benefit from a cross-fertilisation, especially whether NLP evaluation can be fertilised by ideas and methods from the field of HCI.

Experiments with altogether over eighty participants were carried out which compared the state-of-the-art intrinsic evaluation methods for ASR, MT and TTS such as the word error rate, BLEU or the mean opinion score with extrinsic evaluation methods. In the extrinsic evaluations the map task and the Wizard of Oz technique were employed.

The key conclusions are that NLP evaluation can benefit from methods and ideas inherited from the field of HCI but with several caveats. By carrying out the experiments it was confirmed that evaluation efforts need to keep the user and application context in mind if they aim to produce meaningful results for real-life situations. The main advantage of such extrinsic evaluation is that they often offer a better insight into obstructive situations that cannot be revealed by the results of intrinsic evaluation. Furthermore, this research shows that extrinsic evaluations can be used to fine-tune intrinsic evaluation metrics. Albeit these big advantages the execution of the experiments has confirmed that extrinsic methods are very time-consuming and laborious which includes the difficult task to plan how to capture interesting data and to decide which data is important for analysis. The research in the context of this thesis has also demonstrated that the map task and the Wizard of Oz technique are well suited for extrinsic NLP evaluation under certain circumstances. Improvements on the methods were inspired by the research carried out in this thesis. By doing so it has further been demonstrated that HCI can also benefit from ideas and methods that stem from the field of NLP.

Acknowledgements

I would like to thank a number of people for their support. Without them this thesis would never have been written:

First, I'd like to thank my supervisor Saturnino Luz, for offering me with new perspectives and his numerous valuable comments. Further, I would like to acknowledge Trinity College Dublin for affiliating my research, and Science Foundation Ireland that provided me with an ample grant through the Centre for Next Generation Localisation (CNGL). In some sense I was very lucky and had a second supervisor, who always had an open ear for my problems and who was never tired of discussing my research. Thank you, Micha. I guess, now I owe you a PhD.

The last four years wouldn't have been the same without my like-minded office mate Stephan, who always provided me with a fair share of coffee and cleaned the dishes afterwards. Stephan and I had two fellow coffee drinkers, Ielka and Niki, the two post-docs in our group. Unfortunately, we didn't have the opportunity to spend as much time together. However, big thanks are due for their support especially at the beginning of this big adventure. Besides the hard work, there have also been social moments like lunches or Christmas parties, which would have been pretty lonely without the guys from the green floor: Lilly, Alfredo, Ger, Hector, Erwan, Roman, Oscar, Martin and Francesca. Thank you all for your company and I hope we will stay in touch. Not working on the green floor, but always present at the social events was Daniel, whom I'd like to thank for teaching me a lesson on time management.

The speech group at DCU provided me with synthesised sentences for my experiments. They were also helping me with questions about ASR and TTS and providing literature recommendations. Here, I would like to acknowledge Joao Cabral and John Kane especially, for reviewing the section on ASR.

In the last year of my PhD research I spent the predominant part of my time in Germany, doing my research remotely. In this time Dominic was providing me with a second home in Dublin. Heartfelt thanks are due for this. You'll always find a place at our home. Thanks are also due to Professor Hahn for giving me shelter at his lab at the Friedrich Schiller University in Jena in this last year of my PhD and to my colleagues at the JULIE lab: Erik, Jo, Benjamin,

Kerstin, and Elena.

I am also very grateful for the support of my parents and my brothers. Thank you Max for being available at any time and for keeping things off my back. And thank you Andi for encouraging me to follow into your footsteps. Finally, I want to express my uttermost gratefulness to my small family, Robert and Ian. I know it has been hard on you, especially in the last few months. Without you, I would not have been able to go through the last four years.

Contents

Chapter 1

Introduction

1.1 Introduction

We are living in exciting times. What was pure science fiction just twenty years ago, like spoken interaction with computers, as seen in popular television series such as *Star Trek* for example, has now become a reality and is invading our every day lives. Today people talk to their mobile phones, get directions from a car navigation system or have an electronic book read out to them. Machines have taken over many tasks in the last centuries. Especially in the last decades, due to the steady progress in technology, together with the rapid increase of powerful mobile devices, the use of voice interfaces and other speech-enabled technologies has been promoted. Because speech commands can replace navigation through menus which would require clicks and swipes, speech seems to be ideally suited for mobile computing applications which often also happen to take place in situations where the hands or eyes or even both are occupied otherwise. The clear trend towards speech as input and output modality is palpable in the fact that one of the most hailed features of the new smartphone generation are dialogue systems such as Siri.[1] This trend is now followed by the implementation of dictation services into operating systems. The new *Apple* operating system *Mountain Lion* (Mac OS 10.8), for example, includes the screen-reader software *Voiceover* and a dictation service which is advertised for with the tag line 'Talking is the new typing.'[2]

1.2 Motivation

Speech-to-speech (S2S) translation is one of these emerging language applications. It takes on a fundamental problem of the economic and cultural exchange, the language barrier. Translating

[1] `http://www.apple.com/ios/siri/` (last accessed 19/08/2012)
[2] `http://www.apple.com/osx` (last accessed 19/08/2012)

speech input from one language into speech output in another language is the task that S2S translation research is set to accomplish [Vogel et al., 2006]. This thesis distinguishes between spoken language translation (SLT) and speech-to-speech (S2S) translation. In the literature the difference is often blurred and SLT is used as a synonym for S2S translation in many cases. Indeed SLT can be regarded as a hyponym for S2S, since S2S is the translation of spoken input with speech as output. In this thesis, however, I will follow the approach that the output of the system determines the type, using SLT for systems that translate spoken input into text output and S2S for systems that translate speech input into speech output. In the focus of this thesis are the three components which are chained together in order to produce speech output in the target language based on the speech input in the source language: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). ASR converts the input speech into a text which is then automatically translated by an MT component into a text in the output language. This text is finally converted back into a speech signal by a TTS system.

ASR, MT and TTS have a history of about 50 years of serious research in common which saw a huge leap forward in the last two decades due to a change to probabilistic methods. This paradigm change is grounded in the acceptance that natural language can never be fully analysed and therefore no exhaustive set of rules can be established which could be implemented in a language processing algorithm [Koehn, 2010, p. xi]. The surge in affordable computational power, volatile and persistent storage capacity as well as the advancement of the Internet further triggered this revolution. Moreover, the increasingly ubiquitous access to products and services on mobile devices, where language technology solutions often present distinct benefits such as hands-free and eyes-free interaction can be regarded as a key driver for the rising utilisation of ASR, MT and TTS. The improvements of these three technologies have in turn opened up a variety of new application areas with which we are faced on a day-to-day basis such as voice dialling, in-car navigation systems, interactive voice response, and instant machine translation on mobile devices.

This progress has caused a boost in speech-to-speech translation technology which has gained a maturity in recent years that led to the introduction of *iPhone* and *Android* apps like *SayHi*,[3] *Vocre*,[4] *SpeechTrans*,[5] and *Jibbigo*.[6] However far this evolution may have progressed, it is still evident that the output of the three components, ASR, MT, and TTS, is far from perfect. Automatic speech recognition for instance has gained a high accuracy for spontaneous speech in many languages. However, picking out a voice from a conversation or filtering out background noise remains an unsolved problem [Furui, 2005]. The quality of MT out-

[3] `http://www.sayhitranslate.com/` (last accessed 19/08/2012)
[4] `http://www.vocre.com/` (last accessed 19/08/2012)
[5] `http://speechtrans.com/` (last accessed 19/08/2012)
[6] `http://www.jibbigo.com/` (last accessed 19/08/2012)

put has improved significantly in the last two decades, due to the introduction of data-driven approaches. It reached a level where it is beginning to be more commonly employed in real-world applications. High quality translations, however, are still out of reach, if the domain is unrestricted or the translation language pair is uncommon (e.g. Maltese-Irish), and can only be achieved with post editing by humans [Koehn, 2010]. Today speech synthesis is used in a wide range of applications such as devices that read out aloud to visually impaired people, telephone based conversational agents, or devices that lend a voice to deaf people for instance. It is prioritised in applications where information is better distributed via an acoustic signal, such as car navigation systems. It is also used in the entertainment industry, for instance in computer games. However, TTS systems still struggle with out-of-vocabulary words like proper names or words inherited from other languages [Taylor, 2009]. Furthermore, humans are still well able to distinguish synthesised speech from actual human speech. The main challenge here is to emulate appropriate prosody, by getting sound duration, loudness, emphasis, pitch, and pauses right [Taylor, 2009].

The topic of this thesis is located at the intersection between natural language processing (NLP) and human-computer interaction (HCI). A popular textbook on speech and language processing [Jurafsky and Martin, 2008] introduces natural language processing as the field which aims to provide computers with the ability to process human language. The ultimate goal is to get computers to perform useful language-related tasks such as conversing with a human, translating a document, or answering questions using information from the Web [Jurafsky and Martin, 2008, p. 35]. Human-computer interaction is characterised as the multi-disciplinary subject that involves the design, implementation and evaluation of interactive systems in the context of a user's task [Dix et al., 2004]. In the following chapter it will be argued that HCI and NLP show similar concerns and could benefit from cross-fertilisation, but it will also be discussed that so far only limited interactions between the two fields can be observed. Although, HCI and NLP have the common goal to maximise the naturalness of communication, both fields employ very distinct methods and evaluation frameworks [Ozkan and Paris, 2002]. It will be argued that NLP can profit from a system design perspective by adopting user-centred approaches and that task analysis can help to ensure that subtasks are representative which are chosen for context evaluations. HCI on the other hand can, for instance, profit from more natural interfaces that are speech-enabled. Further, Ozkan and Paris [2002] argue that corpus analysis methods can be employed as a means to characterise and collect user language. With the recent trend to combine different NLP technologies and to include them into commercial applications as well as the growing utility of multi-modal interfaces, the application of usability evaluation complying with HCI standards is of increasing importance to the field of NLP.

This thesis will argue that a useful and immediate contribution that HCI methods can make is to the evaluative work in NLP. The error-prone nature of the three components which are under discussion in this thesis suggest that effective evaluation methods are necessary and it will be argued that the development of valid and useful evaluation measures can benefit from HCI research.

One of the main criteria for an evaluation metric in NLP is its correlation with human judgments, since the output is intended for processing by a human user. Furthermore, a good evaluation metric should be able to distinguish between systems of similar performance even if only small variations are considered [Doddington, 2002]. Moreover, metrics should be intuitive, automatically computable, and reproducible from one evaluation to another. Validity is another important feature which signifies that the evaluation technique should measure what is intended to be measured. The evaluation should work equally well when comparing human and system outputs and when comparing systems of different types with the same function (e.g. comparing rule-based and statistical systems). A good evaluation metric should act as a diagnostic tool that can be used by developers to identify deficiencies within their system. It should enable developers to tune their systems to the metric.[7]

1.3 Research Question

Different metrics that are used in the evaluation of automatic speech recognition, machine translation and text-to-speech will be discussed. It will be observed that it is widespread practice for these technologies to be evaluated intrinsically, without regard to the context of the application (e.g. the task or the user). In recent years there have been calls for paying more attention to extrinsic evaluation in machine translation in particular and natural language technologies in general [Goldstein et al., 2005; Belz, 2009]. Extrinsic evaluation aims to assess the (often indirect) effect of a module on task- and context-dependent variables such as user performance. Furthermore, with the widening context of use (e.g. mobile devices) and the new focus on combining speech with other modalities researchers demand general evaluation frameworks that offer a general toolset with a well defined scope and reflecting a principled approach. Hovy et al. [2002] advocate a global, principle-based approach to MT evaluation which takes into account that no unique evaluation scheme is acceptable for all evaluation purposes. Dybkjaer et al. [2004] argue that for spoken dialogue systems the evaluation and usability of multimodal and mobile systems presents a new challenge to research. In this context the research presented

[7] The description of the perfect evaluation metric are based on notes from the keynote talk by Daniel Marcu at the International Workshop on Spoken Language Translation 2011 (09/12/2011). Further, it should be stated here, that other criteria like speed or storage needed, offline applicability etc. are relevant in practical deployments but will not be the subject of evaluation in this thesis.

in this thesis therefore aims to explore the following research question:

> Can ASR, MT und TTS as main components of S2S translation benefit from extrinsic evaluation methods?

Accordingly this research will compare intrinsic and extrinsic evaluation methods of the three main components of speech-to-speech translation and consider whether extrinsic evaluation results correspond to intrinsic evaluation results. In this context the correlation of intrinsic measures with human judgement will be scrutinised, that is whether intrinsic evaluation metrics adequately capture the variation between system-output and human-generated reference. It will further be examined what extrinsic methods for the evaluation of ASR, MT, and TTS can be implemented.

1.4 Contributions and Key Conclusions

By answering these questions this research aims at contributing to the discussion of whether NLP and HCI could benefit from cross-fertilisation. In particular it examines ways in which NLP evaluation can benefit from methods and ideas developed in the field of HCI. Therefore, this research expands on the existing knowledge on NLP evaluation. By exploring previous work on NLP evaluation in general and the evaluation of automatic speech recognition, machine translation and text-to-speech in particular, this thesis gives a comprehensive overview on the topic of evaluation in NLP. It further investigates the application of the Wizard of Oz and task-based techniques for the evaluation of ASR, MT, and TTS. This thesis concentrates on the three components of S2S translation, however, the findings of this thesis can also be applied to other interactive, possibly multilingual, speech applications such as spoken language translation or (multilingual) dialogue applications.

The experiments confirm the assumption that ASR, MT, and TTS can benefit from extrinsic evaluation, but with several caveats. it has been shown that evaluation efforts need to keep the user and application context in mind if they aim to produce meaningful results for real-life situations. The main advantage of such extrinsic evaluation is that they often offer a better insight into obstructive situations that cannot be revealed by the results of intrinsic evaluation. Furthermore, this research shows that extrinsic evaluations can be used to fine-tune intrinsic evaluation metrics. In spite of demonstrating these advantages of extrinsic evaluation, the execution of the experiments has confirmed that extrinsic methods are very time-consuming and laborious which includes the difficult task to plan how to capture interesting data and to decide which data is important for analysis. The research in the context of this thesis has also

demonstrated that Wizard of Oz and task-based techniques (specifically map tasks) are well suited for extrinsic NLP evaluation under certain circumstances.

The experiments did further reveal that also HCI might benefit from an exchange with NLP. The proposed metric to assess the task performance in map tasks showed several practical drawbacks. Therefore, a new metric is proposed in this thesis which was used to analyse the results from the extrinsic ASR evaluation. The new method proved to be much simpler while still being discriminative enough to determine the amount of miscommunication.

1.5 Outline

This thesis is structured into five chapters. The following two chapters (*Related Work* and *Methods*) will set the scene for the main part of this thesis. The *Related Work* chapter covers the classifications of NLP evaluation methods and defines intrinsic and extrinsic evaluation, the two core concepts of this research. This is followed by a discussion of the literature that the two fields (HCI and NLP) which constitute a frame of reference for this thesis, can benefit from a synthesis. The chapter finishes with an introduction to ASR, MT and TTS, the three core components of S2S translation. In the sections on ASR, MT and TTS the history of those components is used to preface the description of the main approaches which is followed by a detailed discussion of intrinsic and extrinsic evaluation methods. In the *Methods* chapter the intrinsic evaluation methods which were utilised in the experiments described in Chapter 4 (*Intrinsic and Extrinsic Component Evaluation*), and which constitute the fundament of this research, are revisited and discussed in detail. Further, the map task method and the Wizard of Oz technique which form the core for the extrinsic evaluations, will be comprehensively surveyed. Chapter 4 forms the backbone of this thesis. Here the intrinsic and extrinsic experiments around the evaluation of ASR, MT and TTS will be discussed, their set-up outlined, and the results reported. Only a few isolated efforts can be observed where NLP technologies are evaluated in their context of use. In speech-to-speech translation, evaluation usually concentrates on the performance of the components as stand-alone systems or on the performance of the system as a whole (end-to-end evaluation). Initiatives that assess the combination of language technologies which are part of a bigger system, are rare. One of the outcomes of the experiments around the stand-alone assessment of the S2S components is that evaluation efforts should not lose sight of the combination of single components which are part of a large whole in the assessment. Section 4.4 includes the discussion of experiment outcomes where the machine translation component is combined with the TTS component. As a consequence of these outcomes another experiment was carried out which examined a phenomenon unveiled by research into the combination of MT and TTS. This will be outlined in Subsection 4.5. The

results of the experiments and the conclusions that can be drawn from the research described in the previous chapter will be discussed further in the *Conclusion*.

List of Publications

A number of publications resulted from the research presented in this thesis. The following list specifies them in chronological order:

Nikiforos Karamanis, Anne H. Schneider, Ielka van Der Sluis, Stephan Schlögl, Gavin Doherty and Saturnino Luz (2009)
Do HCI and NLP interact? *In Proceedings of CHI '09, pp. 4333-4338.* Boston, USA.

Stephan Schlögl, Gavin Doherty, Nikiforos Karamanis, Anne H. Schneider and Saturnino Luz (2010)
Observing the Wizard: In Search of a generic Interface for Wizard of Oz Studies. *In Proceedings of iHCI 2010, pp. 43-50.* Dublin, Ireland.

Anne. H. Schneider, Ielka van Der Sluis and Saturnino Luz (2010)
Comparing Intrinsic and Extrinsic Evaluation of MT Output in a Dialogue System. *In Proceedings of the 7th International Workshop on Spoken Language Translation, pp. 329-336.* Paris, France.

Stephan Schlögl, Anne H. Schneider, Saturnino Luz and Gavin Doherty (2011)
Supporting the Wizard: Interface Improvements in Wizard of Oz Studies. *In Proceedings of the 25th BSC Conference on Human-Computer Interaction, pp. 509-514.* Newcastle upon Tyne, GB.

Anne H. Schneider and Saturnino Luz (2011)
Speaker Alignment in Synthesised , Machine Translated Communication. *In Proceedings of the International Workshop on Spoken Language Translation 2011, pp. 254-260.* San Francisco, USA.

Anne H. Schneider and Saturnino Luz (2012)
Spoken Language Translation Evaluation– Overview of Error Classifications and Applications. *In Proceedings of Workshop on Innovation and Applications in Speech Technology 2012.* Dublin, Ireland.

1.6 Summary

This introduction embedded the research that will be described in the following chapters into the context of the two research areas natural language processing (NLP) and human-computer interaction (HCI) which constitute a frame of reference for this thesis. Further, the relevance of this research to the NLP and HCI communities was motivated and the research questions as well as the key contributions and conclusions of this work were discussed. The chapter concluded with an outline of this thesis and an overview of the publications that resulted from the research presented. The next chapter will set the scene for the main part of this thesis and discuss related work. It will cover the classifications of NLP evaluation methods and define intrinsic and extrinsic evaluation, the two core concepts in this thesis. It will be discussed that HCI and NLP can benefit from a synthesis. Further, ASR, MT and TTS, the three core components of S2S translation are examined more deeply. Their history will be used to preface the description of the main approaches which will be followed by a detailed discussion of intrinsic and extrinsic evaluation methods.

Chapter 2

Related Work

2.1 Introduction

The main objective of this thesis is to contribute to the research in NLP evaluation by comparing intrinsic and extrinsic evaluation methods for the three core components of speech-to-speech translation. This chapter aims to discuss the context of this thesis and to establish the basic concepts and knowledge necessary to understanding the research that will be outlined in the following chapters. The chapter is therefore divided into two parts: a general introduction into NLP evaluation (Section 2.2 and Section 2.3) and the characterisation of the three components of speech-to-speech translation which have been subject of the comparison between intrinsic and extrinsic evaluation (Section 2.4 to 2.6). The first section in this chapter will discuss classifications of NLP evaluation methods and define intrinsic and extrinsic evaluation, the two core concepts of this thesis. The next section will argue that NLP and HCI share similar concerns and importing methods or ideas from one discipline into the other can be beneficial. Although, HCI and NLP have the common goal to maximise the naturalness of communication, both fields employ very distinct methods and evaluation frameworks [Ozkan and Paris, 2002]. It will be argued that NLP can profit from a system design perspective by adopting user-centred approaches, and that task analysis can help to ensure that subtasks chosen in context evaluation are representative. HCI on the other hand can fall back on NLP applications to design more natural interfaces. Ozkan and Paris [2002] further argue that, for instance, corpus analysis methods can be employed as a means to characterise and collect user language. This research aims at investigating how NLP evaluation can profit from methods and ideas imported from the field of HCI.

The second part of this chapter can be found in the following three sections which will give an overview of the three main components of S2S systems: automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS). ASR, MT and TTS are

technologies that have been well researched for about 50 years but made great progress in the last two decades, mostly due to greater use of probabilistic methods which was triggered by improvements in technology and the emergence of the World Wide Web. However, all three technologies are still error prone, deeming good evaluation methods essential. The discussion of the history and technology behind those three components will therefore be followed by a focused discussion of evaluation methods and efforts which aims to accentuate the respective efforts in intrinsic and extrinsic evaluation.

2.2 NLP Evaluation

The evaluation of NLP systems serves two purposes: one is to enable researchers to compare different systems with each other, the other is to obtain information in a development cycle whether the system under analysis has improved [Carstensen et al., 2001]. Evaluation plays a crucial role for users and developers and is guiding research in natural language processing [Hirschman and Thompson, 1997].

A good overview of different types of NLP evaluation can be found in Eli Kumar's book on natural language processing [Kumar, 2011]. He lists three well accepted distinctions in NLP evaluation, namely black-box vs. glass-box evaluation, automatic vs. manual evaluation, and intrinsic vs. extrinsic evaluation (see Figure 2.1). For a *black-box evaluation* the NLP system is run on a given data set and a number of parameters related to the quality of the process (e.g. speed and resource consumption) and the results (e.g. the accuracy of a translation) are measured. In *glass-box evaluation* the design of the system, the implemented algorithms and the used resources and other components are examined which makes it especially interesting to designers and developers. For glass-box evaluation either all modules except for one are kept constant and several runs of the system are performed or a single component is evaluated in isolation. Glass-box evaluations are usually more informative with respect to error analysis but it is more difficult to predict the performance of a system based only on glass-box tests. In *automatic evaluation* a procedure is defined that automatically compares the output of an NLP system with the desired output or a gold standard, whereas for *manual evaluation* human judges estimate the quality of a sample of system outputs or a system in general. The development of a gold standard is a complex task and can be costly but once it is established the automatic evaluation can be repeated as often as needed. The problem with human judges is that very often their ratings show considerable variation. Biological, cognitive and social differences (e.g. handedness or gender) can be factors that influence the judgement process [Schütze, 1996]. Further, attention has to be paid to careful experiment design, for example experiment descriptions can be interpreted differently by participants to what experimenter

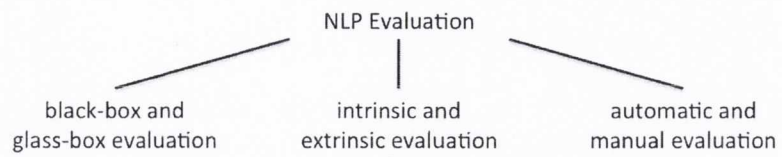


Figure 2.1 – Classification of NLP evaluation types according to Kumar [2011].

intended [Schütze, 2005]. Therefore, manual evaluation is sometimes also called *subjective evaluation*, while automatic evaluation is referred to as *objective evaluation*. The distinction of *intrinsic* and *extrinsic evaluation* is based on the context in which a component is evaluated. Intrinsic evaluation measures the performance of an isolated NLP system, whereas extrinsic evaluation considers the NLP system in a more complex setting. The definition will be discussed in more detail later.

Another classification of NLP evaluation methods can be found in Hirschman and Thompson [1997]. They distinguish three types of NLP evaluation (see Figure 2.2). *Adequacy evaluation* determines whether a system fits its purpose and is therefore typically user oriented. *Diagnostic evaluation* generates a performance profile of the system for a set of possible inputs. It is therefore targeted at the developer, but can also be offered to the end-user. Finally, *performance evaluation* measures the system's performance in one or more specific areas and is usually employed to compare similar approaches or systems with each other. The last evaluation type is also targeted at developers. The authors further point out the importance of distinguishing between system evaluation and component evaluation. For the latter type they state that "a further distinction between intrinsic and extrinsic evaluation must be respected – do we look at how a particular component works in its own terms (intrinsic) or how it contributes to the overall performance of the system (extrinsic). At the whole system level, this distinction approximates the distinction between performance evaluation and adequacy evaluation, where intrinsic is to extrinsic as performance evaluation is to adequacy evaluation." [Hirschman and Thompson, 1997, p. 386]

No account of NLP evaluation is complete without a mention of *shared tasks* or *evaluation campaigns*. A shared task is a competition between interested teams, where the goal is to achieve the best performance possible on a well-defined problem that everyone agrees to work on. In many areas of speech and language processing shared tasks have evolved to be an instrument for objective comparison of results between different groups, sometimes with the

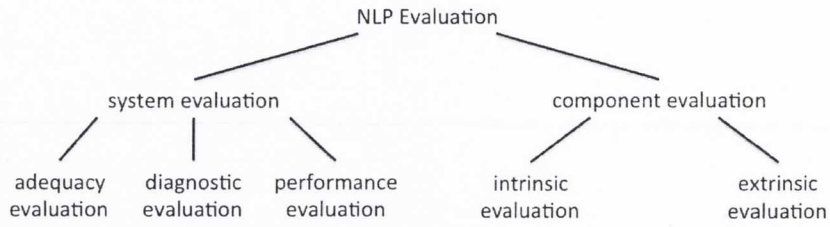


Figure 2.2 – Classification of NLP evaluation types according to Hirschman and Thompson [1997].

effect of creating new subfields [Rayner et al., 2008b]. Well known shared tasks in NLP are among others the *BioCreAtIvE* challenge with the aim to evaluate information extraction and text mining advancements in the biological domain [Hirschman et al., 2005], the *Text REtrieval Conference* (TREC) which focuses on the information retrieval research and dates back to 1992 [Harman, 1993] and *Semantic Evaluation* (SemEval), which developed from the Senseval evaluation series for word sense evaluation and tries to evaluate computational semantic analysis systems [Agirre et al., 2007].

Of special interest for this thesis are the two core concepts, intrinsic and extrinsic evaluation. Their definition in the literature will therefore now be discussed in more detail. According to Kumar *intrinsic evaluation* is based on measuring the performance of an isolated NLP system which is mainly characterized by the performance with respect to a gold standard result which in turn has been predefined by the evaluators. Extrinsic evaluation considers the NLP system in a more complex setting, i.e. as part of a system or as precise function for a human user, and is therefore also called *evaluation in use*. The performance of the NLP technology is then rated with regards to its utility in the complex system or for the human user. Kumar's categorisation stems from Spärck Jones and Gallier's book on the evaluation of natural language processing systems. According to them there are two main types of criteria for performance evaluation [Spärck Jones and Galliers, 1995, p. 19]: "Intrinsic criteria are those relating to a system's objective, extrinsic criteria those relating to its function i.e. to its role in relation to its setup's purpose."

Jurafsky and Martin also discuss the distinction between extrinsic and intrinsic evaluation in their chapters on the evaluation of N-grams and word sense disambiguation. They define an intrinsic or *in vitro* evaluation metric as "one that measures the quality of a [language] model independent of any application" [Jurafsky and Martin, 2009, p. 129]. They also constitute that evaluating "component NLP tasks embedded in end-to-end applications is called extrinsic

evaluation, task based evaluation, end-to-end evaluation, or in vivo evaluation." [Jurafsky and Martin, 2009, p. 678]. According to the authors, only extrinsic evaluation can indicate an improvement in performance on a real task, however, it comes at a higher cost and time expense and may not be generalisable to other applications since the evaluation are limited to the context of the application under test. Therefore, according to Jurafsky and Martin, intrinsic evaluation is the most common type in NLP evaluation practice.

The four definitions by Kumar, Spärck Jones and Gallier, Hirschman and Thompson, and Jurafsky and Manning agree that intrinsic evaluation considers the performance of an isolated system or component whereas extrinsic evaluation considers the component as part of a more complex system. Kumar, Spärck Jones and Gallier, and Jurafsky and Martin also discuss that extrinsic evaluation includes the functionality or utility of the system for the human user. This dependence of the performance on the task which a system is set to is left out in the discussion of the two terms by Hirschman and Thompson because they distinguish between system and component evaluation and assign the distinction of extrinsic and intrinsic evaluation to component evaluation only. For system evaluation they have the same distinction between task dependence and independence but they call these methods adequacy evaluation and performance evaluation respectively.

The inclusion of the application in which the technology is employed into the remit of extrinsic evaluation, in my opinion, does not go far enough, because it lacks recognition of the influence of the context in which the system is used. Factors such as time constraints or distractions for example that arise in interactive situations can be expected to change the perception of the systems performance and can therefore not be disregarded. In this thesis I will use the term *intrinsic evaluation* to refer to evaluation methods that observe a single NLP component in isolation independent of their context or task. *Extrinsic evaluation* will be used to describe methods that aim to assess the (often indirect) effect of a component, on task- and context-dependent variables such as user performance, through its monitoring as part of a functioning system.

With the development of shared tasks the development of evaluation methods evolved. Typically the evaluations in shared tasks are based on intrinsic methods, however, the aim to judge a performance for a special task at hand is arguably an extrinsic evaluation. In recent years there have been renewed calls for more attention to be paid to *extrinsic evaluation* in natural language technologies [Goldstein et al., 2005; Belz, 2009]. Hirschman and Thompson [1997] point out that a limitation of current evaluation methods in NLP is the limited focus on how the user interacts with a system. Since the field of human- computer interaction (HCI) is characterized as the multi disciplinary subject that involves the design, implementation and

evaluation of interactive systems in the context of the user's task [Dix et al., 2004] the next section will discuss how HCI and NLP could benefit from a synthesis.

2.3 The Relation between NLP and HCI

Although the term human-computer interaction cannot be found in the index of at least two of the most commonly read NLP textbooks [Jurafsky and Martin, 2008; Manning and Schütze, 1999], the relationship between HCI and NLP has been investigated in some detail by members of the NLP community. It has been argued, that HCI and NLP share similar concerns [Ozkan and Paris, 2002] and could benefit from a synthesis [Winograd, 2006], however, only limited interaction between those two fields can be observed. Similar remarks were made by Dybkjaer and Bernsen [2000] and Larsen [2003]. HCI researchers have also called for a synthesis between HCI and Artificial Intelligence which encompasses NLP (see [Winograd, 2006] for an overview).

My colleagues and I investigated whether NLP and HCI have come any closer since these remarks were made [Karamanis et al., 2009].[1] Since the work which falls within the realm of HCI is vast, it can be challenging to try to identify the subset that has impacted on NLP and vice versa. In order to get a general idea of the level of overlap between the two fields, we performed a bibliometric analysis of research in NLP and HCI, extending a preliminary survey by Reiter where he analysed the citations in two major NLP journals in 2005 to identify which fields have impact on recent NLP research [Reiter, 2007]. Our investigation extends this study by distilling additional citations from articles published in 2007 in five major NLP and five major HCI journals. Major here means the journals with the highest impact factor. For these we assessed how many times each journal cites (a) itself, (b) the other four journals in the same category and (c) the five journals in the other category. The results can be found in Table 2.1, normalised by the total number of citations to journals in the *ISI Web of Knowledge* database.[2]

These results agree with Reiter's results, showing very limited influence from HCI on NLP and the other way around. The small amount of cross-citations is mostly related to work on speech and dialogue processing (from Speech Communication to the International Journal of Human-Computer Studies and vice versa).

With the recent trend to combine different NLP technologies and to include them into commercial applications as well as the growing utility of multi-modal interfaces, the application of

[1] The paper resulted from a collaborative effort in which Nikiforos Karamanis and I reviewed the publications of ten journals published in the year 2007 and analysed the results together. A first draft of the paper was written by Nikiforos and than reviewed and refined by myself and other colleagues who are mentioned as authors on the paper.

[2] `http://wokinfo.com/` (last accessed 07/07/2012)

		self citation	NLP	HCI
NLP	CL	54.55%	5.05%	0.00%
	CSL	16.09%	43.53%	0.00%
	SC	19.86%	18.38%	0.46%
	LRE	7.46%	65.67%	0.00%
	IEEE	26.84%	12.37%	0.00%
HCI	HCI	25.40%	0.00%	7.94%
	UMUAI	20.14%	0.00%	9.03%
	IJHCS	12.58%	0.77%	8.28%
	IWC	14.36%	0.00%	22.67%
	BIT	32.64%	0.00%	30.56%

Table 2.1 – Cross citations between five major NLP and HCI journals (2007 issues). The journals considered are Computational Linguistics (CL), Computer Speech and Language (CSL), Speech Communication (SC), Language Resources and Evaluation (LRE), IEEE Transactions on Audio Speech and Language Processing (IEEE), Human-Computer Interaction (HCI), User Modeling and User-Adapted Interaction (UMUAI), International Journal of Human-Computer Studies (IJHCS), Interacting with Computers (IWC) and Behaviour and Information Technology (BIT).

usability evaluation complying with HCI standards is of increasing importance to the field of NLP. The bibliometric survey has indicated that HCI researchers may not be familiar enough with ongoing NLP research, and thus with the opportunities emerging from natural language processing. HCI textbooks view NLP mostly as contributing towards the development of yet another mode of interaction. This is reflected in the discussion of Sharp et al. who briefly examine the differences between text and speech-based interaction and provide some general design guidelines for language-based interfaces [Sharp et al., 2007, pp. 113-114]. Dix et al. contrast language-based interaction with direct manipulation which is considered to be a more attractive alternative [Dix et al., 2004, pp. 138-139]. More generally, it seems that in situations where the use of NLP provides the only reasonable way to accomplish a task, NLP becomes a research challenge for HCI. While standard interaction design assumes a certain amount of component reliability, the inherent inaccuracy of NLP technology (of which NLP researchers are very much aware) might call for new HCI methods to be developed.

It is striking that one useful and immediate contribution HCI methods could make to the evaluative work reviewed above relates to modelling: both at the level of underlying human factors and at the higher level of task analysis. Performance modelling of basic selection tasks could, for instance, be used to inform the design and contextualise the results of studies. Performing task analysis can furthermore ensure that the subtasks chosen for evaluation that consider the context are indeed representative of the work that is carried out on a daily basis.

In addition to evaluation, HCI methods are relevant from a system design perspective.

Introducing user-centred approaches [Beyer and Holtzblatt, 1998], for instance, would shift the focus from adding functionalities to existing interfaces into placing more emphasis on the overall process and context of work. This could shed light on some of the observed complex interdependencies and help clarify under which circumstances NLP does provide added value. Looking at user's strategies to overcome errors in more detail may contribute additional insight with respect to these issues and help feed evaluation back to overall system design.

This research advocates that HCI can fertilise research in NLP by introducing methods such as task and error analysis as well as contextual inquiry which can provide a sound basis for the development of NLP-enabled systems. However, the bibliometric survey has indicated that HCI researchers may not be familiar enough with ongoing NLP research, and thus with the opportunities emerging from this field. This is reflected in the discussion of NLP in the leading HCI textbooks. Sharp et al. briefly discuss the differences between text and speech-based interaction and provide some general design guidelines for language-based interfaces [Sharp et al., 2007, pp. 113-114]. Dix et al. contrast language-based interaction with direct manipulation which is considered to be a more attractive alternative [Dix et al., 2004, pp. 138-139]. A similar view is held in the more detailed account of NLP research by Shneiderman and Plaisant [2005, pp. 331-340]. However, they also add that HCI studies focused on discovering and analysing the tasks and situations for which NLP-enabled applications are most beneficial and can make their use more widespread [Shneiderman and Plaisant, 2005, p. 332].

The previous two sections concentrated on NLP evaluation in general to establish the context of this thesis. The following sections will focus on the three main components of speech-to-speech translation, which have been subject of the comparison between intrinsic and extrinsic evaluations. The review will start with automatic speech recognition which is followed by the discussion of machine translation. Finally text-to-speech synthesis will be characterized. All three components are discussed starting with a general introduction and history of the research in the area. This is followed by an outline of the most important approaches. In the end evaluation methods for the three components are discussed, starting with intrinsic methods followed by extrinsic methods.

2.4 Automatic Speech Recognition

The goal of automatic speech recognition (ASR), also known as speech-to-text, is to develop systems that are capable of mapping from an acoustic signal to a string of words [Jurafsky and Martin, 2009, p. 319]. ASR systems can be distinguished based on their ability to recognize different types of utterances. Four groups are differentiated:

1. **Isolated word recognition** requires the speaker to pause between the words or utterances that should be recognised.

2. **Connected word recognition** is similar to isolated word recognition but allows for the pauses between the utterances to be minimal.

3. **Continuous speech recognition** allows the speaker to use almost natural speech (dictation style) while the recognition system determines the utterance boundaries.

4. **Spontaneous speech recognition** handles natural, unrehearsed speech and therefore has to deal with false starts, filled pauses, hesitations, ungrammatical constructions and alternative pronunciations for example.

The history of ASR may be broadly partitioned into three generations. The earliest attempts to solve the problem to enable machines to recognise words date back to the 1950s. The first generation of speech recognisers were hardware-based and only able to recognise isolated digits. At the Bell Laboratories for example a machine was built that was able to recognise isolated digits for a single speaker [Davis et al., 1952]. In the 1960s several Japanese labs entered the recognition area with a strong focus on special purpose hardware like the vowel recogniser developed at the Radio Research Lab in Tokyo in 1961 and the phoneme recogniser created at the University of Kyoto in 1962.

In the 1960s and 1970s the second generation of ASR systems which was able to perform isolated word recognition surfaced. This period was dominated by the use of dynamic programming algorithms that were necessary to align pairs of speech segments, since second generation ASR systems were mainly based on comparing reference speech samples (e.g. phonemes or words) with test speech. Dynamic programming was first used in ASR by a research group from the Soviet Union that proposed the method to align pairs of speech utterances [Vintsyuk, 1968]. The 1970s was the decade when isolated word recognition became usable due to research in Russia which helped to advance the use of pattern recognition ideas, while research in Japan demonstrated the successful application of dynamic programming.

The third generation of ASR systems is able to recognise large vocabulary, continuous speech independent of the speaker and is based on machine learning techniques. The research in the field of continuous speech recognition was triggered by the speech recognition programme at the Carnegie Mellon University with a focus on tracking phonemes dynamically [Reddy, 1966]. Amongst others the Nippon Electrical Corporation (NEC) and the Bell Labs developed a variety of connected word-recognition algorithms in the early 1980s [Sakoe, 1979; Myers and Rabiner, 1981]. The 1980s also saw a shift to statistical modelling methods like the Hidden Markov model which was first employed at IBM and the *Institute for Defense Analysis*

(IDA) and Dragon Systems before it became widely applied. The idea of introducing neural networks to solve problems in speech recognition which had failed in the 1950s, re-emerged due to a better understanding of their strengths and limitations [Lippmann, 1987; Waibel et al., 1989]. A good account of the 'dawn of statistical ASR' can be found in Fred Jelinek's acceptance speech for the *ACL Lifetime Achievement Award* [Jelinek, 2009]. Based on these developments the widespread commercialisation of speech recognition began in the 90s. Early versions of commercial ASR systems had to make compromises, they were either dependent on a particular speaker, had a small vocabulary, or used a very formal and rigid syntax. Since then many practical applications have been developed and the research in ASR has advanced to the point where speaker independent continuous speech recognition of spontaneous speech with a large vocabulary, to some extent even in a noisy background, is possible. A comprehensive account of the history of speech recognition until 2005 is given by Furui [2005].

Depending on the application, the goals of automatic speech recognition is to recognize words in real-time with a 100% accuracy, independent of the characteristics of the speaker, the speaker's accent, background noise or the size of the vocabulary to be recognized in the application. To be fair it should be stated that this is not even achieved by humans. Commercial dictation systems claim a recognition performance of around 99%.[3] However, human recognition performance is still out of reach for ASR systems. According to Lippmann [1997] the error rate of human speech recognition for spontaneous speech is set below 5% and does not increase very much in degraded conditions. As is shown in experimental settings for most recognition tasks, human subjects produce one to two orders of magnitude fewer errors than machines. Especially in degraded conditions (e.g. background noise) the ASR performance deteriorates dramatically. Two important challenges in speech recognition are still not met which are essential to achieve human-like performance [Lippmann, 1997]. One is the ability to distinguish speech from environmental sounds [Cooke et al., 2010] which can be the challenge to separate speech from environmental noise sounds or to separate speech from different speakers when their speech overlaps. The second important challenge is to recognise and learn unknown words [Yazgan and Saraclar, 2004]. At the University of California, San Francisco experiments were performed on patients undergoing brain surgery to discover how humans filter out single voices which are buried in noise [Mesgarani and Chang, 2012]. The authors claim that if the process to filter out extraneous noise is fully understood this may inspire more efficient and generalizable solutions in current speech engineering approaches. Another factor that has to be considered to decrease the gap between machine and human speech recognition is prosody. *Prosody* refers to the acoustic structure of words or segments in spoken language

[3] `http://www.nuance.ie/technology-primer.asp` (last accessed 15/04/2012)

which conveys important information, such as syllable stress, intonation, and rhythm [Zue and Cole, 1997]. In some situations these signals might be used to recognise and detect sounds such as hesitations or laughter. A recent account on issues of expressive speech processing can be found in [Campbell, 2010].

2.4.1 Approaches to Speech Recognition

There are three different approaches to speech recognition, namely the acoustic-phonetic approach, the pattern recognition approach and the artificial intelligence approach [Rabiner and Juang, 1993, p. 42]. The *acoustic-phonetic approach* [Hemdal and Hughes, 1967] is a method that is based on the theory of acoustic phonetics. This theory hypothesises that in spoken language there is a fixed number of distinctive phonetic units, and each of these units can be represented by a set of acoustic features (e.g. nasality or frication) that can be determined from the speech signal. Broadly speaking the system sequentially decodes the speech signal by assigning the appropriate phonetic label, the phonetic symbols, to every sound. This is done in four steps which can be seen in the graph in Figure 2.3. First the speech analysis system performs a spectral analysis of the speech signal. This spectral representation is then converted into a set of features which describe the acoustic properties of the phonetic units. Therefore, a set of feature detectors determines in parallel for each unit features like formant location, frication, nasality etc. In the segmentation and labelling phase the system breaks the signal into sections where the features change only little over time and associates phonetic labels to these stable sections based on the best matching set of features. The output of this step is a so called phoneme lattice which includes possible combinations of phone sequences, from which finally in the lexical access procedure the best matching word or sequence of words is chosen. The acoustic phonetic approach did not establish in practical systems for several reasons, among them the vast necessary knowledge about the acoustic properties of phonetic units and lack of an automatic procedure to tune the feature detection which uses the phonetic knowledge to map features to phones [Rabiner and Juang, 1993, p. 50].

The *pattern recognition approach* [Rabiner, 1989] has been proven to be of greater practical value than the acoustic-phonetic approach due to the possibility to automate this method, its robustness with respect to different speech vocabulary and the high performance that it has shown in practical applications [Rabiner and Juang, 1993, p. 44]. As illustrated in the schematic representation in Figure 2.4, the statistical pattern recognition approach can also be divided into four steps. Feature measurement is the first step analogous to the acoustic phonetic approach. The result of this step is a test pattern. The two essential steps in this paradigm following next: the pattern training and the pattern classification. On one hand, in the pattern-training phase

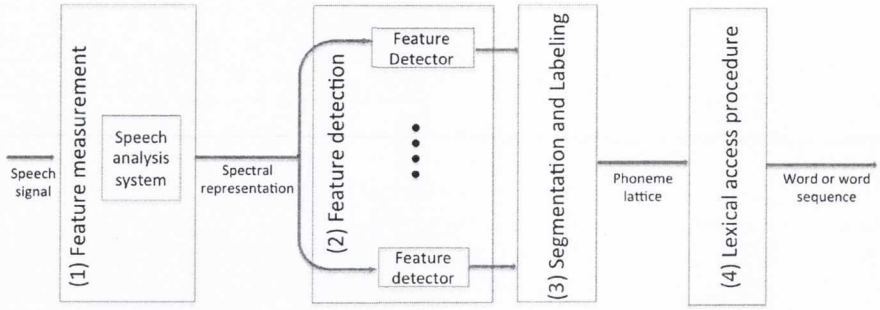


Figure 2.3 – A schematic representation of the acoustic phonetic approach.

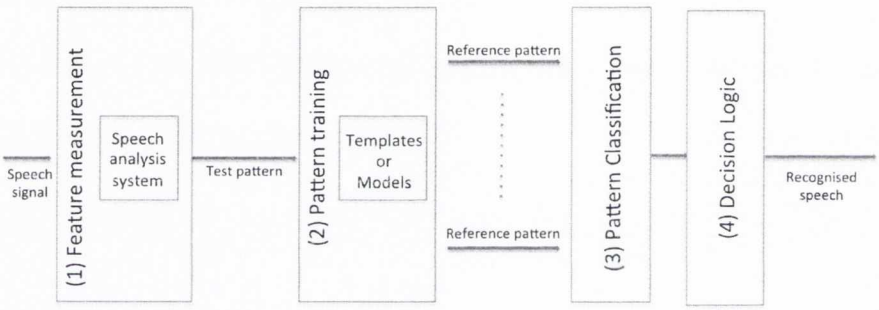


Figure 2.4 – A schematic representation of the pattern recognition approach based on [Rabiner and Juang, 1993].

so called reference patterns are generated from a set of labelled training samples. On the other hand, in the pattern-classification stage each possible pattern learned from the classification stage is compared with the unknown test pattern and a similarity score is computed. In the final step, the decision logic, the similarity score decides the best match between test and reference pattern. Depending on the used feature measurement method, the model for the reference pattern and the technique that is used to create the reference patterns different pattern recognition approaches are distinguished.

The *artificial intelligence approach* is a hybrid method of the acoustic phonetic and the pattern recognition paradigm, where the key idea is to incorporate knowledge from different sources (e.g. phonetic, linguistic, syntactic, semantic and pragmatic knowledge) to solve the speech recognition problem. Several approaches to integrate knowledge in speech recognizers are employed. In the "bottom up" paradigm low-level processing steps like feature extraction

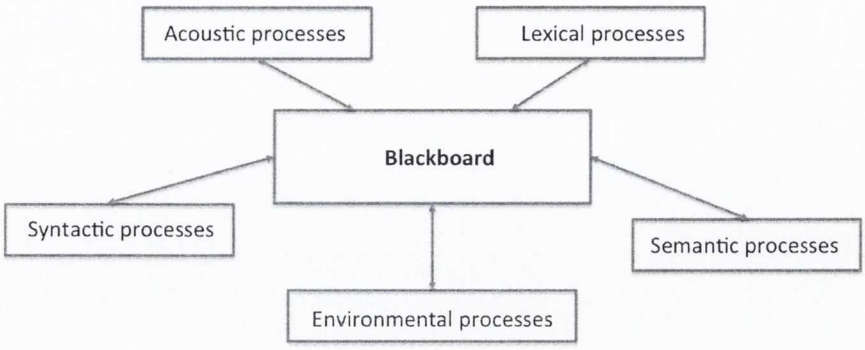


Figure 2.5 – A schematic representation of the blackboard paradigm based on [Rabiner and Juang, 1993].

and segmentation precede higher-level processes such as lexical decoding and sentence verification. In the "top-down" paradigm on the other hand the language model generates hypotheses for words that are matched against the speech signal with the aim to build syntactically correct and semantically meaningful sentences. In the blackboard approach the knowledge sources are studied independently by a central hypothesis-and-test paradigm, the so-called black board. An elaborate rating mechanism accounts for the combination and propagation of hypotheses for possible words and outputs the recognized speech. A schematic representation of the blackboard paradigm is shown in Figure 2.5.

2.4.2 Evaluation of ASR systems

As aforementioned automatic speech recognition systems are typically evaluated using intrinsic evaluation methods. The most widely used of these is the *word error rate* (WER) which is based on the *Levenshtein distance* [Levenshtein, 1966]. For some languages such as Chinese or Japanese the character error rate (CER) is also commonly used. The word error rate is defined as the proportion of word errors to recognised words and therefore is calculated as a distance measure (edit distance), between the series of recognised words and the reference word sequence (i.e. correct word sequence), normalised by the length of the reference text. The *edit distance* is the minimum number that is determined by adding the number of insertions to the deletions and substitutions that are necessary to transform the series of reference words into the recognised word sequence. Therefore, the word error rate is defined as follows:

$$WER = \frac{\text{edit distance}}{\text{length of reference}}$$

WER is an objective and direct metric to measure the accuracy of an ASR system, however, there are several practical disadvantages which affect the interpretation and comparability of the results. One problem of WER is that it is not a true percentage since it has no upper bound. It is possible that the results of an ASR evaluation with WER can exceed 100%, as will be discussed in Chapter 3, and therefore a statement about the quality of a system is only possible in comparison to another system's performance within the same conditions (e.g. same speech data set used in the evaluation). The word error rate will be revisited in the *Methods* Chapter and discussed in greater detail, since it is the intrinsic evaluation method employed in the ASR evaluation that is reported on in Chapter 4.

A simple and direct way to bypass the problem posed by the unbounded numerator is a normalisation by the maximum possible value that WER can reach [Morris et al., 2004]. Let N_{REF} be the number of words in the reference and N_{ASR} be the number of words in the recognized output, then the maximum value that WER ($maxWER$) can reach is:

$$maxWER = \frac{max(N_{REF}, N_{ASR})}{N_{REF}}$$

Therefore, the alternative normalised word error rate ($altWER$) would arise as:

$$altWER = \frac{\text{edit distance}}{max(N_{REF}, N_{ASR})}$$

The *match error rate* (MER) is another alternative to WER which circumvents the normalisation problem [Morris et al., 2004]. It calculates the probability that a given match is incorrect. Let H be the number of correctly recognised words (hits), S the number of substitutions, D the number of deletions and I the number of insertions. The edit distance is the sum of substitutions, deletions and insertions $editdistance = S+D+I$ and let N denote the number of matched word pairs between the reference and the ASR output:

$$N = H+S+I+D$$

The match error rate MER can therefore be calculated as:

$$MER = \frac{\text{edit distance}}{N}$$

It has been shown that in test runs with a high error rate, WER starts to misbehave and more weight is given to insertions than to deletions [Morris, 2002]. The alternative *word information preserved* (WIP) measure is therefore proposed. WIP is an approximation of the proportion of information that is preserved by the recognition from the reference [Morris, 2002]. WIP is

WER	alternative metric that tackles the problem
no true percentage	altWER, MER, WIP, precision/recall/F-measure
deletion-insertion asymmetric	WIP, precision/recall/F-measure
insensitive to semantics	precision/recall/F-measure

Table 2.2 – Overview of WER's drawbacks and alternative evaluation metrics.

defined as:

$$WIP = \frac{H^2}{(H+S+D)(H+S+I)}$$

Since $(H+S+D)$ equals the length of the reference (N_{REF}) and $(H+S+I)$ equals the length of the ASR output (N_{ASR}) we can also define WIP as:

$$WIP = \frac{H^2}{N_{REF}N_{ASR}}$$

The ASR metrics that have been introduced so far are all intrinsic evaluation methods, since they are task independent. WER also does not allow a performance analysis with respect to the semantic importance of the words that have been wrongly recognized. One can argue that there are words in a conversation that are more important to be recognized correctly than others, especially if a specific task has to be achieved. For the evaluation of the spoken retrieval recognition track in the Seventh Text Retrieval Conference (TREC-7) which evaluated system combinations of automatic speech recognizers and retrieval systems, three alternative metrics have been proposed which take into account the information-carrying words that are relevant to the retrieval task [Garofolo et al., 1998]. The first metric is based on the recognition rate of named entities. The second uses information retrieval approaches like stop-word filtering and stemming to determine the words that are considered in the evaluation [Garofolo et al., 1998]. The third metric calculates WER only for words that were in the test topic [Garofolo et al., 1998].

Another approach that takes the semantic content of the recognized words into consideration takes up the concepts of recall, precision and F-measure from the information retrieval domain [McCowan et al., 2005]. *Recall* is the fraction of relevant information units that have been retrieved, *precision* is the fraction of retrieved information units which are relevant and *F-measure* is the most widely used combination of precision and recall [Baeza-Yates and Ribeiro-Neto, 2010, p. 75]. An overview of the drawback's of WER and which alternative metrics tackle these problems can be found in Table 2.2.

An attempt to evaluate the performance of ASR systems in the context of a special task can be found in Halverson et al. [Halverson et al., 1999]. The authors compare three commercially

available ASR systems for dictation with a focus on error correction strategies. A similar aspect is examined in an experiment on human-human communication which explores error recovery strategies after miscommunication in order to figure out the implications for spoken dialogue systems [Skantze, 2005]. In the experiment a speech recognizer is used to corrupt the speech in one direction. In another experiment the readability of machine transcripts is compared to human transcripts by asking the subjects to answer comprehension questions [Jones et al., 2005a,b].

In *spoken language translation* (SLT) [Waibel and Fügen, 2008], in which automatic speech recognition is combined with machine translation, new approaches to evaluation are also emerging as part of competitions such as in the *International Workshop on Spoken Language Translation*. Especially the sub-domain of medical SLT recently saw several efforts to take the task at hand into account for the evaluation of the developed systems [Rayner et al., 2008b]. The *Converser* system, a bidirectional spoken language communication system for the healthcare area [Dillinger and Seligman, 2006] for instance has been evaluated with a strong focus on usability aspects of the system, multimodal input and output [Seligman and Dillinger, 2006], and possible interactive monitoring and correction methods for ASR and MT [Seligman and Dillinger, 2011]. However, results of the ASR performance are not reported independently of the machine translation output. Another system that was extensively studied is the medical spoken language translator *MedSLT* [Rayner et al., 2008a]. An extensive study has been carried out with the system at the *Dallas Children's hospital* where a set of automatic metrics was applied together with a human metric and the results of these assessments were compared [Starlander and Estrella, 2011]. The system was also used to compare the performance levels of novice users with expert users in a simulated situation [Chatzichrisafis et al., 2006]. The evaluation focused on the system as a whole and did not evaluate the subcomponents separately, therefore no intrinsic measures of the ASR performance were reported.

2.5 Machine Translation

Machine translation (MT) is the use of computers to automate translation from a source language into a target language [Jurafsky and Martin, 2009, p. 895]. Translation is different from the other two language processing tasks that are under consideration in this thesis, since it is a language profession in which even humans need special training unlike language understanding and production which is usually acquired naturally while growing up. Some aspects of natural languages seem to be universal to all languages, for example every spoken language seems to have vowels and consonants [Christiansen et al., 2009]. However, there can be big differences between languages concerning for example the character set, word order or different degrees

of morphology (for an overview please refer to [Jurafsky and Martin, 2009, p. 898 ff.]). These differences are the reason why learning a foreign language is difficult for humans and also why automatic translation can be challenging.

Machine Translation has a tradition of nearly sixty years and the history has often been told. The following overview of the history of MT research is based on Koehn's account in his textbook on statistical machine translation [Koehn, 2010]. Translating languages is similar to breaking a code. Therefore it can be stated that the research of machine translation has its roots in the attempt to crack the German Enigma code in World War II. A further factor influencing the rise of this research area was the advent of electronic computing around the same time. At first, major funding went into the research of automatic translation. The great optimism in the mid fifties that the problem of MT will be solved soon was triggered by the success of *IBM*'s *Georgetown Project* [Hutchins, 2004]. This demonstrated the rule-based translation from Russian into English for about sixty sentences. However, the *ALPAC report*[4] in the mid sixties stopped almost all major funding. The report assessed that no advantage can be gained from machine translation because it was cheaper to provide full human translation than to post-edit MT output. Especially, since there was no shortage of human translators. However, foundations for commercial translations were laid in the decade after the ALPAC report. One of them is the founding of one of the most successful translation companies, *Systran*, in 1968, which is still in business today providing technology for *Yahoo! Babel Fish* and until 2007 for *Google*'s language tools. The *Meteo* research project at the University of Montreal developed a fully functioning translation system that was able to translate weather reports from one language into another [Thouin, 1982] and has been operating since 1976. In the eighties the focus of machine translation research shifted to interlingua-based approaches, where meaning is represented in a formal way independent of a specific language. The *CATALYST project* at the *Carnegie Mellon University* (CMU) and the German *Verbmobil* project [Bub and Schwinn, 1999] are only two examples of research ambitions to develop interlingua systems. Another idea that was introduced in this decade was to learn how to translate from past translation examples and especially in Japan the first *example-based translation* systems came up. The *IBM Research Labs* also steered in the direction of data driven approaches to machine translation with their *Candide* project [Berger et al., 1994] which modelled the translation task as a statistical optimisation problem. However, it was not until the beginning of the new century that the idea of statistical machine translation gained momentum with large research projects like *TIDES* (Translingual Information Detection Extraction, Summarization), and *GALE* (Global Autonomous Language Exploitation) which were funded by the *Defense Advanced Research*

[4] An electronic version of the ALPAC report can be accessed online through the *National Academic Press* via `http://www.nap.edu/openbook.php?record_id=9547` (last accessed 01/05/2012).

DICTIONARY

reading	...	Lesen, lesend, Messwert, Lektüre, Lesung, Messung
a	...	ein, eine, a
book	...	Buch, bestellen, reservieren, buchen, ausbuchen
is	...	is
fun	...	Spaß, Vergnügen, Scherz

Figure 2.6 – Example dictionary.

Projects Agency (DARPA).

In the early nineties the pure research activities in MT changed into practical applications such as computer aided translation systems for human translators. The quality of MT output has improved significantly in the last two decades, due to the introduction of data-driven approaches and reached a level where it is beginning to be more commonly employed in real-world applications. A significant trend in this direction in the last five years was the use of MT in combination with speech technologies. For example Paul et al. [Paul et al., 2008] report on two multilingual translation services for mobile phones. Yamashita and Ishida [Yamashita and Ishida, 2006] investigate the way in which MT affects referential communication between speakers of different native languages. Shimizu et al. [Shimizu et al., 2008] describe a handheld speech-to-speech translation system. The main objective, *fully-automatic high-quality machine translation*, has only been achieved for limited domains such as the translation of weather forecasts. More widespread is the use of machine translation to get a gist of what the content of a document is. In order to get the meaning across translation does not have to be perfect. In post-editing machine translation is used to increase the efficiency of human translators. Here a rough machine translation of a text is obtained and then adjusted by a professional human translator who corrects mistakes and assures fluency.

2.5.1 Approaches to Machine Translation

Machine translation research broadly falls into three approaches, namely rule based MT, statistical MT and example based MT. This section illustrates the basic ideas behind these three approaches by way of examples. Consider the English input sentence "reading a book is fun" which has to be translated into German. The most intuitive way to approach the task of translation is to use a dictionary, look up every translation and write them down, word by word. An example dictionary for the example sentence can be found in Figure 2.6. This very basic method is implemented in the simplest variant of rule based MT called *dictionary-based machine translation.*

Figure 2.7 – Example sentence with word by word translation and adjusted correct translation.

This approach might be suitable for the translation of long lists of phrases such as inventories or catalogues of products but translation on a sentence level is bound to fail for several reasons. The first problem with this simple approach arises when the dictionary proposes more than one word for the translation, like for the word *'book'* in our example which can be the verb *'buchen'* as in booking, or the noun *'Buch'* which is a printed literary work, amongst others. Many words have more than one meaning (*lexical ambiguity*) and the correct meaning is determined by the context. The second shortcoming of this simple approach is that languages may have different word orders, as is the case for English, with a typical sentence structure where the subject precedes the verb (V) which in turn is followed by an object (S-V-O), and German, where the sentence structure is different. Here the subject usually precedes the object which in turn precedes the verb (S-O-V). In Figure 2.7 the word-by-word translation of the example can be found and then an adjusted version which corrects the different sentence structures. The example shows not only that words in the translation have to be reordered, but also that there are words that need to be inserted in order to obtain grammatically correct sentences.

The more sophisticated rule-based approach which is called *transfer-based machine translation*, tries to straighten out these problems with the help of rules which find an alternative in all cases were the simple word-by-word translation is not enough. Consider the lexical ambiguity problem for example where the word book can be translated in many different ways. In this case the rule-based approach uses linguistic knowledge and the context of the word that needs to be translated in order to find out which sense of 'book' is appropriate for the translation. In the example case the linguistic knowledge that the 'book' must be a noun rather than a verb since the determiner 'a' precedes it, is used to disambiguate the meaning of the word.

There are many different ways in which transfer-based machine translation systems can be implemented. The common pattern is that in a first stage the morphology and syntax of the

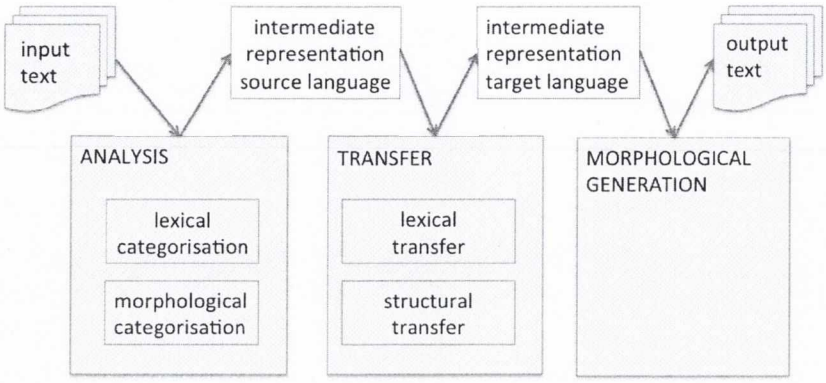


Figure 2.8 – Schematic representation of the transfer-based machine translation approach.

input text is analysed and a syntactic abstract representation is created. In the transfer step this is converted into an abstract syntactic representation in the output language. These steps are illustrated in Figure 2.8. The analysis stage consists of two steps, the morphological analysis in which the part-of-speech tags (e.g. noun, verb, adverb) and sub-categories (such as number, gender, tense) are determined and the lexical categorisation in which the right meaning of the word in its context is determined based on the part of speech tags. The transfer step also involves two steps: one is the lexical transfer, in which every word is mapped to its translation with the correct meaning in the given context, and the structural transfer in which then the larger constituents are considered and word order and gender agreement are adjusted. In the morphological generation stage the output of the structural transfer is basically a mapping of the abstract representation that has been generated to the appropriate word in the output language.

Interlingua machine translation is similar to the transfer-based approach. The difference is that in the interlingua-based approach the intermediate representation, the so-called interlingua, is independent of the source and target language. In the interlingua approach the source language is transformed to the interlingua and from there the target text can be retrieved. The advantage of this approach is that fewer components are required to translate between all word pairs. Since there is a universal interlingua developers are only concerned with the transformation from a language to the interlingua and back from the interlingua to a language, independent of the language pair. In other MT approaches, for instance the transfer-based approach, translation components need to be developed for each language pair, resulting in $n(n-1)$ components for n different languages. In an interlingua-based approach on the other hand only $2n$ components need to be developed for n languages (see Figure 2.9). Consequently, if another language would be added in the interlingua approach only two new components need to be developed,

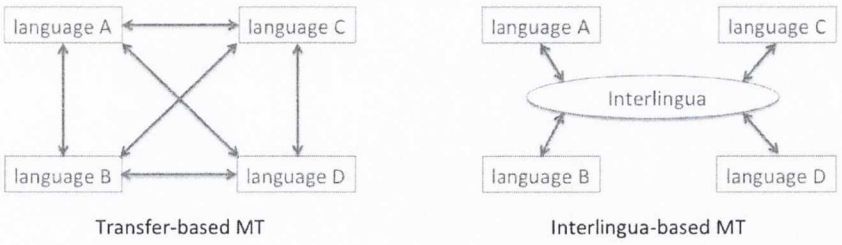


Figure 2.9 – Comparison between the number of components (arrows) that are needed in usual MT approaches (e.g. transfer-based MT) and the interlingua approach.

whereas in the transfer-based approach two components for each already existing language would be necessary. In the example in Figure 2.9 this would result in eight new components.

The key problem with the interlingua approach is that it is almost impossible to define an adequate interlingua for bigger domains. The interlingua needs to be abstract and independent of the source and target languages and it must be able to represent all characteristics of all languages. Therefore it is ideally suited in situations where multilingual translations in a restricted domain are required.

The problem with rule-based approaches remains however that language is so complex and rich and evolves over time so that it is a challenging, arduous and time-consuming task to try to fully analyse languages manually. This is where *statistical MT* comes into play. The idea behind statistical machine translation is to learn how to translate from past translation examples. The lexical ambiguity problem for example is solved on the basis of word counts in large parallel corpora. A parallel corpus is a collection of texts in one language with corresponding translations in another language. The ambiguous word is looked up in the corpus and the most likely translation is calculated statistically. In the case of the example in Figure 2.7 this means to look up the word 'book' in a large text collection of English texts and their German translations and count how often the different translations (e.g. 'Buch' or 'buchen') occur. On the basis of these counts the *lexical translation probability distribution* is then estimated using the function:

$$p_{book} : \mathrm{e} \rightarrow p_{book}(\mathrm{e})$$

where e is each possible German translation for book:

$$\mathrm{e} \in \{\text{Buch, bestellen, reservieren, buchen, ausbuchen}\}$$

Together with an alignment function that maps each input word to an output word these probabilities define a model that generates a number of different translations for each sentence with different probabilities. This very simple model is called *IBM's Model 1*. Since this model is weak in cases where reordering and dropping or adding of words is needed, four models with more sophisticated approaches have been proposed, called IBM Models 2 through 5 [Vogel et al., 1996]. The IBM models are word based but the introduction of phrase based models [Koehn et al., 2003], where sequences of words with different lengths are translated, resulted in significant improvements in translation quality. A further improvement in methodology in statistical machine translation is the introduction of syntax-based translation, where the phrases that are translated are syntactic units in order to be able to use all necessary syntactic information in the translation process [Charniak et al., 2003]. For a detailed account of the method of statistical machine translation and the mathematical background please refer to [Koehn, 2010].

Challenges in statistical machine translation that remain are the alignment of sentences in parallel corpora, idioms, different word order and out of vocabulary words which are words or phrases that cannot be found in the training data and can therefore not be translated.

The third approach to machine translation, *example based machine translation* (EBMT), is related to phrase based machine translation. In recent systems the distinctions between the two methods are blurred [Koehn, 2010, p. 152]. EBMT is also based on the use of parallel corpora and uses the idea of translation by analogy. The idea of example based machine translation was first introduced in 1984 by Makoto Nagao [Nagao, 1984]. EBMT systems look up similar sentences to the input sentence and then make the necessary adjustments to the target sentence. In the case of the introduced example imagine the sentence *'Reading a book is fun.'* and its translation *'Ein Buch zu lesen ist ein Vergnügen'* can be found in parallel corpus. Consider a new input sentence *'Reading Shakespeare is fun'*. In EBMT the sentences are segmented in their substantial units and the translation of these units have to be learned. In the example the EBMT system would need to learn the following three units:

- *'**X** zu lesen ist ein Vergnügen.'* is the translation of *'Reading **X** is fun.'*
- *'Ein Buch'* is the translation of *'a book'*.
- *'Shakespeare'* is the translation of *'Shakespeare'*.

By recomposing these units the translation of the new input sentence can be determined to be *'Shakespeare zu lesen ist ein Vergnügen'*. Similar to this is the translation of new sentences by the recombination of units of sentences that have been translated already. For example if the system has learned the translation of the two sentences *'At the age of 18, Shakespeare married Anne Hathaway, with whom he had three children.'* and *'James Joyce published several books*

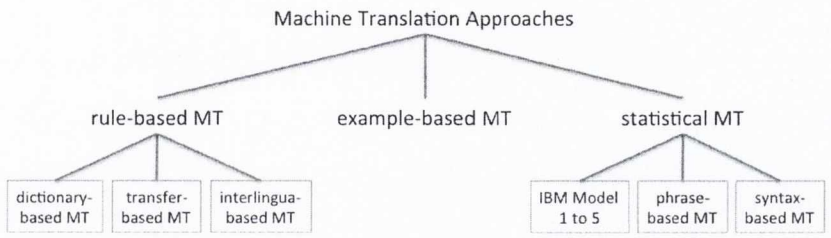


Figure 2.10 – Synopsis of approaches to machine translation.

with George Roberts.' the system can combine these into the translation of the new input sentence *'At the age of 18, James Joyce published several books.'*

A synoptic overview of the machine translation approaches discussed can be found in Figure 2.10.

2.5.2 Evaluation of MT Systems

For MT components to be effectively deployed, it is important for developers to be able to reliably assess the quality of the translation output. In order to judge the performance of a translation system the quality of the translated output text is usually assessed. This task is very difficult for several reasons [Flanagan, 1994]. Often the simple count of wrong words is not meaningful. First of all, there is no single 'correct' translation. A text can have several acceptable translations and a quality rating in this case is rather a matter of taste than of what is right or wrong. Furthermore, the boundaries of errors are hard to determine. Repeatedly errors involve whole phrases or even discontinuous expressions. Moreover, one error can lead to another. Finally, the cause for an error in the MT output is not always apparent which hinders the discovery of what went wrong in the translation.

In the early days of machine translation many evaluation attempts were based on human judgements (adequacy, fluency and fidelity) [Hovy, 1999], due to the difficulty of the task to assess the quality of a machine translation automatically. These were replaced by automatic procedures for the obvious advantages that they are quicker, cheaper, repeatable and objective. The aim of these automatic metrics is to correlate with human judgements as much as possible and therefore they are based on comparing the translation output to human reference translations. Unlike other metrics that are reference based (e.g. WER in automatic speech recognition) MT metrics cannot rely on a single reference because translations can vary in terms of word choice and the ordering of phrases.

Analogous to the word error rate in speech recognition is the *translation error rate*[5] (TER) [Snover et al., 2006] for machine translation which has been introduced by the Global Autonomous Language Exploitation (GALE) research program. TER calculates the minimum number of edits that are needed to change a hypothesis so that it exactly matches the closest reference from a set of references. This number is normalised by the average length of the references. Edits can be insertions, deletions, substitutions and shifts and are summed up only for the closest reference. TER is defined as:

$$TER = \frac{\text{number of edits for closest reference}}{\text{average length of references}}$$

For the human-in-the-loop alternative *HTER* [Snover et al., 2006] human annotators produce the closest reference. The TER score is then automatically computed on basis of this human corrected reference.

The best known and most widely applied automatic evaluation metric is *BLEU* (BiLingual Evaluation Understudy) [Papineni et al., 2002]. To determine the BLEU score of a translation the number of N-gram matches, of varying length, between the system output and a set of reference translations is calculated. Typically BLEU scores are calculated for sentences. The score can range from 0 to 1, a higher score indicates a closer match. A score of 1 is assigned if the translation exactly matches one of the reference translations. Criticism on BLEU has been raised because it is relatively unintuitive and, in order to correlate with human judgments, it relies on a large number of references and sentences [Snover et al., 2006]. In 2003 in an experiment the performance of N-gram based metrics like BLEU on professional human translations was compared to machine translations and it was shown that some machine translations outscore the professional human translations [Culy and Riehemann, 2003]. Callison-Burch et al. also showed that BLEU might not correlate with human judgements. They proved that an improved BLEU score does not necessarily indicate an improvement in the translation quality and vice versa. Systems with an improved translation quality did not receive a higher BLEU score [Callison-Burch et al., 2006]. Since BLEU is of high interest to this thesis it will be revisited in more detail in Chapter 3.

The *NIST* metric [Doddington, 2002] is a variant of BLEU that has been introduced in 2002 by the *National Institute of Standards and Technology* (NIST). Unlike BLEU, the NIST metric calculates how informative a particular N-gram is. More weight is given to rarer correct N-grams, for example the N-gram "that is" is more likely than "correlate with" and will therefore achieve a lower score. The brevity penalty which is introduced in BLEU in order to compensate

[5] Snover et al. point out that the name translation *error* rate is regrettable for its implication that it is a definitive MT measure and therefore renamed it to translation edit rate.

for too short translation segments (see Chapter 3), is also altered accounting for the fact that the impact of small variations which have a higher influence in the BLEU score.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics which was inspired by BLEU and was developed to evaluate automatic summarization [Lin, 2004]. It can also be used for MT evaluation. Like BLEU, ROUGE compares automatically produced summaries or translations against a set of human produced references. There are five different versions of ROUGE implementing different approaches. ROUGE-N for example is N-gram based, as is ROUGE-S and ROUGE-SU, with the difference that the latter two use skip-bigrams and a combination of skip-bigrams and unigram based co-occurence statistics respectively [Lin and Och, 2004]. ROUGE-L and ROUGE-W are based on the least common substrings.

N-gram based evaluation metrics like BLEU and NIST enabled researchers to validate and optimise translation methods quickly. However, with the development of better MT systems they are not discriminative enough anymore to reflect the translation quality of today's systems [He and Way, 2009], therefore several different approaches have been explored.

An alternative automatic metric to BLEU and NIST is *METEOR* [Banerjee and Lavie, 2005] which tries to improve on their matching schemes. The METEOR score counts the number of exact word matches between the system output and reference. In a second step, words that cannot be matched are stemmed and matched again. If reordering of words is necessary to pair the translation output with the reference, penalties are incurred and reflected in the score. Furthermore, METEOR results are refined by looking up synonyms from *Wordnet*. A similar approach is taken up by *ParaEval* [Zhou et al., 2006]. Like ROUGE, ParaEval is also employed to evaluate summaries and is a paraphrase-based method which employs external data resources in order to incorporate paraphrases into the evaluation.

Apart from these attempts to improve on N-gram based metrics there are two main approaches to enhance MT evaluation, syntactic methods and machine learning methods. Syntactic metrics incorporate syntactic information such as constituent labels, or dependencies into the evaluation process (e.g. [Liu and Gildea, 2005], [Giménez and Màrquez, 2008]). Machine learning approaches aim at learning human judgements, either classification based [Corston-Oliver et al., 2001], regression based [Albrecht, 2007] or ranking based [Ye et al., 2007].

All the methods discussed so far are intrinsic methods. They are 'decontextualized' in the sense that they do not take into account the task being supported by MT, and therefore might not produce meaningful results in real-life situations [Karamanis et al., 2010]. In such applications time constraints, distractions, accommodation to certain types of mistakes by the user, awareness of translation errors in spoken language as opposed to text, and other issues that arise in interactive situations might influence the perception of MT output.

Most extrinsic evaluation efforts which assess the effect of task and context-dependent variables on the user performance of MT systems date back to the late 90s, when two new MT research trends emerged [Laoudi et al., 2006]. One was task-based experiments conducted by developers of MT systems. Levin et al. [2000], for example designed a task based evaluation method for the JANUS speech-to-speech MT system. They compared the results of this evaluation with their accuracy-based evaluations. Phillip Resnik [1997] proposed a method to evaluate multilingual gisting based on its role in decision support. The second evaluation stream was task-based experiments assuming an ordering of task difficulty for text-handling tasks such as proposed by Taylor and White [1998] and White et al. [2000]. In 2006, the idea of task-based evaluation was picked up again by a research group around Laoudi and Voss who conducted experiments on the text-handling task of extracting information from MT output [Laoudi et al., 2006; Voss and Tate, 2006].

Several researchers also tried to target the problem that not all MT errors are created equally and depending on the task at hand certain errors might have a bigger impact than others. For example Rayner et al. argue that the mistranslation of a greeting might not have strong clinical consequences whereas the mistranslation of a negation might [Rayner et al., 2008b]. In order to account for the different implications that MT errors might have, different categorisation schemes have been proposed in the literature. These categories are associated with scores based on how fatal the error is for the task in order to determine the performance score. Bederson et al. for example describe a new iterative translation process that combines MT and the collaboration of monolingual speakers in order to establish translations [Bederson et al., 2010]. They introduce a new classification scheme for MT errors which distinguishes three types of MT errors: (1) errors that are detectable and correctable (due to linguistic redundancy and shared context), (2) errors that are detectable but not correctable and (3) undetectable errors.

A linguistic approach to the categorisation of errors is proposed for the evaluation of *Compuserve*MT systems [Flanagan, 1994]. Based on an evaluation of the most frequent error types in English-to-French translation (e.g. wrong spelling, incorrect accent, wrong verb inflection) that have been observed, 18 categories are proposed. Notably, wrong translation of negation is one of the categories proposed. For each language pair a different error classification is necessary depending on the special features of the languages involved. Flanagan argues that the error categories can be ranked in order to user priorities which equates to assigning a score to the categories, but leaves the details open to the reader. One downside of the proposed categorisation is that multiple or linked errors pose problems for the category assignment.

Especially in end-to-end evaluation of medical SLT systems the approach to categorise MT errors has been taken up, due to the safety critical nature of the application area. In their

proposal of a shared task for medical SLT Rayner et al. [2008b] describe a categorisation of translation errors according to the clinical consequences that arise when the system fails (i.e. does not produce a translation or does not produce a correct translation). The seven categories are as follows: (1) perfect translation which has an implication on the clinical consequences, (2) perfect translation which has no implication on the clinical consequences (e.g. greetings), (3) imperfect translation which is not dangerous in terms of clinical consequences, (4) imperfect translation which is potentially dangerous, (5) nonsense, (6) no translation produced, but later rephrased in a way the system handled adequately, and (7) no translation produced and not rephrased in a way the system handled adequately. Since medical SLT is a safety critical area the authors argue that potentially dangerous mistranslation has to be associated with a negative score that is large compared to the positive score for a useful correct translation and that the scale has to be normalised so that failure to produce a translation is counted as zero.

Another categorisation targeted at medical SLT systems is introduced by Starlander and Estrella [2011]. In order to prioritise precision over recall their proposed classification focuses on the meaning of a sentence. It is based on whether the message came across instead of syntactic or linguistic aspects. The four proposed categories are: (1) completely correct: the entire meaning of the source is present in the target sentence, (2) translation is not completely correct: meaning is slightly different but it represents no danger of miscommunication (between doctor and patient), (3) translation does not make any sense: translation is not correct in the target language, it is gibberish, and (4) translation is incorrect and the meaning in the target and source are very different: a false sense, dangerous for communication (between doctor and patient) is transmitted.

There are a number of noteworthy evaluation campaigns for machine translation. The oldest and most well known is the *NIST* evaluation. The translation between European languages is in the focus of the Workshop on Statistical Machine Translation and the International Workshop on Spoken Language Translation (*IWSLT*) is mainly concerned with speech translation.

2.6 Text-to-Speech Systems

The core task of *text-to-speech (TTS) synthesis*, often also called *speech synthesis*, is to take a sequence of written text and produce an acoustic waveform (speech) as output [Jurafsky and Martin, 2009, p. 283]. The field of speech production is one of the oldest in the history of natural language processing and dates back to the 18th century. In 1791 Wolfgang von Kempelen build the first speaking machine, a model of the human vocal tract which was able to produce vowels and consonants [von Kempelen, 1970]. In the 1950s the first computer based speech synthesisers emerged. Technologies that rank among this first generation of modern speech

synthesisers are articulatory synthesis, formant synthesis and classical linear prediction. Most modern synthesisers are based on concatenative synthesis. The foundations of this technique were laid out in the early fifties, when Cyril M. Harris proposed to recombine recorded phones into new words [Harris, 1953]. He demonstrated this technique with recordings on magnetic tape. However, the first concatenative approaches were based on simple phones and the idea of unit selection including the use of diphones which surfaced at the end of the 1950s was only implemented in systems in the 1970s [Dixon and Maxey, 1968]. With the invention of the unit selection technique at ATR in Japan in the 1980s high quality, natural sounding and intelligible synthetic speech for virtually any task was accessible [Hunt and Black, 1996]. In the nineties the *Hidden Markov Model (HMM)* synthesis paradigm emerged which can be called the third generation of text-to-speech synthesis. Today speech synthesis is used in a wide range of applications such as devices that read out text to visually impaired people, telephone based conversational agents, or devices that lend their voice to people who have lost their own voices. It is prioritised in applications where information is better distributed via an acoustic signal, such as in-car navigation systems. It is also used in the entertainment industry, for example in computer games. However, still a high number of commercial applications that use speech output today do not use TTS, but rather make use of a set of pre-recorded prompts [Dybkjaer and Bernsen, 2000] because this offers perfectly natural results and a low-tech solution [Taylor, 2009]. A limitation of pre-recorded prompts is the fixed set of utterances. In order to extend or revise the set of utterances the same speaker is required to record the new prompts. TTS systems offer the most flexible way of generating output speech, but they are still facing difficulties with the pronunciation of proper names or with phrases borrowed from other languages [Dybkjaer and Bernsen, 2000].

Apart from the challenge of how to get the prosody right recent research proposed to synthesise filled pauses and disfluencies into fluent synthesised sentences in order to enhance the naturalness of synthesised speech [Adell et al., 2012]. The authors justify this approach with the explanation that instead of simulating the way people read it is necessary to simulate the way people speak since the application of future TTS systems will vary and reading style speech will not be appropriate in all cases.

In the *MIT technology review*[6] it has further been reported that a prototype was presented at a demonstration by *Microsoft* where the TTS system is trained on the sound of the users voice in order to use it on translated texts. This research is especially useful for the development of speech-to-speech translators.

[6] `http://www.technologyreview.com/news/427184/software-translates-your-voice-into-another/` published 09/03/2012 author Tom Simonite (last accessed 09/05/2012)

2.6.1 Approaches to Text-to-Speech Synthesis

All approaches to text to speech synthesis basically contain two steps. As can be seen in Figure 2.11, in the first step, the *text analysis*, the input text is converted into an internal phonemic representation. This representation gets then converted into a waveform in a second step called *waveform synthesis*. The mapping between text and phonemes in the text analysis is a standard procedure that consists of three steps. In a first step the text input is normalized which includes sentence tokenisation, dealing with non-standard words (e.g. numbers and acronyms) and homograph disambiguation (e.g. read in present tense and past tense). In the next stage a pronunciation for each word is generated. This step is called the phonetic analysis. For standard words a simple dictionary look-up in a pronunciation dictionary (e.g. *CMU Pronunciation Dictionary*[7]) is attempted, but more sophisticated methods (e.g. see [Fackrell, 2004]) have to be implemented in cases where names or other unknown words are encountered. In cases where the pronunciation cannot be looked up from a pronunciation dictionary grapheme-to-phoneme conversion algorithms are applied. These can either be rule-based or rely on machine learning methods. The third and last step in the text analysis stage is the prosodic analysis, in which the text is analysed with a focus on the prosodic structure, prosodic prominence and tune and these suprasegmental features are added to the internal phonemic representation.

The phonemic internal representation which is a list of phones with a specified duration and F0 target,[8] is the basis for the second step in text to speech synthesis, the waveform synthesis. There are several different paradigms for waveform synthesis which will be introduced in the following section.

Before the late 1980s three main techniques dominated the field of speech synthesis which were based on the model of the vocal tract. These *first generation techniques* were articulatory synthesis, formant synthesis, and classical linear-prediction synthesis.

Articulatory synthesis is the most obvious way to synthesise speech by modelling a human vocal tract directly, as was done by von Kempelen in the 18th century. Modern articulatory synthesisers digitally simulate the flow of air through a representation of the vocal tract and are based on the acoustic tube model [Taylor, 2009, chapter 11] which approximates a complex tube shape by a number of smaller uniform tubes. Since it is difficult in practice to determine the necessary rules and models for the articulatory model there is only little engineering work in articulatory synthesis. *Gnuspeech* is a software package under the GNU General public license which implements articulatory speech synthesis based on the NEXT system which was originally developed at the *University of Calgary*.

[7] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict` (last accessed 09/05/2012)

[8] The fundamental frequency, often abbreviated with F0, is the frequency of the vibration of the vocal fold [Jurafsky and Martin, 2009, p. 267].

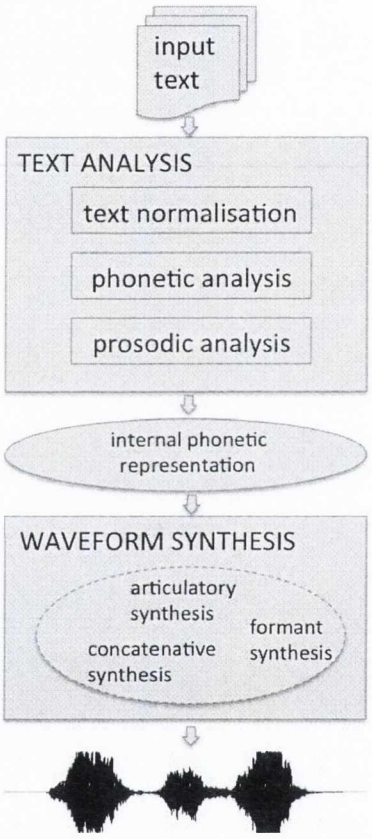


Figure 2.11 – Two steps to text-to-speech synthesis (based on the hourglass metaphor by Paul Taylor [Taylor, 2009]).

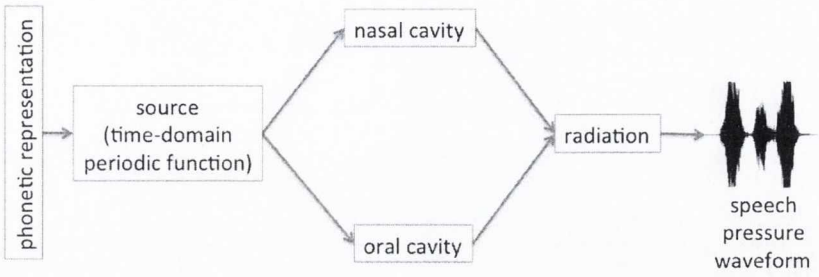


Figure 2.12 – Schematic representation of the formant synthesis model.

Formant synthesis was the most dominant synthesis technique until the 1980s. Alternatively it is also called synthesis by rule in order to distinguish the technique from other methods that reconstruct waveforms which were common at the time when formant synthesis surfaced. However, there are also data-driven approaches to formant synthesis. The technique is also based on the tube model but it is common that in formant synthesis the nasal cavities and the oral cavities are modelled separately and the output of both components are later combined in a radiation component which simulates the articulatory modulation of nose and lips. A schematic representation of the formant synthesis approach can be found in Figure 2.12. Formant synthesisers typically produce intelligible speech which is not natural sounding. The implementation of the rules for formant synthesis and the setting of the correct parameters cannot be done automatically. The Klatt synthesiser is probably the most well known example of a formant synthesiser [Klatt, 1980].

Classical linear prediction is very closely related to formant synthesis. It is mentioned here just for the sake of completeness. For more details on the method itself please refer to the textbook on text-to-speech synthesis by Paul Taylor [Taylor, 2009].

In the late eighties a new generation of TTS systems emerged, and again their rise was made possible by new technological developments. First generation synthesis systems were able to generate intelligible but not natural sounding speech. But when memory became cheaper it became feasible to store vast amounts of waveforms with higher sampling rates. Rather than using wave-forms to acquire model parameters, techniques were developed where high quality speech was recorded and new word sequences synthesised by recombining pre-recorded phones. Most modern synthesiser are based on the *synthesis by concatenation* technique. The idea behind concatenation is to record a human voice, then cut the recordings into small pieces and store these pieces in a database together with their phonemic representation. Once the in-

ternal phonemic representation for a text that needs to be synthesised is determined, the correct sequence of recorded snippets is retrieved from the database and spliced together. Concatenative synthesis produces the most natural sounding speech, since it is based on the human voice. For the same reason, however, it is limited to a single voice.

There are two main approaches to concatenative speech synthesis: one is unit selection the other is diphone synthesis. In the *unit selection approach* [Hunt and Black, 1996] a large database of recordings is dissected into smaller units of different length (i.e. phones, diphones, syllables, morphemes, words, phrases and sentences). These units are stored in a database together with their acoustic parameters (e.g. duration, neighbouring phones). The desired text is synthesised by selecting the best sequence of candidate units, hence the name unit selection, retrieving them from the database, and splicing them together.

Phones are the basic units of speech sounds. Diphones are units like phones but they start in the middle of one phone and end at the middle of the following phone. Hence, they are rather the transition sounds from one phone to the next phone. Different languages have different numbers of phones. Combinatorial, the number of diphones is roughly the square of the number of phones. However, as some phone-phone pairs are not used, the number of diphones is usually less than the square of the number of phones (e.g. approximately 2500 in German). In *diphone synthesis* the recordings are dissected into diphones, with the advantage that the necessary speech database is small in comparison to the one employed in the unit-selection approach. Diphones are favoured over phones for their ability to simulate coarticulation which denotes the fact that the sound of a phone is dependent on the preceding and succeeding phone, and can therefore differ slightly depending on its context. The simple splicing of diphones, however, is not sufficient. Signal processing methods have to be employed to produce the desired prosody. The most widely used are the *pitch-synchronous overlap and add (PSOLA)*, where individual pitch periods are isolated from the original speech and modified in order to resynthesise the final wave-form [Moulines and Charpentier, 1990]. An alternative to PSOLA is MBROLA [Dutoit et al., 1996]. Diphone synthesis, however, suffers from the fact that there are more global effects in phonetics that go beyond the unit of diphones. This is why unit selection synthesis creates more natural sounding speech.

First generation speech synthesis techniques basically try to simulate the human vocal tract. In order to do so, these methods attempt to guess a set of parameters for a given synthesis specification. To determine these parameters is the limiting factor of these techniques [Taylor, 2009, p. 435]. The speech that first generation synthesis systems produce is intelligible but often not natural sounding. Second generation synthesis approaches circumvent the parameter problem by measuring the values from speech samples directly. However, there will never be enough

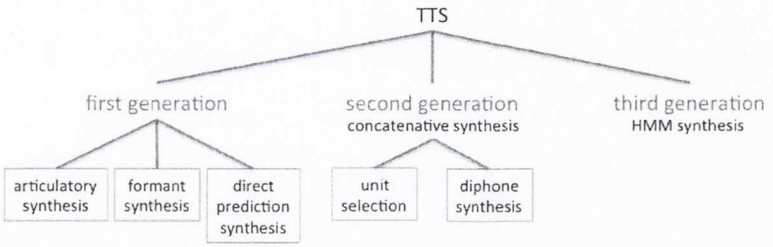


Figure 2.13 – Synopsis of approaches to text-to-speech synthesis.

data to cover all effects, and synthesis is basically reordering the original material, therefore restricted to recreate what has been recorded. The idea behind *Hidden-Markov synthesis* is to learn the mapping between specification and parameters using statistical machine learning methods. This idea has the benefit that the model can be adjusted so that the original voice can be transformed into a different voice. Furthermore, only the parameters of the model need to be stored, hence the technique requires less memory. Hypothetically other machine learning approaches can be used for speech synthesis but most work has concentrated on hidden Markov Models (HMMs) [Rabiner and Juang, 1986]. A detailed description on HMM synthesis can be found in Paul Taylor's book on text-to-speech synthesis [Taylor, 2009, chapter 15]. A synoptic overview of the TTS approaches discussed in this section is presented in Figure 2.13.

2.6.2 Evaluation of TTS systems

The assessment of TTS systems and the development of standardised evaluation methods has seen much attention in the last two decades with European projects like *Sam*, *Eagles* and *Eurococosda* concentrating on the topic. The evaluation of TTS systems is fundamentally different to the assessment of other speech technologies, and these differences set limits to what can be expected of TTS technology [van Santen, 1997]. One problem that research in TTS faces is that there are not any universally agreed evaluation criteria for TTS systems. The quality of a TTS system is an abstract measure of the naturalness, fluency or clarity of the speech [Jurafsky and Martin, 2009, p. 314]. When it comes to evaluation of language technology researchers are always faced with the problem that natural language is a complex systems in which a wide range of linguistic variation has to be considered in tests. This is especially true when evaluating systems that emulate the voice, because rating a human voice alone is a subjective matter and this is only one factor that influences the perception of a TTS system. Most features of spoken text such as voice or prosody are subjective in their nature and might receive varying ratings from

different judges. Furthermore, it is nearly impossible to draw the line separating one feature from another. All aspects of the system under inspection have to be taken into consideration but since TTS systems consist of multiple components such as grapheme-to-phoneme conversion, accentuation, and phoneme intelligibility for example, it is not always clear which component is responsible for a given problem. The evaluation of TTS systems is still dependent on human judges, because so far no automatic evaluation method has been developed to judge TTS performance. In order to assess the intelligibility of concatenative speech synthesisers for example it would be necessary to find texts that use every possible concatenative unit, which is very difficult. Even worse is the problem to find texts that contain all possible neighbouring pairs of units. And even if it were possible to establish a complete set of utterances covering all possible neighbouring pairs of units, TTS evaluation would remain dependent on human judges who can only process a few hundred utterances. Furthermore, listeners become familiar with synthetic speech and therefore the learning effect influences the results for one listener over time. When it comes to the evaluation of prosody, speaker style and emotional characteristics there is a genuine lack of proper tests [Pols, 1997]. Moreover, a factor that should not be underestimated in TTS evaluation is that the quality depends on the technical equipment that is used to play back the synthesis. For these reasons Jan van Santen goes as far as stating that TTS is not evaluation driven as is speech technology in general but rather has been driven by a mixture of the progress made in underlying sciences, enabling technologies and priorities that were dictated by individual researcher's determination to address weak components [van Santen, 1997, page 242]. The author argues that speech quality is rarely assessed formally because of practical and fundamental reasons such as conceptual multidimensionality and coverage.

In TTS, two evaluation types are differentiated: on the one hand system tests which are equal to black-box testing and which consider the overall performance of the system and on the other hand unit tests which examine a particular unit or component and which are glass-box testing methods [Taylor, 2009, p. 522]. I will concentrate here on system tests since unit test have the aim of improving a specific aspect or component of the TTS system and since the components differ from system to system, it is beyond the scope of this thesis to provide an exhaustive description of testing methods.

Two factors are normally focussed on by system tests, firstly the naturalness and secondly the intelligibility of the speech. Another factor that can be evaluated is the suitability for the used application. In reading aids for visually impaired people intelligibility and speech rate are more important features than naturalness for example.

Naturalness refers to how human-like a speech synthesiser sounds which might not be a desirable goal for all speech synthesisers. In order to assess the naturalness of the synthesised

speech, judgement testing is used which involves asking humans to judge the performance of a TTS system by explicitly rating specific attributes and characteristics of the system (e.g. the speakers voice). This method is subjective in nature. In 1994 the *International Telecommunication Union* (ITU) defined the ITU-T Recommendation P.85, a method for subjective performance assessment of the quality of synthetic speech [ITU, 1994]. The recommendation is better known as *mean opinion score* or *mean opinion scale* (MOS) and was originally targeted at telephone-based conversational services. It is a listening test in which a message is aurally presented to a human judge, who has to answer specific questions on a five-point rating scale. Each message is presented twice, first the judge has to answer specific content related questions about the message and after a second play back the judge is asked to rate the speech quality. Speech quality is assessed with the help of several recommended rating scales which can be overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness and acceptance. The use of MOS can clearly indicate if the speech quality is acceptable for a specific application but it has only little diagnostic value. The mean opinion score will be discussed in Chapter 3 in greater detail since it is the basis for the intrinsic evaluation employed in the experiments which are presented in Section 4.4. Apart from several variants such as R-MOS and X-MOS [Polkosky, 2003] which are further discussed in the section that re-visits MOS, *degradation MOS* (DMOS) is a kind of opposite version of MOS. DMOS is an impairment grading scale which measures how different disturbances in the speech signal are perceived. Therefore, the rating scales are altered to the opposite. An example is the MOS scale that evaluates the overall impression and which ranges from *excellent* over *good*, *fair*, and *poor*, to *bad* is in DMOS *imperceptible*, *perceptible* but *not annoying*, *slightly annoying*, *annoying*, and *very annoying* respectively.

Categorical estimation is another method to judge the naturalness of synthesised speech, similar to MOS. In the categorical estimation method several attributes of speech like pronunciation, speed, and distinctness are independently judged. An overview of the categories and proposed rating scales can be found in Table 2.3.

Intelligibility is the ability of a human listener to correctly interpret the words and meanings of the synthesised utterance [Jurafsky and Martin, 2009, p. 314] and can be determined with the help of comprehension tests in which humans are asked to listen to synthesised words and either choose the right one from a list or transcribe them [van Santen, 1997]. These comprehension tests can be carried out on different levels (i.e. phoneme level, word level, or sentence level). There are several tests that concentrate only on one particular phoneme in a word. The advantage of these segmental evaluation methods is that they can be carried out relatively quickly, naive listeners can be participants, the learning effect can be discarded and reliable

category	rating scale
pronunciation	not annoying - very annoying
speed	much too slow - much too fast
distinctness	very clear - very unclear
naturalness	very natural - very unnatural
stress	not annoying - very annoying
intelligibility	very easy - very hard
comprehensibility	very easy - very hard
pleasantness	very pleasant - very unpleasant

Table 2.3 – Categories and proposed rating scales for the Categorical Estimation method.

The birch canoe slid on the smooth planks.
Glue the sheet to the dark blue background.
It's easy to tell the depth of a well.
These days a chicken leg is a rare dish.
Rice is often served in round bowls.

Table 2.4 – The first five sentences of the set of Harvard Psychoacoustic Sentences.

results can be obtained with small groups (10 to 20 people). The basis for such tests is a word list that is carefully chosen with respect to the concatenative units it contains. Various word lists have been published for intelligibility tests.

One test that uses word lists to test the intelligibility on the phoneme level is the *modified rhyme test* (MRT) [House et al., 1965]. The MRT is based on 50 sets of six words which are similar in initial or final consonant (e.g. led, shed, red, bed, fed, wed). Given one word of the list the listener has to decide which of the six words has been synthesised by the system. The performance is measured for the first and final phoneme separately. Similarly to the MRT, the *diagnostic rhyme test* [Voiers et al., 1972] assesses the intelligibility of initial consonants. Listeners get the synthesis of one word out of a pair of rhyming words that differ only in the first consonant (e.g. dense, tense) and have to indicate which it is. The MRT and diagnostic rhyme test, however, have the drawback that they give information only for one or two phonemes per trial. Word level evaluation is performed in the *Bellcore test* [Spiegel et al., 1988], where the listener has to transcribe the whole word that has been synthesised.

Sentence length material is often employed to quantify the overall intelligibility of a system. The *Harvard Psychoacoustic Sentences* are a set of 100 sentences which are chosen because they represent a wide range of phonemes of the English language. The first five sentences from the Harvard sentence set can be found in Table 2.4. This fixed set of sentences can only be presented to the participants once due to the learning effect. Furthermore, the use of meaningful sentences is problematic in intelligibility tests since these can provide cues through syntax and

The wrong shot led the farm.
The black top ran the spring.
The great car met the milk.
The old corn cost the blood.
The short arm sent the cow.

Table 2.5 – The first five sentences of the Haskin sentence set.

semantics. The effect of these cues cannot be filtered out of the performance measure. One method that attempts to address this problem that was introduced by Kalikow et al. is called *SPIN* (SPeech In Noise) was created to assess hearing-impairment for speech [Kalikow et al., 1977]. In the test participants are asked to listen to between five and eight word sentences which end in a common monosyllabic word. There are two types of these sentences. In high probability sentences the context of the sentence enables the listener to predict the last word (e.g. 'Eat your soup with a spoon.'), whereas in low probability sentences the final word is unpredictable (e.g. 'We spoke about the tree.'). Each final word is presented twice to the participant, once in a high probability and once in a low probability sentence. The percentage of correct transcriptions for each final word is calculated to measure the performance. The use of SPIN limits the evaluation to final words. Another idea proposed to circumvent the problem with semantic cues is realised in the set of *Haskin Sentences* which are constructed in a way that words cannot be predicted from the sentence context. The first five sentences of the set of Haskin sentences can be found in Table 2.5. The drawback of this and other fixed sets of sentences is that participants can only be used once due to the learning effect.

A method which implements the construction of *semantically unpredictable sentences* and which also circumvents cues for word prediction is proposed by Benoît et al. [Benoît et al., 1996]. Semantically unpredictable sentences are syntactically acceptable but semantically anomalous (e.g. 'She ate the house.'). These sentences are randomly generated using a syntactic structure and a fixed word list, therefore enabling the creation of a large number of different sentences from a fixed vocabulary. Semantically unpredictable sentences usually have a simple syntactic structure and a length of around eight words in order to avoid fatigue of the listener.

Naturalness and intelligibility is evaluated at the same time in the *word pointing paradigm* which is a method to detect 'bad' units for concatenative speech synthesisers [van Santen, 1993]. The experimenter generates a text of semantically unpredictable sentences that covers as many acoustic units as possible and the listener is asked to point to words that seem problematic and to rate the seriousness of the problem.

All TTS evaluation methods discussed so far have in common that they are independent of the system in which they will be used or the task to which the system will be applied. Hence,

they are categorised as intrinsic methods. A more detailed account on intrinsic TTS evaluation methods can be found in [Goldstein, 1995].

In the following paragraphs extrinsic evaluation efforts will be reviewed. Apart from speech-to-speech translation a typical application in which TTS is employed as a component are *spoken dialogue systems* (SDS), i.e. interactive systems that conduct spoken dialogue with a user [Dybkjaer and Bernsen, 2000]. The successful performance of spoken dialogue systems depends on a large number of interaction factors and there have been various efforts to identify these factors and fit them into a useful framework to facilitate SDS evaluation [Walker et al., 2000; Möller, 2005; Möller and Ward, 2008]. Among these factors is the quality and intelligibility of the output speech. Performance analysis identifies factors and strategies that affect the quality of a dialogue. The best known approach is the PARADISE framework [Walker et al., 2000], in which the behaviour of a SDS is quantified with the help of interaction parameters (e.g. number of turns, number of help requests, elapsed time). More recently Möller and Ward [2008] provided an overview of interaction parameters, including the ones suggested in the PARADISE framework. Those are then categorized and evaluated with respect to their correlation to subjective quality judgements and their predictive power about interaction quality. It is impossible to pre-record prompts for dynamic contents in SDS but the quality of synthesised speech is still lagging behind the quality of recorded human speech. Therefore designing SDS with dynamic content always involves a trade-off between the quality and costs of recorded prompts against the flexibility of synthesised speech [Morton et al., 2011]. Several experiments have been conducted around the use of a mixture of pre-recorded and synthesised speech. In a study by McInnes and Edgington, 100 subjects were asked to assess ten different schemes for the generation of spoken output (among them pure recordings, pure synthesis and various mixtures of both) with a six item questionnaire [McInnes and Edgington, 1999]. The results showed more positive ratings for the system that mixed human speech with synthesis than the pure synthesis version. However, the experiment was criticised for the fact that it only included self reports and no assessment of the users' task performance [Gong and Lai, 2003]. Gong and Lai conducted an experiment between participants, where pure synthetic speech output was compared with a mix of speech synthesis with pre-recorded prompts. Additionally to attitudinal responses and self-ratings of task performance independent judges were asked to rate the users' performance. The results showed that although the participants self ratings were better for the mixed condition, the performance analysis of the independent judges showed that the participants actually performed better in the pure synthesis condition [Gong and Lai, 2003]. Similar experiments by Forbes-Riley et al. compared two versions of an intelligent tutoring spoken dialogue system, one with recorded prompts and the other with

a synthesised tutor voice [Forbes-Riley et al., 2006]. They investigated student learning gains, system usability, and efficiency and report that the voice quality has only a minor impact on learning achievements but the usability and efficiency is decreased in the system with the synthetic voice. To evaluate the speech output component of the smart-home, a spoken dialogue system to control different technological devices in a home, car or office environment, several studies have been carried out [Möller et al., 2004; Möller et al., 2006; Möller et al., 2007] using amongst others the *Wizard of Oz* technique which will be explained in detail in Chapter 3. The results of these studies show that when compared to recorded speech, synthesised prompts are not necessarily rated worse [Möller et al., 2006]. The same experiments were further used to examine the reliability and validity of MOS and the SASSI (Subjective Assessment of Speech System Interfaces) questionnaire by using interaction parameters for performance comparison [Möller et al., 2007].

The *Subjective Assessment of Speech System Interfaces* (SASSI) questionnaire was developed in order to facilitate user-centred evaluation of speech technologies [Hone and Graham, 2000]. The questionnaire uses a seven-point Likert scale and comprises 44 declarative statements which can be combined in six dimensions (viz. system response, accuracy, likeability, cognitive demand, annoyance, habitability and speed).

Since 2005 there is a shared task that aims to improve comparability of research techniques in building corpus-based speech synthesizers, the *Blizzard Challenge*. In the 2005 challenge for instance a dataset of 1200 utterances from a single speaker was released which could be used by the participating teams to build speech synthesisers. Later, two further databases were released together with a set of 50 texts that had to be synthesised for the evaluation. Three types of listeners judged the submissions in order to gain a ranking of the systems under test: speech experts, volunteers, and US English-speaking undergraduates.

2.7 Summary

This chapter intended to set the scene for the research described in this thesis. Initiated was the discussion with an examination of the literature about NLP evaluation which resulted in a classification of NLP evaluation methods. Intrinsic and extrinsic evaluation, the two core concepts of this research were defined. Following this, a discussion was presented that dissected the question whether HCI and NLP, the two fields that constitute a frame of reference for this thesis, can benefit from a synthesis. The main part of the chapter constituted an introduction to ASR, MT and TTS. Each application was discussed with regards to their history, main approaches, and a detailed examination of intrinsic and extrinsic evaluation methods. The next chapter addresses the methods that have been used in the research that is described in this the-

sis. It will revisit the intrinsic evaluation methods, word error rate, BLEU and mean opinion score which were mentioned in this chapter and utilised in the experiments described in Chapter 4 (*Intrinsic and Extrinsic Component Evaluation*). These will be examined in more detail. Further, the map task method and the Wizard of Oz technique which were employed in the extrinsic evaluations will be comprehensively surveyed.

Chapter 3

Methods

3.1 Introduction

The previous chapter argued that human-computer interaction (HCI) and natural language processing (NLP) share similar concerns and could benefit from a synthesis. However, as pointed out in Chapter 2, only limited interaction between those two fields can be observed. In NLP, technologies are mainly assessed intrinsically, that is with the help of automatic metrics, without consideration of their use in an application. Only a few isolated efforts can be observed where NLP technologies are evaluated in their context of use. There have been calls for more extrinsic evaluations which take into account the application for which the technology under observation is developed [Belz, 2009]. With the recent trend to combine different NLP technologies and to include them into commercial applications as well as the growing utility of multi-modal interfaces, the application of usability evaluation complying with HCI standards is of increasing importance.

This thesis aims at contributing to NLP assessment efforts by exploiting already existing extrinsic methods supplied by the field of HCI. In order to do so a series of three studies has been carried out concerning the evaluation of the three components of speech-to-speech translation, namely automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS). In these studies the state of the art intrinsic NLP methods are compared with extrinsic methods inherited from the field of HCI. An objective of this work is to investigate the correlation between HCI methods and state of the art NLP evaluation techniques and to contribute to the improvement of the intrinsic methods that are traditionally used in the assessment of natural language technologies.

This chapter introduces the methodologies that were employed in the experiment series which will be reported in the next chapter (Chapter 4). First the intrinsic evaluation methods for ASR, MT and TTS that were employed in the experiment series will be discussed in more

detail. These are the word error rate, BLEU and the mean opinion score respectively which were already introduced in Chapter 2. In the following two sections the map task and the Wizard of Oz technique are introduced which were employed in the extrinsic evaluations.

3.2 Intrinsic Evaluation Methods for ASR, MT and TTS

In this chapter the intrinsic evaluation methods for ASR, MT and TTS, that were applied in the experiments (see Chapter 4) will be revisited and characterized in more detail. This is accompanied by a more detailed discussion of their strengths and weaknesses. First, the word error rate (WER) will be described which is the most widely applied evaluation technique for automatic speech recognition, followed by BLEU, the automatic evaluation metric for machine translation. Finally the mean opinion score (MOS) for text-to-speech systems will be reviewed.

3.2.1 Word Error Rate in Detail

In Section 2.4.2 it was explained that WER is defined as the ratio of the number of word errors to the number of processed words. For isolated word recognition (IWR) this can be simply calculated as the number of substitutions (S) divided by the number of matched word pairs (N) which is the sum of substitutions and hits ($S+H$):

$$WER_{IWR} = \frac{S}{S+H} = \frac{S}{N}$$

This is relatively easy due to the nature of the task where a mapping between reference and recognition output is straightforward. In continuous speech recognition (CSR), however, this mapping is not always straightforward and has to be determined before the WER can be calculated. Therefore the reference and output sequences have to be aligned. Consider for example a person uttering the sentence “Therefore it is red." which is recognised by an ASR system that outputs the following sentence “Their four is red". In order to minimise the total count of errors ($S+D+I$, where I is an insertion) typically the Viterbi algorithm is applied [Viterbi, 1967] to align ASR output with a reference [Morris et al., 2004]. Figure 3.1 shows an illustration of an alignment for the example utterance and ASR output and the assigned labels (i.e. insertion, deletion, hit, and substitution). Based on the alignment the number of substitutions, deletions and insertions is established.

The word error rate for continuous speech recognition (WER_{CSR}) can then be calculated as the number of errors (edit distance = $S+D+I$) divided by the number of words in the reference ($N_{REF} = S+D+H$):

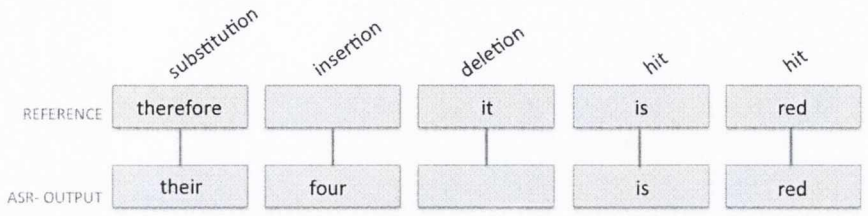


Figure 3.1 – An illustration of an alignment for the example presented in the text with assigned labels (i.e. insertion, deletion, hit, and substitution).

$$WER_{CSR} = \frac{S+D+I}{S+D+H} = \frac{\text{edit distance}}{N_{REF}}$$

Since WER for continuous speech recognition is not input-output symmetric it has no upper bound of 1. Instead the upper bound can reach $\frac{max(N_{REF}, N_{ASR})}{N_{REF}}$, with N_{ASR} being the length of the ASR output. Another drawback is that the objective of WER is to minimize the edit costs. In most applications, however, the objective is to communicate and therefore the costs for corrections might not be meaningful. Two good examples for this problem can be found in Morris et al. [2004], who suggest improved versions of WER. In the first example the authors calculate the WER for a hypothetical system that outputs a wrong word for each input word. With the number of substitutions equal to the reference length and no hits, insertions and deletions the WER score for the example is $WER = \frac{N_{REF}}{N_{REF}} = 100\%$. In their second example they consider a system that outputs two wrong words for every input word. The WER for this system is therefore $WER = \frac{N_{REF}+N_{REF}}{N_{REF}} = 200\%$. Obviously both systems convey no information at all and are useless. WER is an objective and direct metric to measure the performance of an ASR system but these examples demonstrate the drawbacks that WER is facing. First of all WER is no true percentage with an upper bound of 100% and then it is completely insensitive to semantics. However, WER still probably is the most widely applied evaluation metric for automatic speech recognition because it is a valuable tool to compare different systems and to evaluate improvements within one system.

3.2.2 BLEU in Detail

BLEU [Papineni et al., 2002], an automatic metric to assess machine translation was introduced in Section 2.5.2. It is determined by counting the number of N-gram matches, of varying length, between the system output and a set of reference translations.

In BLEU the basic unit of evaluation is a sentence, hence, the translation of a sentence is

MT output: The unit of military intelligence, serving in the thousands of soldiers, is responsible for the interception of foreign states.
Reference 1: The unit of the military intelligence service, in which thousands of soldiers do their service, is responsible for intercepting foreign states. **Reference 2:** The military unit of the intelligence apparatus, in which thousands of soldiers serve, is in charge of intercepting alien nations. **Reference 3:** The unit of the military intelligence service, in which thousands of soldiers serve, is in charge of intercepting foreign states. **Reference 4:** The unit of the military intelligence corps, in which thousands of soldiers do their service, is responsible for the interception of foreign nations.

Table 3.1 – Example of an automatic translation with a set of four reference translations.

unigram The - unit - of - the - military - intelligence - service - , - in which - thousands - soldiers - do - their - is - responsible - for - intercepting - foreign - states - . - apparatus - serve - charge - of - intercepting - alien - nations - corps – interception
bigram: soldiers serve - . the - unit of - intelligence service - foreign nations - do their - in which - nations . - foreign states - apparatus , - military unit - of the - responsible for - intercepting foreign - service , - soldiers do - their service - of intercepting - , in - serve , - the military - the unit - the interception - which thousands - is responsible - intercepting alien - intelligence apparatus - in charge - intelligence corps - , is - states . - corps , - charge of -of soldiers - for the - the intelligence - is in - military intelligence - thousands of - for intercepting - interception of - of foreign - alien nations
trigram: intelligence service , - do their service - the interception of - of soldiers do - military intelligence service - the intelligence apparatus - , in which - the unit of - nations . the - interception of foreign - responsible for the - service , is - their service , - intercepting foreign states - of intercepting foreign - foreign states . - for the interception - thousands of soldiers - of foreign nations - responsible for intercepting - states . the - is in charge - for intercepting foreign - unit of the - serve , is - soldiers serve , - , is responsible - alien nations . - of intercepting alien - intelligence apparatus , - is responsible for - of the intelligence - of the military - which thousands of - charge of intercepting - intercepting alien nations - of soldiers serve - corps , in -intelligence corps , - the military intelligence - military intelligence corps - , is in - apparatus , in - military unit of - in which thousands - . the unit - . the military - the military unit - soldiers do their - service , in - in charge of

Table 3.2 – Unigrams, bigrams and trigrams (separated by -) extracted from the reference translations.

compared with a set of reference sentences (see example in Table 3.1). Punctuation marks are treated as separate tokens. The basis for the calculation of the BLEU score is a modification of the precision measure [Baeza-Yates and Ribeiro-Neto, 2010, p. 75] (see also Chapter 2). For every sentence (S) in the test corpus the number of matching N-grams up to length n ($ngram_{matched}$) is counted and summed up. Note that N-gram matches that appear more than once in the different references (e.g. in the example "foreign states") are only counted once. Table 3.2 shows an overview of all unigrams, bigrams and trigrams which have been extracted from the example references in Table 3.1. In order to assess the *modified precision* score (p_n) the sum of matching N-grams is then divided by the number of N-gram candidates ($ngram$) in the test corpus (C).

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} \text{count}(ngram_{matched})}{\sum_{S \in C} \sum_{ngram \in S} \text{count}(ngram)}$$

In order to factor recall [Baeza-Yates and Ribeiro-Neto, 2010, p. 75] (see also Chapter 2) into the calculation and to compensate for translations with a high precision that are too short, the *brevity penalty* (BP) is introduced into the calculation of BLEU. The brevity penalty is computed over the entire corpus so that length deviations on terse sentences are not punished too harshly. In a first step the best match length for each sentence is determined which is the closest reference sentence length. This is then summed up for all candidate sentences in the corpus in order to determine the *effective reference length* (r). With the help of the total length of the candidate corpus (c), the brevity penalty can be calculated as follows:

$$BP = \begin{cases} 1 & \text{if } c > r, \\ e^{1-r/c} & \text{if } c \leq r. \end{cases}$$

Finally, with the help of the modified precision (p_n) and the brevity penalty (BP), the BLEU score can be calculated using N-grams up to a length of N and positive weights ($w_n = 1/N$) as follows:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n)$$

The BLEU can take a value between 0 and 1, with a higher score indicating a closer match to the references. A score of 1 is assigned if the translation exactly matches one of the reference translations or if the candidate translation has no brevity penalty and matches all N-grams.

BLEU has been criticised because it is relatively unintuitive and, in order to correlate with human judgments, it relies on a large number of references and sentences [Snover et al., 2006]. Furthermore, it has been shown that BLEU may not correlate with human judgements: in some

cases an improved BLEU score does not necessarily indicate an improvement in the translation quality and an improved translation quality does not receive a higher BLEU score [Callison-Burch et al., 2006]. However, BLEU is still widely applied and a better alternative has yet to be developed.

3.2.3 Mean Opinion Score in Detail

The mean opinion score (MOS) to evaluate the naturalness of a TTS system was introduced in Section 2.6.2. It will be revisited in this section and described in more detail, because it is the basis for the intrinsic evaluation that is reported in Section 4.4.

MOS is a more popular name for the ITU-T Recommendation P.85 which was approved by the *International Telecommunication Union* (ITU) in 1994 and which defines a method for subjective performance assessment of the quality of synthetic speech [ITU, 1994]. Originally it was targeted at telephone-based conversational services. The mean opinion score is a listening test type in which a message is aurally presented to a human judge, who has to answer specific questions on a 5-point rating scale. Each message is presented twice. First the judge has to answer specific content related questions about the message and after a second playback the judge is then asked to rate the speech quality. Speech quality is assessed with the help of several questionnaire items and matching rating scales. Aspects that the standard proposes are overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness and acceptance. An overview of the recommended wording for the questions and the suitable rating scales can be found in Table 3.2.

The recommendation further stipulates that the test messages should be related to a practical application and the duration of the speech should be 10 to 30 seconds in order to offer enough content for a meaningful query related to the message. Furthermore, different voice sources should be tested with a reference condition where natural speech is possibly corrupted with degradation.

Discussion arose about the items that are included in MOS and the employed scales (e.g. [Viswanathan and Viswanathan, 2005] and [Alvarez and Huckvale, 2002]). *MOS-R* is a revised version of the MOS scale developed at *IBM Voice Systems* [Lewis, 2001]. The authors propose to increase the number of scale steps for each item from five to seven, and to add two new items which they expect to correlate with the naturalness scale. The two items that were recommended for addition are voice naturalness ('Did the voice sound natural?' with a scale from *very unnatural* to *very natural*) and ease of listening ('Would it be easy to listen to this voice for long periods of time?' with a scale ranging from *very difficult* to *very easy*). Two other new items were also suggested by Polkosky and Lewis [2003], namely social impression

Overall impression: How do you rate the quality of the sound of what you have just heard?

- excellent
- good
- fair
- poor
- bad

Listening effort: How would you describe the effort you were required to make in order to understand the message?

- no effort required
- no appreciable effort required
- moderate effort required
- effort required
- no meaning understood with any feasible effort

Comprehension problems: Did you find certain words hard to understand?

- never
- rarely
- ocassionally
- often
- all of the time

Articulation: Were the sounds distinguishable?

- yes, very clear
- yes, clear enough
- fairly clear
- no, not very clear
- no not at all

Pronunciation: Did you notice any anomalies in pronunciation?

- no
- yes, but not annoying
- yes, slightly annoying
- yes, annoying
- yes, very annoying

Speaking rate: The average speed of delivery was:

- much faster than preferred
- faster than preferred
- preferred
- slower than preferred
- much slower than preferred

Voice pleasantness: How would you describe the voice?

- very pleasant
- pleasant
- fair
- unpleasant
- very unpleasant

Acceptance: Do you think that this voice could be used for such an information service by telephone?

- yes
- no

Figure 3.2 – Overview of recommended wording for questions in the ITU-T P.85 recommendation and the suitable rating scales.

and prosody. Furthermore, it has been shown that the inclusion of natural speech affects the mean ratings of TTS systems [van Santen, 1993] which is an argument against including natural voice into the set-up of the MOS experiment. An evaluation of the reliability of the ITU recommendation has also shown, that although the test can deliver consistent results with small enough variability to enable the comparison between systems, a large correlation across scales can be found. This implies that the proposed questionnaire items do not really test different aspects of the system [Alvarez and Huckvale, 2002].

3.3 Extrinsic Evaluation Methods

The previous sections revisited the intrinsic evaluation methods that form the basis for the intrinsic assessments in the experiment series which is in the focus of this thesis. This chapter will now introduce two methods that were applied in the extrinsic evaluation parts of the experiment series. The first method is the map task which was used in the extrinsic ASR evaluation (Section 4.2). This is followed by the description of the Wizard of Oz technique which was employed in the extrinsic MT evaluation (Section 4.3) and the extrinsic TTS evaluation (Section 4.4).

3.3.1 The Map Task

The *map task* was introduced in the eighties by Anderson et al. [Anderson et al., 1984, p. 70] along with several other methods as a cooperative task with the aim to improve the ability of pupils to use spoken language. In the map task a speaker instructs a hearer how to follow a route on a map with the obstacle that the maps do not match exactly. The map task differs from the other proposed task, where one speaker is authoritative and aims to transfer information to the other person, in that the information that is required to complete the task is distributed between the pupils and therefore they have to cooperate in order to achieve a shared goal [Anderson et al., 1984, p. 111].

Simultaneous to the revival of corpus linguistics in the nineties which was triggered by the increase in computational power, volatile and persistent storage capacity as well as the emergence of the World Wide Web, the map task saw a renaissance as a method to trigger task-oriented dialogues with a special aim. Anderson et al. discuss that some language phenomena may only be represented sparsely in naturally occurring corpora and therefore the attempt to draw solid conclusions may fail [Anderson et al., 1991]. To circumvent this problem researchers employ approaches with scripted monologues or extended texts, but these have the drawback that they lack spontaneity. Therefore, the map task serves by offering a controlled

environment in which it is possible to collect spontaneous spoken language corpora. The conditions are the same for both participants since no real maps are used and none of them should be acquainted with the landmarks and the route on the map. Another advantage is that the dialogues are more constrained and therefore more predictable due to the task oriented nature of the map task. Since the observer in the experiment shares the domain knowledge (about the map) with the participants it is easier to construe the speakers' communication intention, the knowledge about the mismatches enables the observer also to predict which parts of the route should be straightforward and which difficult.

For a map task experiment two participants share a schematic map which usually cannot be seen by the other. One is assigned the role of instruction giver; the other is the instruction follower. The map of the instruction giver comprises a route. The participants have to collaborate to reproduce this route on the map of the instruction follower. In order to guarantee relative spontaneity no restrictions are placed on what either can say. Both participants know the starting point, but there are mismatches between landmarks. The participants are explicitly instructed that their maps are not identical. In Figure 3.3 one of the instructor maps used in the map task experiment which is described in Chapter 4 can be found. The respective follower map is shown in Figure 3.4. One landmark in which these two maps for example differ is the bird below the hare which is missing in the instructor map. Different sets and restrictions to this basic set-up of the method are imaginable and have been used. Maps can be designed with a research goal in mind. The landmarks on the maps can be depicted only or named and landmark names can be chosen according to phonological interest which has been done in the extrinsic ASR evaluation in Chapter 4. Another factor in the set-up that can be changed for example is whether the participants can have eye contact or not.

A corpus collected in a map task setting is the *Human Communication Research Centre (HCRC) map task corpus* [Anderson et al., 1991]. A total of 64 participants, all native speakers of the standard Scottish English of Glasgow, produced 128 dialogues which have been recorded, transcribed and annotated. Sixteen different pairs of maps were used and each participant performed the task in the role of instructor twice and in the role of follower twice. Variables that were changed in the experiments were the familiarity of the participants (i.e. familiar and unfamiliar) and the eye contact between instructor and follower (i.e. with and without eye contact). Therefore, a total of 32 dialogues for each category were produced which was released for research purposes.[1]

Several independent analyses have been carried out on the material of the HCRC map task corpus. Boyle et al. for example analysed non-verbal behaviour in communication by com-

[1]The material can be accessed online at `http://groups.inf.ed.ac.uk/maptask/` (last accessed 12/06/2012).

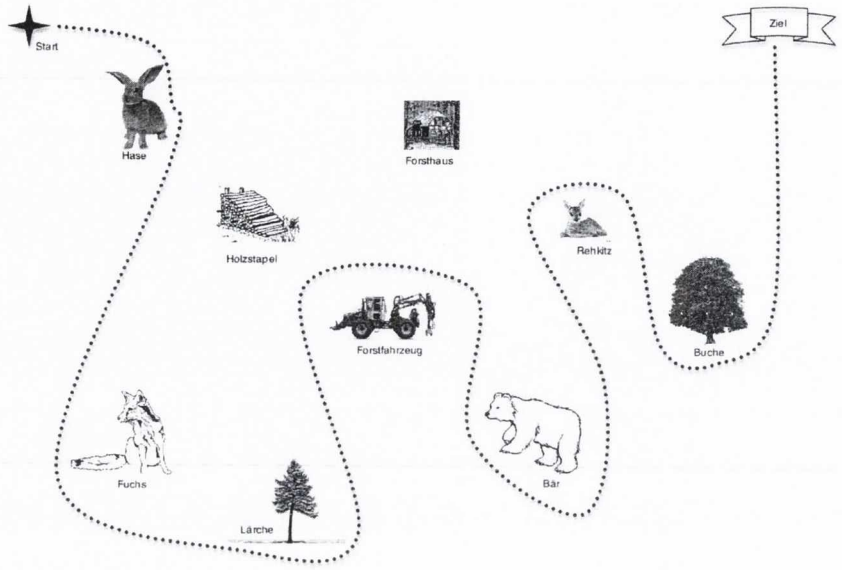


Figure 3.3 – An example of the map of the instructor which was used in the extrinsic ASR evaluation described in Chapter 4.

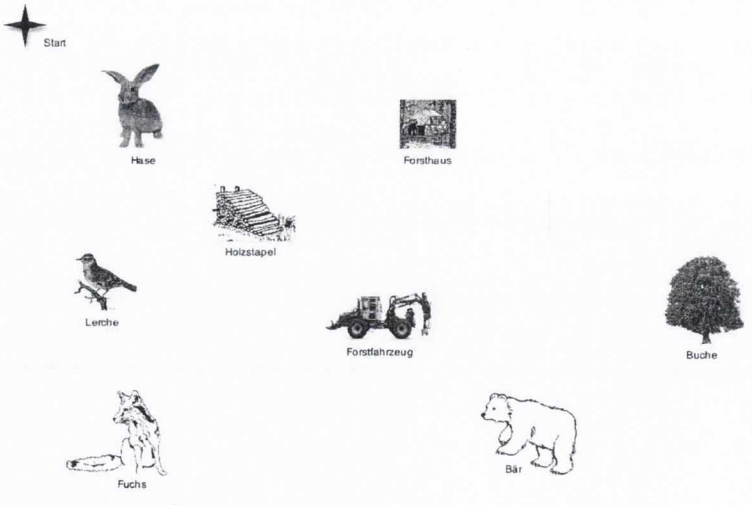


Figure 3.4 – The follower map, respective to Figure 3.3.

paring various dialogue parameters (e.g. task performance, number of turns, number of word tokens, words per turn) for dialogues in which the participants were able to see each other with those where they were not [Boyle et al., 1994]. The results showed that information transfer and turn taking is improved in situations where the visibility of the conversational partner was granted, due to visually transmitted non-verbal signals. In the disguised mode the participants interrupted their partners more often and the verbal feedback was increased because the participants used more back channel responses. Carletta describes the use of the map task corpus to identify human strategies to take on different levels of specificity for referring expressions in order to gain a more 'natural' dialogue structure in the *JAM* system by simulating human risk choices [Carletta, 1992]. Carletta et al. also used the map task corpus to collect information about self-repairs and describe a new coding scheme for these [Carletta et al., 1993]. Another approach, where spatial descriptions are treated like referring expressions and the map task corpus is used as domain model is described by Varges [2005]. In 2006 Reitter et al. used the Switchboard and HCRC map task corpus to examine structural repetition in order to investigate priming effects involving arbitrary syntactic rules in spoken dialogue [Reitter et al., 2006b]. The term *priming* is used to describe the fact that in human communication, dialogue partners often align their linguistic behaviour in terms of lexical and grammatical choices. Alignment, entrainment, and lexical convergence are other expressions which are used to denote this widespread and well researched phenomenon. In Chapter 4.5 this phenomenon will be revisited, because it was also observed in experiments which form the basis for this thesis. Reitter et al. also discovered that between-speaker priming is stronger in task oriented dialogues analysing map task dialogues [Reitter et al., 2006a].

The HCRC corpus has also been replicated several times with different languages or input-output modalities. In such cases the HCRC maps were re-used and the setting was reproduced in order to achieve comparability between the corpora. This was done for example for a number of other forms of English such as American English and Australian English as well as for a number of other languages like Dutch, Italian, Japanese, Swedish, Occitan, and Portuguese.[2] An overview of the different replications and their extent can be found in Table 3.3. The HCRC map task setting was further used to examine the communicative outcome of interaction between non-native speakers [Yule and Powers, 1994], with a special focus on the approach of speakers to represent their world of reference, how the non-native speakers recognise discrepancies between their world of reference and how they negotiate solutions to these disagreements. The Japanese replication of the HCRC map task corpus is used by Koiso et al. [1998] to investigate syntactic and prosodic features at moments of turn-taking and when

[2] A collection of other language representations can be found at `http://prosodia.upf.edu/atlasintonacion/Map-Tasks/index-english.html` (last accessed 25/04/2012).

language	extent	number of participants
Glaswegian English	128 dialogues	64
American English	16 dialogues	8
Australian English	19 hours	204
British English	36 hours	117
Dutch	8 dialogues	4
Italian	44 dialogues	
Japanese	128 dialogues	64
Portuguese	64 dialogues	32
Swedish	50 minutes	4
Catalan	6 hours	30
Occitan	1 hour	30

Table 3.3 – Overview of corpora which use the HCRC maps and setting as described by Anderson et al. [Anderson et al., 1991].

back-channels occur.

The *Defence and Civil Institute of Environmental Medicine (DCIEM) Map Task Corpus* reuses the HCRC maps in order to examine effects of sleep deprivation and drug treatment on speech production [Bard et al., 1996]. 216 dialogues were produced by Canadian army reservists in three conditions (baseline, sleepless and recovery period) by groups which were treated with placebos or drugs reputed to counter effects of sleep deprivation. The DCIEM corpus is also available to researchers.

A research topic for which the map task seems to be very beneficial is the analysis of referring expressions which have the goal to distinguish a target from a concurrent distractor set. The aim of referring expression research is to develop algorithms that are able to single out an item by a set of attributes (e.g. type, size, colour) as human-like and efficient as possible. As previously discussed, Carletta [1992] used the HCRC map task corpus to analyse human strategies taking on different levels of specificity for referring expressions. For some experiments the HCRC map was altered to gain a more controlled experiment setting, e.g. by introducing many landmarks of the same type but of different size and colour. Guhe et al. [2007] for example introduced new maps with 'ink blots' to obscure the colour of some objects in the maps of the instruction followers and analysed the use of colour terms in referring expressions. The so collected *iMap corpus* consists of 256 dialogues recorded from 64 participants. Their experiment showed that dialogue partners adapt to the property of the task environment by using fewer colour terms over time. These results are later refined by an analysis of how and why the feature terms in referring expressions change over time [Guhe and Gurman Bard, 2008a,b]. The *iMap corpus* was later re-used in an analysis of subsequent references and the impact of the reference history on forming conceptual pacts in discourse context (i.e. other entities that have been mentioned earlier in the dialogue) and visual context (i.e. visually available objects

near the intended referent) [Viethen et al., 2010].

Simultaneous to the emergence of new forms of communication such as chatting another research branch materialised for which the map task is applied widely: computer-mediated communication. In map task experiments concerning computer mediated communication the input and output modalities are usually varied. Newlands et al. for example studied how people adapt to text-based communication, with a focus on strategy changes of novice users [Newlands et al., 2003]. Therefore, the participants had to carry out three map tasks over three consecutive days. The authors used the *conversational game analysis* as framework to investigate pragmatic function and content of the messages. The results show that although in the text condition the participants showed poorer performance compared to the spoken condition, the participants improved over time as they gained experience. Louwers et al. [2006] addressed further communication channels such as eye-gaze, facial expressions and torso movement. For these experiments the participants were separated to ensure they could not see and hear each other directly. They communicated through microphones and headphones and were able to see the upper torso of the other participant which was video recorded and appeared on a monitor in front of them which also showed the map. Dialogue moves, eye-gaze (i.e. times of total fixation and number of blinks) and the intensity limit and duration of pauses were analysed. The results show that eye-gaze, facial expressions and pauses correlate at certain points and can be identified by the speaker's intention behind the dialogue move. A similar setting was applied in an experiment by Clayes and Anderson, where participants had to execute four experimental tasks, among them the map task [Clayes and Anderson, 2007]. There were two conditions in the experiments. In the first condition the participants were represented by a basic static avatar, in the second condition they could see the other's video image. The aim of the study was to explore the ways through which users communicate via avatars. Eye-tracking data was collected and interpreted and the results show that participants gaze more often on the video images which helps to reduce interruptions but is not necessarily an advantage in a problem-solving scenario. It was further shown that the number of gazes is dependent on the communicative context.

Automatic speech recognition as input was used in an explorative experiment to elicit human error recovery strategies in computer mediated communication [Skantze, 2005]. In a simulated telephone conversation in which an operator had to describe a route to an information seeker, speech recognition was used to corrupt the speech of the information seeker. Forty dialogues were collected and transcribed. Dialogue acts were annotated manually on the spoken, not the recognized, utterances and each user utterance was categorised in one of four groups which indicated how well it was immediately understood by the operator. The analysis of the

error recovery situations reveal that it is a common strategy to ask task related questions that confirm the speaker's hypothesis about the situation rather than to signal that something was not understood and to inquire a repetition of what was said.

3.3.2 The Wizard of Oz Technique

The *Wizard of Oz (WOZ)* technique is a well established early stage prototyping method by which a system can be tested without first having to build it [Faulkner, 2000]. Kelley, who simulated a calendar application that could be navigated with natural language input [Kelley, 1983] dubbed the method the 'Oz paradigm' with reference to the well-known children classic *The wonderful wizard of Oz* [Baum, 2008]. In the story a tiny man mocks the main characters with the help of mirrors and other utilities and feigns he is the mighty Wizard of Oz with immense magical power. Equivalent to this, the WOZ technique presents the user with what appears to be a working system, while a human operator, the so called wizard, who is not visible to the user, simulates the role of the system as a whole or a part of the functionality. The advantage is obvious: by using the WOZ technique researchers can simulate the functionality or the user experience of a system that has not been implemented yet, it can even be applied if the proposed system goes beyond what is feasible. Early feedback through prototyping is important to build usable products [Gould and Lewis, 1985]. In the design of graphical user interfaces this can easily be achieved with methods such as sketching or wire-framing, but these methods prove insufficient when it comes to the evaluation of speech technologies [Schlögl et al., 2011]. In these cases the WOZ technique has proven particularly useful. It is a powerful method when the input modality has a high computation/cognition ratio in the sense that it can be only partially decoded by computers but is easily understood by humans [Dybkjaer et al., 1993]. This is the case for most speech technologies. Practically for all speech technologies the real performance of the systems is still error prone and implementing the technology involves an extensive engineering effort. Hence, the WOZ technique has a long tradition in the design of speech systems [Gould and Lewis, 1983]. It has also been adopted by the field of human-computer interaction. One of the pioneers of HCI, Bill Buxton, grants the WOZ technique the same positive features as paper prototypes by enabling designers to explore interactive systems before they are implemented [Buxton, 2007]. Therefore, they are relatively cheap compared to high fidelity prototyping techniques, quick to realize, disposable and they have only sufficient fidelity to serve the intended purpose [Buxton, 2007]. While the user is made to believe he operates a functional product the wizard can examine the functionality that is needed. This examination allows designers to gauge user expectations and obtain feedback on the operation of the product under development.

The history of Wizard of Oz experiments is long. More than four decades ago Erdmann and Neal [1971] simulated a self-service airline ticket kiosk using a human which was disguised from the participants. More renowned is the experiment by Gould and Lewis [1983], who examined the idea of automatic speech recognition with the 'listening typewriter'. The idea of WOZ prototyping has also been taken up to discover possible commands from novice users in order to make the use of command line interfaces more intuitive [Good et al., 1984]. Novice users were asked to use an electronic mail command line interface without instruction or help texts. To create the illusion of an interactive system a wizard intercepted and corrected the commands of the users if necessary. In the process of this iterative development cycle the set of commands was refined in every iteration. Other early applications of the method examined the use of help texts as implemented in software to advise a user on how to recover from an error. Hill and Miller examined the cognitive processes of an advisor giving help recommendations [Hill and Miller, 1988]. In the setting the advisor acted as wizard sending natural language answers to the participant who could direct natural language queries to the simulated help system. Similarly Carroll and Aaronson conducted a WOZ study to explore the usability of an intelligent advisory interface [Carroll and Aaronson, 1988]. Ten years later, Davis [1998] took up the subject again and investigated the benefits of the two help paradigms, 'active help', where the user is interrupted when appropriate and 'back channel' communication, where the user can send messages to the help system. The results indicate that active help is preferred by users, however, the author points out that since the development of active help systems is associated with high development effort the issue should be investigated further and he recommends the WOZ technique for such studies.

The area of applications for WOZ experiments is very wide. As mentioned before, especially the field of speech technology is rich in WOZ experiments. The use of WOZ was very extensive in two main application areas, namely the collection of corpora and the design of human-computer dialogues. Often these go hand in hand since corpora are often collected in a special context with the aim to improve dialogue strategies for special situations. One example where this holds true is the collection of mathematic tutorial dialogues in German with the objective of developing mathematical tutoring dialogue systems that employ an elaborate natural language dialogue component [Benzmüller et al., 2003]. German conversations of multiple parties were recorded and collected in the *PIT corpus* at the *University of Ulm* with the objective of developing intelligent and user-friendly human-computer interaction in multi-user environments [Strauß et al., 2008]. The corpus comprises 75 dialogues where two dialogue partners are assisted by a simulated computational assistant to discuss their choice of a restaurant. Experiments done at *LIMSI* in the design, development and evaluation of spoken language

dialogue systems for information retrieval tasks (e.g. for a train or flight information system, a multi-modal multimedia service kiosk, etc.) are reported by Lamel [1998]. A comprehensive overview of the data collection with the help of the Wizard of Oz technique to bootstrap systems and the following processing steps of the collected data is given. Fabbrizio et al. propose the use of an automated wizard which reuses previously labelled and transcribed data from similar domains to improve the informativeness of the collected data in order to tune speech recognisers more task specific [Fabbrizio et al., 2005].

In the design of human-computer dialogues, Whittaker et al. [2002] presented a study where a wizard is used to discover dialogue strategies in the restaurant domain. Furthermore, dialogue design for voice controlled telephone dialogues have been in the focus of a study which investigated four distinct navigation structures with different cognitive workload (e.g. tree oriented) with the aim of developing more user-friendly audio navigation structures for voice-controlled telephone services [Goldstein et al., 1999]. Likewise, for a speech activated city guide the effect of tree oriented, hierarchical dialogue structure was compared with an office filing metaphor which offers different metaphor-related menu options at each level [Howell et al., 2005].

At the centre of attention of early WOZ studies were applications with pure speech or text interaction, but in the early nineties a trend has started to use WOZ to evaluate multi-modal systems. Among the first studies on multi-modal interaction was an experiment with 36 participants, where the use of gestures in combination with speech to manipulate graphic images on a computer screen was analysed [Hauptmann, 1989]. The results found an extraordinary amount of uniformity in the way subjects communicated with gestures and speech and therefore encouraged the development of multi-modal interfaces. WOZ has furthermore been used to evaluate the perception of different input modalities by Rajman et al. [2006]. Speech as input modality for a flight booking system was examined by Karpov et al. [2008]. The WOZ technique was further employed to guide the creation of an augmented reality interface that uses multimodal input like natural hand interaction and speech commands [Lee and Billinghurst, 2008]. Mäkelä et al. employed the WOZ method to simulate a virtual doorman [Mäkelä et al., 2001] and Bradley et al. assessed people's reactions to multimodal, intelligent, personal natural language interfaces, called companions [Bradley et al., 2009, 2010]. Furthermore, human-robot interactions with children between four and eight years has been explored [Saint-Aimé et al., 2011]. Similar experiments have been conducted to analyse speech based design ideas, such as the smart home [Gödde et al., 2008] and VICO, the driving assistant [Geutner et al., 2002].

With the increasing use of the Wizard of Oz technique several tools have been developed to support researchers in conducting WOZ studies. Often the systems are custom made and

developed specifically for one experiment, so that they can be regarded as throwaway tools. However, several tools have been developed to conduct more general kinds of Wizard of Oz experiments. The *CSLU* toolkit [Sutton et al., 1998], the *Olympus* dialogue framework [Bohus et al., 2007] and the *Jaspis* dialogue management system [Turunen and Hakulinen, 2000] focus on the dialogue management aspect of Wizard of Oz studies and therefore usually provide the experimenter with an interface to specify a dialogue flow and the possibly to incorporate other natural language components such as automatic speech recognition. They are mostly employed to explore language-based interactions between humans and a system. The objective behind using these tools in experiments is typically the improvement of the dialogue or to test and improve language technologies. These tools have in common that they are in a way generic enough to be reused by other researchers with similar objectives. Several WOZ tools have been developed that have a very narrow focus on a special aspect under observation. The *SUEDE* toolkit [Klemmer et al., 2000] is a speech interface prototyping tool based on the concepts of example-based test scripts and prompt/response state transitions. The toolkit also supports the collection of the test data. The *NEIMO* platform [Coutaz et al., 1996] even supports the observation of multimodal interaction and also supports the collection and analysis of the experiment data. *SUEDE* and *NEIMO*, however, have the problem that they use software that is no longer in use, leading to incompatibility problems. As part of the *OpenInterface* software framework which is an open-source platform for developing interfaces that communicate intelligently through several modalities, *OpenWizard* aims at enabling developers to evaluate non-fully functional multi-modal prototypes using the Wizard of Oz technique [Serrano and Nigay, 2010].

The development of a generic WOZ interface is in the focus of research conducted at *Trinity College Dublin* [Schlögl et al., 2010b,a]. *WebWoz* is a platform that aims to support the conduct and analysis of experiments, and offers the flexible integration of language technology components such as ASR, MT or TTS into WOZ experiments. The tool also offers components that provide ways of capturing and analysing contextual information and domain knowledge while carrying out WOZ experiments. Recently *WebWoz* was combined with *MySpeech*, a prototype of a computer-assisted pronunciation training system to test *MySpeech* functionality before implementing it and gathering user data to improve the pronunciation analysis model [Cabral et al., 2012a,b]. The *WebWoz* platform was also utilized in the extrinsic evaluation experiments which were carried out in the processes of the research which forms the objective of this thesis.

A crucial part of Wizard of Oz studies is the conduct of the wizard. The role of the wizard can range from 'controller' who simulates a technology to 'moderator', who monitors output

and decides which output is presented. It can also span to 'supervisor' who overrides undesired output [Dow et al., 2005]. One of the main goals in WOZ studies is to make sure that the human intervention is imperceptible to the test user interacting with the system. In order to convey the illusion of a real system it sometimes is necessary for the wizard to show flawed behaviour [Peissner et al., 2001]. Another way to deal with this challenge is to provide the wizard with an appropriate interface that offers sufficient support to fulfil the task efficiently. The task of the wizard and how it can be supported is therefore in the focus of ongoing research that is coupled with the development of *WebWoz* [Schlögl et al., 2010a, 2011]. It has also been proposed that the task of the wizard should be divided into subtasks which are carried out by multiple wizards at the same time [Salber and Coutaz, 1993]. This setting is also supported by experiments to evaluate and design the *Archivus* multi-modal meeting browsing and retrieval system, where the wizard's role was split up between an 'input' and an 'output' wizard [Ailomaa et al., 2006; Melichar and Cenek, 2006].

The Wizard of Oz technique has been employed in two of the experiments that will be described in the following chapter. In the first case it served as extrinsic method for the evaluation of machine translation. In the context of machine translation, studies that use WOZ are rare. In the *Verbmobil* project which aimed at developing a system capable of interpreting dialogues, the method was employed to explore strategies and phenomena in interpretation [Krause, 1997]. Bederson et al. [2010] also employ WOZ experiments to explore the idea of an iterative translation process supported by a combination of machine translation with monolingual human speakers. In the second case the WOZ technique has been employed as extrinsic evaluation method for the text-to-speech system under observation. In parallel to the research that is described in this thesis an experiment has been carried out by Janarthanam and Lemon [2009] which has a similar setting to both experiments described in Section 4.3 and Section 4.4. This experiment, however, was targeted at collecting and annotating data for referring expression generation. The task of the participants was to set up a broadband Internet connection with the help of instructions from a WOZ system with speech output. No machine translation was involved in the scenario. The utterances were intercepted by a wizard, whose task it was to annotate the content of the participant's utterances.

3.4 Summary

This chapter explained the theoretical background of the evaluation methods that have been employed in the experiments which will be outlined in the next chapter. The word error rate, BLEU, and MOS are at present the state-of-the-art evaluation methods for ASR, MT and TTS respectively. All metrics have been criticised for their drawbacks. The word error rate which

was employed in the intrinsic ASR evaluation, is a valuable tool to compare different systems and to evaluate improvements within one system. However, it is completely ignorant of semantics and it is no true percentage with an upper bound of 100%. BLEU is the traditional evaluation metric for machine translation. It was criticised for its missing intuitivity and the fact that a large number of references and sentences are needed in calculations to correlate with human judgments. Furthermore, it has been shown that BLEU may not correlate with human judgements in some cases. BLEU is still widely applied and a better alternative has yet to be developed. TTS evaluation is especially difficult and no widely employed automatic metric has been developed to present. The mean opinion score, was reviewed in more detail since an adapted version was employed in the intrinsic TTS evaluation. MOS is a listening test and discussion arose mainly about the items that are included and the recommended scales. The extrinsic evaluation of ASR employed the map task. The map task is a cooperative task which enables participants to involve in spontaneous dialogues. Therefore, it was mainly used to collect corpora that form the foundation for analysis of certain language aspects, such as the use of referring expressions. The Wizard of Oz technique was outlined in this chapter because it was employed in the extrinsic MT and the extrinsic TTS evaluation. WOZ is an early stage prototyping method by which a system can be tested without first having to build it. The method is a powerful technique when the input modality has a high computation/cognition ratio in the sense that it can be only partially decoded by computers but is easily understood by humans and has therefore a long tradition in the design of speech systems. Neither the map task nor the Wizard of Oz technique were employed in the evaluation of ASR, MT or TTS similar to the experiments that will be described in the following chapter. The next chapter forms the backbone of the research for this thesis. The intrinsic and extrinsic experiments around the evaluation of ASR, MT and TTS will be discussed, their set-up outlined, and the results reported.

Chapter 4

Intrinsic and Extrinsic Component Evaluation

4.1 Introduction

This chapter focuses on the four experiments that are the backbone of the research presented in this thesis. Three self-contained experiments were carried out comparing intrinsic and extrinsic evaluation methods for each of the main components of speech-to-speech translation. Although each experiment is part of a well articulated whole, the common topic of this research is the comparison of intrinsic and extrinsic evaluation methods. The fourth experiment expands on observations made in the third experiment which indicate that stand alone component evaluation or end-to-end evaluation might be insufficient and do not unveil important aspects which influence the output quality of the system.

The automatic speech recognition (ASR) experiment (c.f. Section 4.2) compares results gained from a word error rate analysis with outcomes of an extrinsic evaluation that employed the map task method. In the focus of this experiment is an attempt to improve the state-of-the-art evaluation method, word error rate. The machine translation (MT) experiment (c.f. Section 4.3) employed an adapted version of the translation error rate and compares these intrinsic results with results that were gained in an extrinsic study which employed the Wizard of Oz technique. The third experiment concerns text-to-speech (TTS) evaluation (c.f. Section 4.4). An adapted version of the mean opinion score was employed to assess a set of synthesised utterances. These intrinsic results are compared with interaction data gained from an extrinsic evaluation that also employed the Wizard of Oz technique. An additional aspect of this experiment is the combination of the MT component with the TTS component. This reveals that the usual evaluation practice in speech-to-speech translation which either concerns stand-alone

evaluation of the components or end-to-end evaluation, might not be enough to tell the full story. An aspect that also needs consideration is the combination of component pairs and their evaluation. In the experiment that combined the MT with the TTS component it was observed that participants regularly repeated erroneous machine translation output when answering the system's questions. Motivated by this observation a fourth experiment was carried out. A targeted question-answering study with thirty participants was performed which investigated the effect of combining MT with TTS in a speech-to-speech translation setting. The experiment aimed at uncovering whether people align their answers to the wording presented in the questions or if they correct the given alternatives or engage in some sort of attempted repair. The outcome of this experiment alerts to the possible negative effect of MT errors in S2S translation on the ASR component.

The chapter is organised according to the components starting with automatic speech recognition, followed by machine translation and finally text-to-speech. For each of the components first the intrinsic evaluation is described before the extrinsic evaluation is outlined. Further, the outcomes of the experiments are discussed and compared. The following section outlines the targeted question-answering study, before the results of all experiments are recapitulated in the summary of this chapter (c.f. Section 4.6)

The experiments which are described in this chapter, went through an ethical approval process with the ethics committee of the *School of Computer Science and Statistics* at *Trinity College Dublin* since all involved human participants and included some form of live recording (e.g. audio records). Furthermore, for each study three pilot tests were carried out which informed the experiment set-up and gave valuable insight into possible error sources.

All studies, except the experiment that tests alignment in synthesised machine translated communication (Section 4.5), are based on a dialogue scenario. A dialogue setting was chosen because in Human Computer Interaction spoken dialogue offers an immediate and human-like means of communication that suits many applications in for instance hands-busy, eyes-busy situations. In addition, the state of the art in spoken dialogue systems has grown to a point where such applications are feasible. Text used in a dialogue differs considerably from written text [Biber, 1988]. Dialogues are, for example, more interactive and contain direct references (i.e. to the addressee). Language production in dialogue is prompt whereas in written text typically more consideration is put into the formulation. Even in situations which deviate from natural, face-to-face dialogue settings, such as spoken-language dialogues between remotely located participants or interactive, time-constrained situations where different modalities are employed, language use varies with respect to standard written text. It is therefore not obvious that, for example, a machine translation system that performs well on written text will also yield

good usability results in interactive systems. Similarly, poor machine translation performance in text does not necessarily imply poor usability in interactive contexts.

4.2 Automatic Speech Recognition

Automatic speech recognition is traditionally evaluated through the intrinsic evaluation metric *word error rate* which has been introduced in the Chapter 2 and was described in detail in the *Methods* Chapter. These chapters discussed that the word error rate has several practical drawbacks. One is, for example, that it is not a true percentage with an upper bound of 100%, the other is that it is completely insensitive to semantics. The latter handicap impacts task specific evaluations which are rare and mostly concern the evaluation of ASR as a component in spoken dialogue systems or speech-to-speech translation.

This section reports on a study to examine ASR output intrinsically and extrinsically using the map task evaluation method to gather data from dialogues that were mediated by a speech recognition system. For the intrinsic evaluation, the collected dialogues have been transcribed and the transcriptions were used as reference to calculate the word error rate. In the extrinsic evaluation logging data and the small corpus of dialogues was used to assess the performance of the ASR system in a specific context.

It can be expected that in cases where the ASR accuracy is low, interaction measures like the number of words uttered or the time to complete the task are negatively affected since communicative problems will occur. Therefore, the hypothesis of the experiment is that where the WER is especially bad, this can be reflected in the logged data from the interactions. In the following section the correlation between the word error rate and such interaction measures will be analysed.

Materials: For each of the four scenarios (i.e. office, wild west, river, forest) a pair of instructor and follower maps was designed (all maps can be found in Appendix B). The resulting eight maps have between eight and twelve landmarks, labelled in German. Each scenario includes a pair of homophone landmarks in the instructor and follower map. These homophones differ in their position between instructor and follower maps. In the wild west scenario for example there is a pair of scales (German *'Waagen'*) on the left hand side of the follower map, whereas the instructor map showed two wagons (German *'Wagen'*) on the right hand side (cf. the map in Figure B.3 in Appendix B). One landmark on the instructor maps of the office and the wild west scenario was duplicated, a flip-chart and tepee respectively. Apart from this all other landmarks in the four maps were unique. The instructor maps show a route with a designated starting point and goal. In their maps the route and the goal have been left out. The follower maps comprise, among several landmarks, only of the starting point in the same posi-

	follower map	instructor map
Wild west		
homophones	wagon (Wagen)	scales (Waagen)
# of landmarks	10	12
duplicate landmarks	-	tepee
	shared properties	
# of identical landmarks	9	
route length	136 cm	
map orientation	portrait	
Office		
homophones	faeces (Kot)	code (Code)
# of landmarks	8	10
duplicate landmarks	-	suit case, flip-chart
	shared properties	
# of identical landmarks	7	
route length	102 cm	
map orientation	portrait	
River		
homophones	berries (Beeren)	bears (Bären)
# of landmarks	8	9
	shared properties	
duplicate landmarks	none	
# of identical landmarks	7	
route length	131 cm	
map orientation	landscape	
Forest		
homophones	larch (Lärche)	lark (Lerche)
# of landmarks	8	9
	shared properties	
duplicate landmarks	none	
# of identical landmarks	7	
route length	116 cm	
map orientation	landscape	

Table 4.1 – Overview of the features of the four maps (e.g. number of landmarks, duplicate landmarks, length of the route).

tion as the starting point of the instructor map. As can be seen in the overview of the features of the different maps and scenarios which can be found in Table 4.1 the office scenario has the shortest of the four routes but comprises of two landmarks that appear twice on the instructor's map and only once on the follower map (suitcase and flip-chart). The two maps share seven out of the ten landmarks that the instructor map shows. The wild west scenario features the longest route, and also includes most landmarks in the instructor's map. It only includes one duplicate landmark on the instructor map (the tepee). Furthermore, out of the twelve landmarks on the instructor's map nine are shared by the follower map. With 131cm the route in the river scenario is only marginally shorter than in the wild west scenario, but it comprises of only nine landmarks on the instructor map and none of them is a duplicate. The forest scenario also shows only nine landmarks on the instructor map (no duplicates), and with 116cm the route length is second shortest. The length of the route, the number of landmarks, homophone and duplicate landmarks are expected to affect the difficulty of the task and therefore impact on the time that is needed to describe the route and the number of turns or words uttered by the instructor and follower.

Subjects: Sixteen participants performed the map task in groups of two. Their ages ranged from 25 to 42, with an average age of 29.06 years (4.52 standard deviation). All were native speakers of the German language, and had no or only minor experience with ASR systems. None had ever encountered a map task experiment before. Twelve of the participants were male, and four were female. It was a requirement that participants who performed the map task together knew each other before the experiment. An overview of the demographics of the participants and their familiarity with ASR systems can be found in Table 4.2.

4.2.1 Intrinsic Evaluation

Chronologically the extrinsic experiment was carried out first. On basis of the utterances that were recorded during the extrinsic experiment the intrinsic evaluation was performed. Owing to the consistency of this thesis, however, the intrinsic assessment and results will be reported before the extrinsic study set-up and results (c.f. next section). The extrinsic experiment used a map task scenario in which the instructor had to dictate directions to an ASR system, and the recognised text was sent to the follower. A protocol of the sent utterances was recorded. The directions that were uttered by the instructor were recorded and transcribed. On basis of this reference transcription and the protocoled ASR transcriptions, the word error rate was calculated for each utterance that was sent of by the instructor in the ASR mode. For each scenario an overall WER was calculated. Since every participant had to be instructor in the ASR mode once, the experiment resulted in one WER value for each participant.

participant	sex	age	experience with map task	experience with ASR system
01	male	29	none	none
02	male	25	none	none
03	male	29	none	minor
04	female	28	none	none
05	female	32	none	none
06	female	29	none	none
07	male	27	none	none
08	male	33	none	minor
09	male	26	none	none
10	male	25	none	none
11	male	42	none	none
12	male	34	none	minor
13	male	25	none	none
14	male	26	none	none
15	female	25	none	none
16	male	30	none	none

Table 4.2 – Demographics and experience with map task and ASR systems for the participants, who took part in the ASR study.

Results: The number of words that instructors uttered in the ASR mode in order to lead the follower through the map ranged from 140 to 923 words per map, with an average of 370.125 (216.8 standard deviation). This number of words uttered by the instructor is the reference length which was the basis for the calculation of the word error rate of each interaction. The word error rate ranged from 0.1148 to 0.4786 with an average of 0.2451 (0.097 standard deviation). A scatter plot (Figure 4.1) of the ASR errors in relation to the length of the reference clearly shows that participant 7 and participant 10 are outliers in the data set. Both needed considerably more words to describe their routes than the other fourteen. The ASR accuracy was especially bad for participant 4, 7, and 16, and very good for participant 8, 10, and 14. However, for participant 5, 12 and 16, the reference length is very low and, therefore, single recognition mistakes have more influence on the WER results than in interactions where the number of words uttered is very high.

4.2.2 Extrinsic Evaluation

The extrinsic part of the ASR experiment employed the map task method which was described in detail in the *Methods* Chapter. In this section the detailed set-up of the experiment as well as the results will be discussed.

Procedure: A pair of participants was invited into the lab to perform the map task. It was required that both were native speakers of German and knew each other before taking part in

participant	map	reference length	recognition mistakes	**WER**
01	office	315	70	**0.2222**
02	forest	311	67	**0.2154**
03	office	472	118	**0.2500**
04	forest	307	111	**0.3616**
05	office	163	36	**0.2209**
06	forest	338	74	**0.2189**
07	office	773	299	**0.3868**
08	forest	275	43	**0.1564**
09	wild west	527	119	**0.2258**
10	river	923	106	**0.1148**
11	wild west	146	49	**0.3356**
12	river	286	50	**0.1748**
13	wild west	429	86	**0.2005**
14	river	265	44	**0.1660**
15	wild west	252	49	**0.1944**
16	river	140	67	**0.4786**

Table 4.3 – Reference length, number of recognition mistakes and word error rate results for each ASR interaction.

the experiment. After a general introduction the participants were separated in two different rooms, each equipped with a computer. The experimenter did leave them in the rooms alone in order to not influence the experiment. Before the experiment started each participant had to perform the standard initial model training with the *Dragon* dictation software,[1] further training mechanisms for the speech model were switched off to enable comparability between subjects. At the beginning of the experiment the instructor and follower received one of the eight maps each (as described in the Materials Section) according to their role and the designated scenario which was printed out in A4 format and presented on paper. *Skype*[2] was used to facilitate a communication channel between the instructor and follower. Furthermore, the experiment was audio recorded and the follower's screen was captured using the *Screenflow* screen casting software.[3] The audio recordings comprised two channels, one for each participant. There were two modes of communication. In the telephone communication mode, both participants were able to talk and listen to each other without being able to see the other participant, comparable to a telephone communication. In the ASR mode, the loudspeakers of the follower's computer were muted and the instructor used the *Dragon* dictation software to input text into the *Skype* chat field. The instructor could follow what was recognised on his computer screen, but was not allowed to make any changes to the recognized text by typing. The participant who acted as instructor was free to decide when to send the recognized utterances to the follower by

[1] The German version of *Dragon Naturally Speaking* Version 11 (academic version) by *Nuance* was used in the experiment.

[2] `http://www.skype.com/` (last accessed 21/08/2012)

[3] `http://screenflow.softonic.de/` (last accessed 21/08/2012)

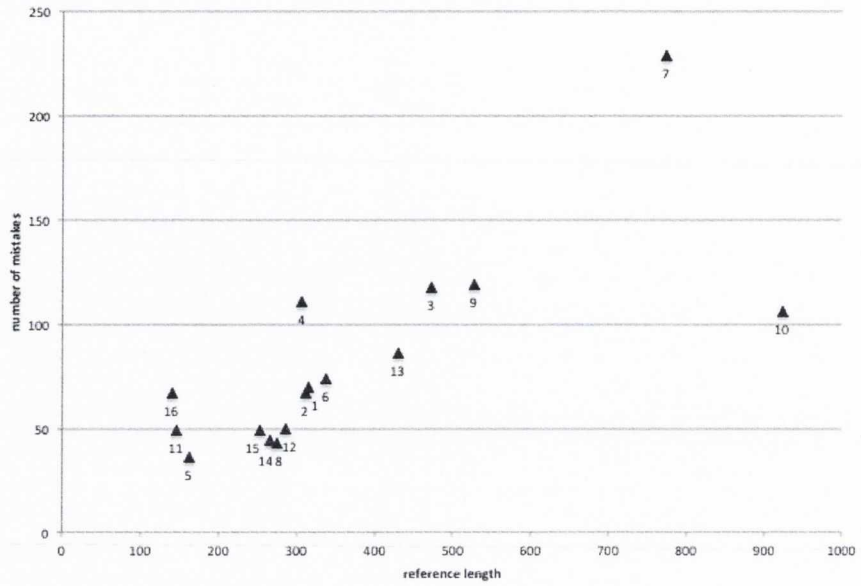


Figure 4.1 – Scatter plot illustrating the relation between reference lenght and number of mistakes for the participants.

hitting the return button. The follower received the chat messages on his or her screen and was allowed to talk freely. The instructor could hear the follower's utterances. Each participant was the instructor in two of the experiment sessions, once in ASR mode, once in telephone mode, and the follower in the other two sessions, also once in ASR mode and once in telephone mode, resulting in four single sessions in which the pair of participants was asked to perform the map task during the course of the experiment. In consideration of map specific bias and order effects the sequence of maps and mode was evenly shuffled. An overview of instructor, map and mode for each of the four parts in the eight experiment sessions can be found in Table 4.4.

Several interaction parameters were logged and analysed for each of the map task sessions. The ASR output was protocoled with a timestamp. The audio recordings were fully transcribed. As seen in the previous section, the transcripts of the instructor's utterances in the ASR mode were used as reference to calculate the word error rate for the intrinsic evaluation. Time logs were gathered from the recordings and logged interaction protocols. On basis of the full transcriptions, the resulting routes on the follower's maps, and the logged interaction data, measures like word count, time to complete task and the performance on the route description were gathered.

Results: There are three factors that will be used in the analysis in order to perform the extrinsic evaluation: One is the performance of the pair to draw the route as accurately as possible, another is the time that was needed to complete the task and the third is the number of

		scenario nr.			
		1	2	3	4
1	instructor	part02	part02	part01	part01
	map	forest	wild west	office	river
	mode	ASR	phone	ASR	phone
2	instructor	part04	part04	part03	part03
	map	forest	wild west	office	river
	mode	ASR	phone	ASR	phone
3	instructor	part 06	part06	part05	part05
	map	wild west	forest	river	office
	mode	phone	ASR	phone	ASR
4	instructor	part08	part08	part07	part07
	map	wild west	forest	river	office
	mode	phone	ASR	phone	ASR
5	instructor	part10	part10	part09	part09
	map	office	river	forest	wild west
	mode	phone	ASR	phone	ASR
6	instructor	part12	part12	part11	part11
	map	office	river	forest	wild west
	mode	phone	ASR	phone	ASR
7	instructor	part 14	part14	part13	part13
	map	river	office	wild west	forest
	mode	ASR	phone	ASR	phone
8	instructor	part16	part16	part15	part15
	map	river	office	wild west	forest
	mode	ASR	phone	ASR	phone

Table 4.4 – Overview of eight experiments (top to bottom) specifying for each of the four parts in the single session the instructor's participant number, the map that was discussed, and the mode that was used.

words uttered in the interaction. The main measure for task performance in map task experiments is how much the route that has been drawn by the follower, deviates from the original route on the instructor's map. One recommendation can be found on the HCRC map task corpus website,[4] which consists in overlaying the follower's map with an one centimetre grid and calculating the deviation score by measuring the area of deviation between the two routes in square centimetres.

However, this system of calculating the deviation has its drawbacks for several reasons. The main problem is that it calculates how far off the follower was from the original route but it seems more plausible to assess whether the communicative goal was met. Hence it has to be evaluated whether the follower understood which landmarks has to be passed by, and what the correct directions around the landmarks are in order to take the same way from the start to the goal. In some cases the participants followed the perfect way (direction wise), but were simply

[4] http://groups.inf.ed.ac.uk/maptask/

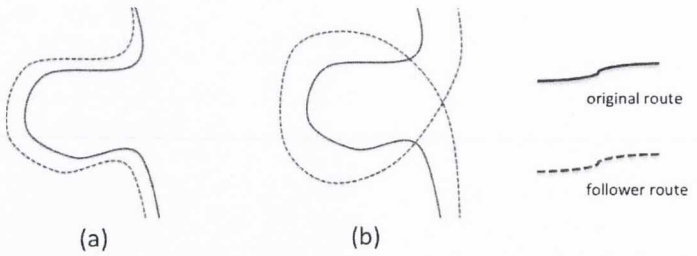


Figure 4.2 – Examples where the proposed calculation of the deviation between follower's route and the original poses a problem.

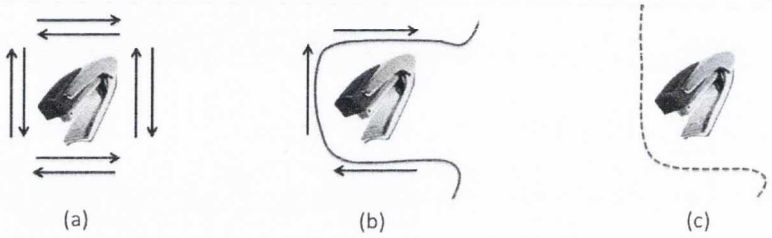


Figure 4.3 – Illustration of possible features to calculate the task performance when considering bearings around a landmark is passed by.

a bit off the original line. Picture (a) in Figure 4.2 demonstrates such a case. This happens especially in cases where the landmark that had to be circled was not in the follower map, and therefore the exact position and size were unknown to the follower. Although participants were aware of a rough position and size, and circled the imagined object most routes deviate in these cases critically. Another difficulty is to calculate the deviation in cases where the follower route loops but the original route does not. An example for such a situation can be found in Picture (b) in Figure 4.2. In such cases it is unclear which area has to be used in order to calculate the deviation.

Because of these obstacles a new, yet simpler, assessment of the maps has been designed and applied which is geared by the landmarks that determine the original route. The method assesses whether the follower approached the relevant landmarks from the correct direction, passed it by on the right side or sides, and left it in the correct direction. To formalise this approach each landmark is associated with eight possible direction vectors in which the route can pass by the landmark. These vectors are illustrated in Picture (a) in Figure 4.3.

For each landmark that is relevant to communicate the original route on the instructor's

map, the direction scores are determined (cf. Figure 4.3 Picture (b)). Each direction vector is associated with a score of one which is summed up for all landmarks in a instructor map in order to determine the gold standard route description. An overview of all direction scores that have been determined for the four maps that were used in the experiment can be found in the appendix (*Materials of the ASR experiments*). Later in the assessment, the correct direction scores for the follower's route are calculated. In case of the example in Picture (c) in Figure 4.3 this would mean that only two direction vectors from the original route in Picture (b) are matched. Therefore, the route would only gain two scores. Furthermore, deviations from the route have to be accounted for. This entails to penalise participants if they navigated around wrong landmarks due to communication failure. In cases where the participant navigated around such wrong landmarks, each direction vector is assigned a penalty score. To assess task performance these mis-navigation scores, or penalties, are subtracted from the achieved navigation scores.

For each of the four instructor maps the direction scores have been assessed. An overview of the direction scores per map can be found in Appendix B. Overall there were 19 scores to be achieved in the office scenario, 18 in the forest scenario, 28 in the wild west scenario, and 21 in the river scenario.

		1	2	3	4	5	6	7	8
office	mode	ASR				phone			
(19 direction scores)	matched features	19	13	19	17	18	16	16	8
	penalty	0	6	0	5	0	2	4	10
	sum	19	7	19	12	18	14	12	-2
forest	mode	ASR				phone			
(18 direction scores)	matched features	10	16	13	13	18	18	18	14
	penalty	8	4	1	3	0	0	0	4
	sum	2	12	12	10	18	18	18	10
wild west	mode	phone				ASR			
(28 direction scores)	matched features	27	28	28	12	28	27	27	22
	penalty	2	2	0	11	0	0	1	3
	sum	25	26	28	1	28	27	26	19
river	mode	phone				ASR			
(21 direction scores)	matched features	21	21	21	21	17	18	13	13
	penalty	0	0	0	0	4	0	0	5
	sum	21	21	21	21	13	18	13	8

Table 4.5 – Overview of the achieved direction score and scored penalty points for each of the eight sessions ordered by the maps. The last row for each scenario (sum) subtracts the penalty points from the achieved direction scores to indicate the overall performance in the single session.

Table 4.5 shows an overview of the direction scores that were gained in each of the eight experiment sessions, as well as the penalty scores assigned for wrong turns in the routes. The overview demonstrates that the best performance was achieved in the telephone mode of the

river scenario where all participants gathered full number of direction scores and were not penalised with a single penalty score. The table also exposes that the pair in session eight performed especially badly when it came to drawing the routes. Table 4.5 is a compressed version of Table B.1, B.2, B.3 and B.4, in Appendix B which enlist the exact number of achieved direction scores for each participant and map.

In Table 4.6 the direction scores are further compressed by averaging over the sessions for each mode, in order to enable the comparison of the performance in the two modes for the four different maps. The table shows that participants in the river and forest map gained less correct direction scores and more penalty scores in the ASR mode, where more miss-communication can be expected. This cannot be reflected in the results of the wild west and office map however. Here the route drawing performance went smoother in the ASR mode than in the telephone mode.

		ASR	phone
office	matched features	17 (89.5%)	14.5 (76.3%)
	penalty	2.75	4
	sum	14.25 (76.3%)	10.5 (55.3%)
forest	matched features	13 (72.2%)	17 (89.5%)
	penalty	4	1
	sum	9 (50%)	16 (88.9%)
wild west	matched features	26 (92.9%)	23.75 (84.8%)
	penalty	1	3.75
	sum	25 (89.3%)	20 (71.4%)
river	matched features	15.25 (72.6%)	21 (100%)
	penalty	2.25	0
	sum	13 (61.9%)	21 (100%)

Table 4.6 – Overview of the direction score calculations for each mode averaged over the eight sessions.

The performance on drawing the routes can be also displayed as percentage of how many of the gold standard direction scores were achieved. This is illustrated in Table 4.7 which shows for each session and each map the percentage of achieved direction scores, as well as the percentage of achieved direction scores from which the penalty scores have been subtracted.

Given these percentages a direction score percentage to each session in the ASR mode can be assigned and these can be compared with the calculated word error rate for the same interaction. A comparison of these two data sets can be found in Table 4.8. On basis of this two data sets the *Pearson correlation coefficient* was calculated [Rodgers and Nicewander, 1988] which can range from -1 to 1. A value of 1 implies that the relationship between one dataset (say X) and the other (say Y) can be described as perfectly linear. That is, all data points can be drawn on a line for which Y increases as X increases. If all data points lie on

		1	2	3	4	5	6	7	8
	mode	ASR				phone			
office	matched features	100	68.4	100	89.5	94.7	84.2	84.2	42.1
	sum	100	36.8	100	63.2	94.7	73.7	63.2	-10.5
forest	matched features	55.6	88.9	72.2	72.23	100	100	100	77.8
	sum	11.1	66.7	66.7	55.6	100	100	100	55.6
	mode	phone				ASR			
wild west	matched features	96.4	100	100	42.9	100	96.4	96.4	78.6
	sum	89.3	92.9	100	3.6	100	96.4	92.9	67.9
river	matched features	100	100	100	100	80.9	85.7	61.9	61.9
	sum	100	100	100	100	61.9	85.7	61.9	38.1

Table 4.7 – Overview of the percentage of achieved direction score for each of the eight sessions (ordered by maps).

a line for which Y decreases as X increases than the coefficient takes on the value of -1. A value of 0 implies no linear correlation between the variables. It is dependent on the data set whether an r-value indicates a low or a high correlation. In the case of this experiment an r-value of lower then 0.3 is considered to indicate no correlation and if the r-value is higher than 0.8 it is considered to imply a strong correlation. The sample Pearson correlation coefficient for word error rate and achieved direction scores minus the penalty scores (cf. Table 4.8) is $r = -0.15$ ($p - value = 0.57$) which implies no linear correlation between word error rate and the performance on drawing the routes.

	Participant	WER	direction score (in %)	direction score minus penalty (in %)
ASR1	1	0.2222	100	100
ASR2	2	0.2154	55.6	11.1
ASR3	3	0.2500	68.4	36.8
ASR4	4	0.3616	88.9	66.7
ASR5	5	0.2209	100	100
ASR6	6	0.2190	72.2	66.7
ASR7	7	0.3868	89.5	63.2
ASR8	8	0.1563	72.2	55.6
ASR9	9	0.2258	100	100
ASR10	10	0.1148	80.9	61.9
ASR11	11	0.3356	96.4	96.4
ASR12	12	0.1748	85.7	85.7
ASR13	13	0.2005	96.4	92.9
ASR14	14	0.1660	61.9	61.9
ASR15	15	0.1944	78.6	67.9
ASR16	16	0.4786	61.9	38.1

Table 4.8 – Comparison of calculated word error rate and the percentage of achieved direction score for each of the sixteen ASR sessions.

Another measure that was gained from the data logs is the time to complete the task. This

is defined as the time from the first utterance of the instructor, with which the route description was initiated, to the last word in the utterance in which the follower confirmed that he had reached the goal. An overview of the task completion times in seconds can be found in Table 4.9. Comparing the time to task completion with the achieved direction scores, it can be shown that both measures do not correlate (sample Pearson correlation coefficient of $r = 0,05$, $p - value = 0.86$).

	1	2	3	4	sum	5	6	7	8	sum
mode			ASR					phone		
office	417	1085	318	1651	3471	340	340	245	146	1071
forest	379	747	825	592	2543	854	172	267	142	1435
sum	796	1832	1143	2243	6014	1194	512	512	288	2506
mode			phone					ASR		
wild west	257	327	225	314	1123	1453	550	695	432	3130
river	377	255	120	505	1257	1308	566	413	326	2613
sum	634	582	345	730	2291	2761	1116	1108	758	5743
sum overall	1430	2414	1488	2973	-	3955	1628	1620	1046	-

Table 4.9 – Overview of the task completion times in seconds for each of the eight sessions ordered by maps.

This data can be compressed by averaging the time to task completion over all ASR and telephone sessions for each map. Table 4.10 combines these results. It can be seen that the time to complete the task is considerably longer in the ASR mode than in the telephone mode. On average the instructor needed 735 seconds (12:15 minutes) in the ASR mode, and 305 seconds (5:05 minutes) in the telephone mode. It is no surprise that times are much higher in ASR mode than in telephone mode since in addition to the time that the instructor needed to utter the words, the time to recognise the utterances, and the time it takes the follower to read the utterances have to be added. It can also be expected that in cases were the WER was high, and therefore the recognition was far from what was originally uttered, the follower needed extra time to understand the probable content of the message.

	ASR		phone	
	sum	avg	sum	avg
office	3471	867.75	1071	267.75
forest	2543	635.75	1435	358.75
wild west	3130	782.5	1123	280.75
river	2613	653.25	1257	314.25
all maps	11757	734.81	4886	305.38

Table 4.10 – Overview of the task completion times for each of the eight sessions (ordered by maps).

The comparison between time to task completion for the interactions in the ASR mode and

the word error rate can be found in Table 4.11 and in the plot of the WER results ofer time in Figure 4.4. In theory, in cases where the word error rate was especially low it is to be expected that this has a negative influence on the time to complete the task. However looking at the data it can be seen that the three fastest instructors were participant 5, 16, and 2 and the slowest were participant 10, 9, and 7. Looking at the word error rate, the three lowest WER results were achieved by participants 10, 3, and 8, and the highest WER results can be found with participants 7, 12, and 4. Participant 16 has the highest WER but is the fastest and participant 10 has the lowest WER but takes the most time to complete the task contradict this hypothesis, however. Furthermore, no linear correlation between the time to complete the task and the word error rate could be confirmed with a sample Pearson correlation coefficient of $r = -0.04$ ($p - value = 0.88$).

	Participant	WER	time
ASR1	1	0.2222	417
ASR2	2	0.2154	379
ASR3	3	0.2500	1085
ASR4	4	0.3616	747
ASR5	5	0.2209	318
ASR6	6	0.2189	825
ASR7	7	0.3868	1651
ASR8	8	0.1564	592
ASR9	9	0.2258	1453
ASR10	10	0.1148	1308
ASR11	11	0.3356	550
ASR12	12	0.1748	566
ASR13	13	0.2005	695
ASR14	14	0.1660	413
ASR15	15	0.1944	432
ASR16	16	0.4786	326

Table 4.11 – Comparison of calculated word error rate and the time to complete task for the ASR sessions.

From the transcripts of the interactions the number of words uttered by the instructor and the follower were gained. Furthermore the number of words uttered for each interaction was calculated. Table 4.12 which gives an overview of the number of words used by instructor and follower in each session, shows for example that pair number five was much more verbose than pair number eight. A comparison of the times to task completion with the number of words uttered by the instructor illustrates a dependency between the two data sets. Pair eight was overall the quickest and pair five the slowest to describe the four routes. With a sample Pearson correlation coefficient of $r = 0.86$ ($p - value = 1.8e^{-5}$) the time to task completion and the number of words uttered in the interaction overall (by instructor and follower) show a

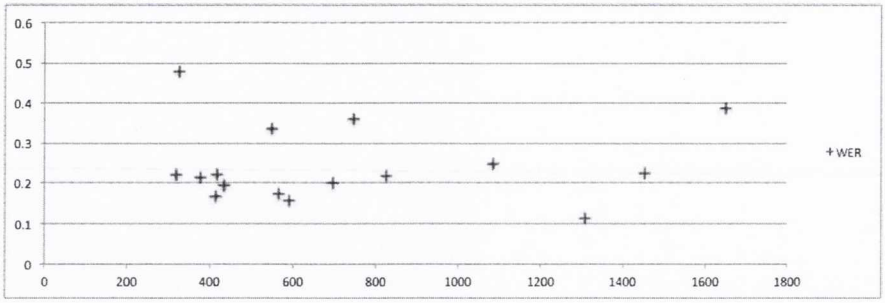


Figure 4.4 – Plot of WER results over time.

high linear correlation. The number of words that was needed to complete the task indicates efficiency and also how smoothly the interaction went. A high verbosity of the instructor in the ASR mode compared to the telephone mode can be expected to indicate that the interaction did not go smoothly, forcing the instructor to repeat words and phrases or to describe words which were misrecognized. In general it can be anticipated that the instructor and follower in the telephone mode are more verbose, speaking more freely not weighing into every word. This behaviour is expected to change in the ASR mode, where the instructor is expected to choose words more carefully and to speak slower, while the follower probably will try to compensate bad ASR output by posing more clarification questions, or even take over the initiative by asking more questions about possible further steps. Therefore, cases where the ratio between instructor and follower are high are hypothesised to indicate smoother interactions, hence in such cases the WER should be low.

There are big differences between the pairs and maps when the ratio between follower and instructor words is examined, but Table 4.13 reveals that the ratio in the ASR mode is below one in most of the cases and in the telephone mode it is usually well above one. This confirms that the participants change their strategy when discussing the routes dependent on the mode. In the ASR mode it is the follower who tries to lead the conversation, whereas in the telephone mode the instructor is more verbose.

Analysing the number of words by the different roles and in sum with the word error rate (see Table 4.14 for a comparison) no linear correlation can be established between WER and the words uttered ($r = -0.04$, $p - value = 0.72$). However, the sample Pearson correlation coefficient for the word error rate and the word ratio implies with $r = -0.44$ ($p - value = 0.57$) a medium linear correlation.

Examining the correlation, in summary it can be seen that the correlation between each of the extrinsic measurements with the intrinsic word error rate has been calculated and it has

session	role	office	forest	wild west	river	sum
1	instructor	314 (ASR)	310 (ASR)	486	726	1835
	follower	374 (ASR)	315 (ASR)	96	231	1016
2	instructor	440 (ASR)	302 (ASR)	598	486	1829
	follower	623 (ASR)	700 (ASR)	378	167	1868
3	instructor	152 (ASR)	337 (ASR)	302	246	1043
	follower	298 (ASR)	763 (ASR)	158	59	1278
4	instructor	774 (ASR)	287 (ASR)	585	781	2427
	follower	941 (ASR)	254 (ASR)	143	421	1759
5	instructor	681	1080	524 (ASR)	917 (ASR)	3202
	follower	98	329	1786 (ASR)	498 (ASR)	2711
6	instructor	499	235	284 (ASR)	148 (ASR)	1166
	follower	200	106	589 (ASR)	278 (ASR)	1173
7	instructor	434	452	421 (ASR)	265 (ASR)	1572
	follower	252	230	619 (ASR)	212 (ASR)	1313
8	instructor	183	235	250 (ASR)	140 (ASR)	808
	follower	89	80	333 (ASR)	197 (ASR)	699

Table 4.12 – Overview of the number of words uttered by the instructor and follower (ordered by maps).

session	ASR	ASR	ASR overall	phone	phone	phone overall	overall
1	0.84	0.98	**0.91**	5.06	3.14	**4.10**	2.51
2	0.70	0.44	**0.57**	1.58	2.91	**2.25**	1.41
3	0.52	0.44	**0.48**	1.91	4.17	**3.04**	1.76
4	0.82	1.13	**0.97**	4.09	1.86	**2.98**	1.98
5	0.30	1.84	**1.07**	6.95	3.28	**5.12**	3.09
6	0.48	0.53	**0.51**	2.50	2.22	**2.36**	1.44
7	0.68	1.25	**0.97**	1.72	1.97	**1.85**	1.41
8	0.75	0.71	**0.73**	2.06	2.94	**2.50**	1.62

Table 4.13 – Overview of the ratio between the number of words uttered by the instructor and number of words uttered by follower.

been shown that there is no, or only small linear correlation. The only exception is the correlation between WER and the ratio between the words uttered by the instructor and the words uttered by the follower which is medium. A deeper inspection of the reference transcripts and the ASR output was performed to gain a better insight in the results. A complete overview of the most frequent mistakes can be found in B.8 in Appendix B. In Table 4.15 a selection of the most frequent mistakes can be found. Looking at the most frequent mistakes it can be observed that compounding mistakes occur quite regularly. Compounds are words that contain more than one stem (e.g. darkroom, bittersweet, doghouse) and are a frequent phenomenon in the German language. However, the composition of stems to form new words is not unrestricted. There are many compositions that will be rejected by native speakers. Such compositions can be found in the list of errors that were output by the ASR system. For example participant 16

	Participant	WER	# words instructor	# words follower	# words both roles	ratio between instructor and follower
ASR1	1	0.2222	314	374	688	1.84
ASR2	2	0.2154	310	315	625	0.7
ASR3	3	0.2500	440	623	1063	1.13
ASR4	4	0.3616	302	700	1002	1.25
ASR5	5	0.2209	152	298	450	0.98
ASR6	6	0.2189	337	763	1100	0.48
ASR7	7	0.3868	774	941	1715	0.68
ASR8	8	0.1564	287	254	541	0.75
ASR9	9	0.2258	524	1786	2310	0.44
ASR10	10	0.1148	917	498	1415	0.3
ASR11	11	0.3356	287	589	876	0.84
ASR12	12	0.1748	148	278	426	0.52
ASR13	13	0.2005	421	619	1040	0.44
ASR14	14	0.1660	265	212	477	0.53
ASR15	15	0.1944	250	333	583	0.82
ASR16	16	0.4786	140	197	337	0.71

Table 4.14 – Comparison of calculated word error rate and the number of words uttered by the instructor, the follower, and both together, as well as the ratio of words between instructor and follower for each of the sixteen ASR sessions.

told the follower to walk into the *direction* of the *cat* (*'richtung katze'*) and the ASR system compounded direction and cat to a new word (*'richtungskatze'*) which is not a proper German compound. Another very frequent type of error was the decomposition of German compounds into their stems. For example the German word for tepee (*'Indianerzelt'*) which can be literally translated to an 'indian's tent' was always split into the two words *'indian'* and *'tent'* (*'Indianer'* and *'Zelt'*). Furthermore, misrecognition often resulted in words that only differ by a small number of characters from the original uttered word (e.g. *'nordlicht'* and *'nördlich'*).

Compound mistakes have a big influence on WER because two words in the reference will be counted as one substitution and one deletion, resulting in a WER of a hundred per cent. Decomposition mistakes account for even 200 per cent, since one word in the reference is accounted as one substitution and one insertion. However, for humans those mistakes often do not influence the comprehensibility, in some cases they might not even be recognised. These observations indicate that an accuracy measure for ASR systems based on words might not be fine grained enough and therefore further error rates to assess the ASR accuracy intrinsically were considered which are based on smaller units than a word.

The next smallest unit after words are syllables. Therefore, a variation of WER was implemented which calculates the number of misrecognised syllables divided by the number of syllables in the reference, the *syllable error rate* (SylER).[5] The implementation segments the

[5]The implementation of the SylER algorithm as well as the algorithms for CER, CER2, SoundER, SoundER2

reference word	ASR output
Compound mistakes	
Haufen Kot	haufenkot (7)
geschlossenen Koffer	Geschlossenenkoffer (5)
Richtung Saloon	Richtungssalon
Richtung Katze	Richtungskatze
Richtung Hühner	Richtungshühner
Richtung Kaktus	Richtungskaktus
Decomposition mistakes	
Indianerzelt	Indianer Zelt (7)
Linksbogen	links Bogen (3)
Vorderbeinen	vorder Beinen (2)
Other mistakes	
Tacker	Acker (8), Tag Herr (3), Packer (3), Targa (3), Tage (2), Tag der (2), Hacker (2), AK (2), Tag, Tag war, Shaker, Trucker, starker, intakter, Tanker
Forstfahrzeug	vors Fahrzeug (3), befasst Fahrzeug, Forst vorzeigt
nordöstlich	nordwestlich (8)
Start	Staat (4)

Table 4.15 – Selection of the most frequent ASR mistakes (numbers in brackets indicate the number of occurances of the mistake).

reference transcriptions and ASR output texts into syllables, using the `HyFo` library,[6] which utilises the same algorithm as the typesetting system TEX [Liang, 1983], and a German specific list of known segmentations. The syllables of the segmented words were treated as a continuous stream (i.e. word segmentation was ignored). SylER was intended as a linguistically plausible upgrade to WER, motivated by the human ability to use contextual clues to process and correct mistakes in language processing. The results of the SylER for the sixteen dialogues can be found in Table 4.16.

The *character error rate* (CER) is often used for languages like Japanese and Chinese which do not have word boundaries. CER differs from WER by counting misrecognised characters instead of words. To assess the influence of segmentation errors such as compound or decomposition mistakes two variants were implemented, one that scored whitespace errors (CER) and one that ignored them (CER2). The latter version results in lower error rates and it is assumed to be a better model for human language processing, since humans are still able to understand texts that lack segmentation. An overview of the CER and CER2 results can be found in Table 4.16.

Another approach that was implemented is to match homophones which show minor dif-

which will be described in the following paragraphs were provided by my colleague from JULIE Lab Jena, Johanne Hellrich, in the course of a collaborative publication effort on the topic of alternatives to WER.

[6] `http://defoe.sourceforge.net/hyfo/hyfo.html` (last accessed 21/08/2012)

participant	WER	SylER	CER	CER2	SoundER	SoundER2
ASR01	0.222	0.158	0.090	0.089	0.197	0.060
ASR02	0.215	0.192	0.121	0.130	0.196	0.110
ASR03	0.250	0.182	0.064	0.046	0.237	0.033
ASR04	0.362	0.304	0.171	0.179	0.339	0.161
ASR05	0.221	0.212	0.116	0.129	0.209	0.115
ASR06	0.219	0.165	0.100	0.104	0.180	0.077
ASR07	0.387	0.293	0.175	0.177	0.364	0.139
ASR08	0.156	0.117	0.066	0.066	0.149	0.052
ASR09	0.226	0.169	0.085	0.090	0.197	0.072
ASR10	0.115	0.080	0.041	0.042	0.106	0.034
ASR11	0.336	0.206	0.096	0.095	0.329	0.072
ASR12	0.175	0.127	0.064	0.066	0.157	0.050
ASR13	0.200	0.163	0.093	0.095	0.175	0.077
ASR14	0.166	0.152	0.088	0.094	0.155	0.083
ASR15	0.194	0.157	0.082	0.084	0.171	0.071
ASR16	0.479	0.456	0.387	0.386	0.464	0.340

Table 4.16 – Overview of the intrinsic evaluation results of the ASR study. CER and SoundER have variants respecting word boundaries (1) and variants ignoring those (2).

ferences in spelling. A German version of the *Soundex* algorithm[7] was utilized which was initially developed to index names by sound, by encoding homophones to the same representation [Stanier, 1990], the so called *Kölner Phonetik* [Postel, 1969]. The implementation of the *sound error rate* is based on an encoding scheme which maps letter combinations representing similar sounds onto the same numerical code, mostly ignoring vowels and repeated letters. Despite being optimised for German the "Kölner Phonetik" offers only limited support for some elements of German phonetics, e.g. it can not handle umlauts. The input was encoded by using the *ColognePhonetic* encoder[8] in the *Apache Commons*. Again two variants were implemented: One comparing the codes for whole words (SoundER) and one treating the encoded input as a continuous string of numbers, calculating the edit distance for swapped numbers (soundER2). The second version is believed to match human comprehension, since word segmentation is not an essential necessity for comprehension. An overview of SoundER and SoundER2 results can be found in Table 4.16.

The resulting five sets of error rate results (i.e. WER, SylER, CER, CER2, SoundER and SoundER2) have been used to compare the correlation with the extrinsic measures. The Pearson correlation coefficients can be found in Table 4.17.

The word error rate has the highest values in the set of the six error rates under inspection. In theory CER could take on higher values than WER if the systems performance deteriorates

[7]The idea was patented in 1918 and 1920 already by Robert C. Russell (Patent US1261167, and US1 435 663).

[8]`http://commons.apache.org/codec/apidocs/org/apache/commons/codec/language/ColognePhonetic.html` (last accessed 21/08/2012)

error rate	time to task completion	direction scores	ratio instructor follower words
WER	0.04 (0.88)	-0.15 (0.57)	-0.44 (0.57)
SylER	-0.10 (0.72)	-0.26 (0.33)	-0.36 (0.17)
CER	-0.18 (0,49)	-0.31 (0.24)	-0.21 (0.42)
CER2	-0.20 (0.45)	-0.29 (0.28)	-0.22 (0.41)
SoundER	0.02 (0.94)	-0.17 (0.53)	-0.39 (0.13)
SoundER2	-0.23 (0.43)	-0.31 (0.26)	-0.17 (0.21)

Table 4.17 – Overview of the *Pearson correlation coefficient* (p-values in brackets) for the intrinsic evaluation methods (WER, CER, CER2, SylER and SoundER, SoundER2) with the extrinsic evaluation results.

for long words in short texts. For example when the ASR system misses out on words (e.g. reference transcription *'he stumbled'*, ASR output *'he* ␣*'* would yield a CER of 0.77 or CER2 of 0.83, but a WER of 0.5). The results of the SoundER are equal to results of WER, while SoundER2 resembles results of CER2. Values for SoundER variants are lower than those for CER variants and SoundER2 is lower than SoundER. This can be demonstrated by the example of the German word *scharf* ('spicy') which is encoded as *873* and the similar sounding word *Schaf* ('sheep') which is encoded as *83*, resulting in a SoundER of 1, whereas the SoundER2 is only 0.5.

4.2.3 Discussion and Conclusion

The comparison between extrinsic and intrinsic ASR evaluation included the application of the map task method which was used to collect interaction data. This was the basis for the extrinsic measures and the calculation of the word error rate on the same material collected from 16 participants. Three interaction factors have been analysed to perform the extrinsic analysis: the performance on drawing the route, the number of words uttered by the different roles, as well as the time to complete the task. The direction score measures how well the participants performed on the task to draw the routes. The time to task completion is an indicator for how efficient the interaction went. The word ratio is believed to be an indicator for how well the recogniser performed since participants would change their strategy in cases with bad recognition performance to a scheme where the initiative would be taken over by the follower, trying to confirm everything that was said by the instructor, or guessing what next steps could be. It has been shown that the extrinsic measures correlate with each other where a correlation is plausible. The hypothesis of the experiment was that the intrinsic word error rate measure which is the state of the art evaluation method for ASR evaluation, would reflect the results from the extrinsic measures. Comparing the extrinsic measures with the word error rate analysis, however, no or only a minor linear correlation could be demonstrated. A qualitative analysis of

the mistakes that could be found in the ASR transcripts showed that many errors only involve single characters per word. Hence, the assumption was made that an intrinsic evaluation on word level might not be fine grained enough to reflect the ASR accuracy.

The human ability to process languages is still unmatched by technology. Human conversation is full of errors, ill-formed sentences, false starts and hesitations. These complicate matters for NLP components. However, if humans are faced with small mistakes they are well able to filter the relevant content out, sometimes even without realising that there were mistakes. This has been confirmed by the experiment several times. For example all instructors that had to explain the office route in ASR mode tripped over the misrecognition of the German word for stapler. Not in a single case did the ASR system manage to recognise the word for stapler correctly. However, looking at the direction scores for this landmark (Appendix B) it can be observed that all participants perfectly managed to circumvent the stapler in the interaction. In all cases they found a way to refer to the object without using the word for stapler.

Also interesting is the fact that it was expected that the participants in the telephone mode would perform better than in the ASR mode, describing the route more accurately. This cannot be confirmed for the wild west and the office scenario where the direction scores on average are lower, and the penalty scores higher than in the ASR mode. Even if the session eight which performed especially badly in the evaluation, is excluded from the calculation, then the performance of ASR and telephone mode are about the same as for the office scenario. The same effect cannot be achieved for the wild west map, since the pair in session eight performed the wild west scenario in ASR mode and the exclusion of the results would only raise the performance even further.

The way the word error rate is calculated allows for no small mistakes and therefore does not take into account the ability of a human listener to compensate small mistakes. In order to perform a more fine grained evaluation which accommodates its score to the human ability to understand error prone input by using context, five further error rates were implemented and the material of the sixteen ASR interactions was assessed with these intrinsic methods.

The results show that WER has the second lowest correlation value with the direction scores and the lowest with time to task completion. For the word ratio, however, the correlation is the highest among the proposed error rates. The SoundER algorithm shows a similar picture, with no correlation with the direction scores and time to complete task measures, and a medium correlation with the word ratio. SoundER2 on the other hand has the highest correlation value for the direction score together with CER and the second highest with time to complete task.

In summary it can be said that the correlation between extrinsic and intrinsic evaluation paints a two-sided picture. On the one hand the extrinsic ratio between instructor and follower

words shows a medium correlation with the established intrinsic WER score, but only a small correlation (or no correlation at all) with the time to task completion and the direction score. On the other hand the new measurements, especially in the variants disrespecting boundaries, show a smaller correlation with the ratio, but achieve (borderline) medium correlation with the direction score. Only the new intrinsic measurements, especially the SoundER variant without boundaries, correlate with the time to task, i.e. show if the systems quality has influence on the time needed to complete the task. However, the correlation between the extrinsic methods and the alternative intrinsic methods still only show a small to medium correlation.

4.3 Machine Translation

In the last two decades machine translation technology has reached a level of quality which warrants its increasing use in real-world applications. A significant trend in this direction has been the use of machine translation (MT) in combination with speech technologies. However, machine translated content is still far from perfect. For MT components to be effectively deployed, it is important for developers to be able to reliably assess the quality of the translation output. This task is difficult for a number of reasons as was discussed in Chapter 2. MT has traditionally been evaluated, through comparable intrinsic evaluation metrics which aim to correlate as closely as possible to human judgement performed on a task independent (and often sentence-by-sentence) basis. Although most MT evaluation is still intrinsic, there have been renewed calls for more attention to be paid to extrinsic evaluation in MT [Goldstein et al., 2005].

This section reports on a study to examine MT output intrinsically and extrinsically, for use in a dialogue system, to aid applications using MT combined with speech technologies. For the intrinsic evaluation, in which the quality of the translated sentences was assessed in isolation, three human judges were asked to indicate which translation they preferred. An annotation scheme was developed to investigate the reasons behind the preferences expressed by the judges. For the extrinsic evaluation the translations were incorporated into a Wizard of Oz (WOZ) system, and an experiment was run with eight German subjects who interacted with the system. Results of these interactions were gathered using questionnaires, interviews and system logs, as well as through an analysis of the small corpus of dialogues collected with the experiment. This section is largely based on [Schneider et al., 2010].[9]

Preparations: The two-fold experiment is based on a scenario in which German speakers

[9]The paper is based on the WOZ experiment that will be described in the following section. I planned and conducted the experiment under the supervision of Saturnino Luz. My colleague Ielka von der Sluis was involved in the planning process. The first draft of the paper was written by myself and then refined with the help of Saturnino and Ielka.

have to find a good offer on Internet connections in Ireland. The knowledge acquisition for this system was done through inspection of the web sites of Irish Internet providers. To collect the necessary set of English system output utterances a simple prototyping method was used. Human-computer dialogues were simulated using a regular chat tool which proved to be a simple yet adequate prototyping approach to test the set of system utterances for completeness. Five participants were asked to use the chat tool to find out about good offers on Internet connections in Ireland. The experimenter 'helped' them using a fixed set of predefined system utterances. After each dialogue the set of utterances was adjusted depending on what appeared to be missing. The mean dialogue length was around twenty turns, but the number of words within the conversations ranged from less than 200 to more than 500. One of these more complex dialogues was caused by a person 'playing around with the system' to figure out its limitations. Another dialogue ended in a breakdown-like situation which showed that the set of utterances was not sufficient and new utterances had to be created. The final set of system outputs consisted of 32 utterances (ten of which had open slots) and a list of slot fillers. The 32 English utterances consisted of one welcome message, three utterances that signal a problem in the interaction and suggest how to proceed, six explicit feedbacks to user input, four information requests, fifteen utterances that provide information on a particular option, one stalling, one break off, and one goodbye message.

A native German speaker firstly translated this set of English utterances to German. This human translation will be from now on called the reference translation. Secondly, the English source utterances were translated using the online machine translation service of *Systran*[10] and *Google*.[11] A complete list of all four sets of utterances can be found in the Appendix C.[12]

The hypothesis is that the translation with the best ratings in the intrinsic evaluation would produce the smoothest interaction in the extrinsic evaluation.

4.3.1 Intrinsic Evaluation

The intrinsic evaluation, in which the quality of the translated sentences was assessed in isolation, was performed in two steps. First three human judges were asked to indicate sentence wise which of the two automatic translations they preferred. To gain a better insight into this preferences an annotation scheme was developed to investigate the reasons behind the decisions of the judges. Intrinsic MT evaluation typically employs automatic metrics like BLEU, TER or METEOR. These quality scores assume large corpora of text in order to correlate with

[10] `http://www.systranet.com/` (last accessed 21/08/2012)

[11] `http://translate.google.com/` (last accessed 21/08/2012)

[12] All automatic translations have been produced on November 7^{th} 2009. *Google's* translation service is subject to constant change, therefore it is possible that reported translation errors do not appear at a later point in time.

human judgements. Further, there is evidence that statistical MT systems receive higher BLEU scores than their counterparts which do not employ n-gram based approaches and hence the use of BLEU to compare systems with different approaches to machine translation is inappropriate [Callison-Burch et al., 2006]. Therefore, it is not meaningful in the context of this experiment to calculate one of these scores for the small set of dialogue utterances that were used in this study. However, some intrinsic evaluations that are related to the MT experiment have been reported by Turian et al. [Turian et al., 2003] who asked human judges for ratings of MT output before the development of automatic evaluation methods.

Materials: The material that was used for evaluation consisted of four sets of 28 system utterances that have been described in the previous section, namely, the English original, the German reference translation, the *Systran* translation and the *Google* translation. It is important to note that the list of slot fillers and the utterances that rendered identical translations by *Google* and *Systran* were excluded.

Subjects: The three judges were native German speakers with language experience in English speaking countries. None of the judges was involved in the collection of the utterances or saw them before the evaluation.

Procedure: As already discussed, existing MT evaluation methods are not appropriate for small sets of dialogue utterances that considerably vary in their syntax and pragmatics and that are meant to be used in a particular context and should also not be used to compare systems with different approaches to machine translation. Instead, the intrinsic evaluation was designed in a way that preserves some of the characteristics of the automatic metrics described in Section 2.5.2, but relies on human judges. In order to get a first impression the judges were asked to indicate their preference for one of the two automatic translations on a sentence level. Therefore, for each of the 28 system utterances the judges were shown the English original and the two machine translations. In order to exclude any bias on the judges sides the system that produced the translation was disguised to the judges and the order of the two machine translated sentences was randomized. The judges were asked to identify which of the two translations they preferred and to point out the mistakes in the less preferred translation which induced their judgement.

Results: An overview of the preference rating (in numbers) shows that in almost two thirds of the cases, the *Google* translation was preferred over the *Systran* translation (cf. Table 4.18). The agreement between the judges has been calculated using *Cohen's Kappa score* [Artstein and Poesio, 2008]. They results show a high agreement between the judges and are presented in Table 4.19.

Further Analysis: To understand the reasons for the preferences of the judges and to gain a

	Google	*Systran*
j1	20	8
j2	18	10
j3	18	10

Table 4.18 – *Google* and *Systran* preferences of judge j1, j2 and j3.

pair	Kappa
j1 + j2	0.6725
j1 + j3	0.8444
j2 + j3	0.8363

Table 4.19 – *Cohen's Kappa score* for three pairs of judges.

better insight in the translation errors a further analysis of the automatic translations against the human reference translation similar to TER was performed. First, the edit distances between each of the two automatic translations and the human reference translation was calculated by summing the number of added words (i.e. words that do not appear in the reference translation) and the number of deleted words (i.e. words that do not appear in the machine translations but that do appear in the reference translation). However, results of this exercise presented at the top of Table 4.20 show no difference between the two machine translations.

An annotation scheme was developed that covers the following translation errors:

- *wrong word order*: the order of words in the translated utterance does not correspond to the order of words in the utterance in the reference;
- *synonym*: noun or verb which is not the same as the word in the reference but which has a similar meaning;
- *wrong word*: word which does not appear in the reference and which is not a synonym;
- *untranslated word*: word from the source utterance which has not been translated.

The same three judges that were asked to give their preference rating were then asked to compare both the *Systran* and *Google* translations with the reference using the annotation scheme.

Results To find out why the judges preferred the *Google* translation over the *Systran* translation, a simple correction of the edit distance was made. In calculating the edit distance, each synonym causes a deletion and an addition while the meaning of the utterance stays the same. Hence, the edit distance calculations were adjusted by taking into account the number of synonyms (i.e. edit distance minus 2 times the number of synonyms). Table 4.20 shows that *Google* and *Systran* perform similarly when taking into account the number of synonyms.

	Google			*Systran*		
add	110			111		
del	116			115		
edit	**226**			**226**		
	A1	A2	A3	A1	A2	A3
syn	25	23	23	27	20	27
adedit	**176**	**180**	**180**	**172**	**186**	**172**
wwo	3	3	3	4	4	4
ww	31	49	21	64	80	21
untr	10	11	11	0	0	0

Table 4.20 – *Google* and *Systran* comparison including *edit* distance, *add*ed words, *del*eted words, adjusted edit distance (*adedit*), *syn*onyms, the number of wrong word orders (*wwo*), wrong words (*ww*) and untranslated words (*untr*) identified by annotators A1, A2 and A3.

A further examination the results of the annotations which can be found in Table 4.20, indicate three observations. First, the use of wrong word orders was similarly distributed in the output of the two translation systems. Second, *Google* leaves some words untranslated, while *Systran* appears to translate everything. And third, compared to the *Google* translation, annotators A1 and A2 found almost double the amount of wrong words in the *Systran* translation.

4.3.2 Extrinsic Evaluation

In Section 2.5.2 it was discussed that machine translation output is traditionally evaluated intrinsically and only a few, isolated evaluation efforts assess the effect of task- and context-dependent variables on the user performance of MT systems. Some of them have been reviewed in the section on MT evaluation in Chapter 2. For the extrinsic evaluation of MT output that will be discussed in this chapter, the automatic translations were incorporated into a Wizard of Oz (WOZ) system, and an experiment was run with eight German subjects who interacted with the system. Wizard of Oz experiments were introduced in Chapter 3 as an early stage prototyping method by which the prospective functionality of a system can be tested before a fully functional system is build [Faulkner, 2000]. In the context of machine translation, studies that use WOZ are rare. In the *Verbmobil* project which aimed at developing a system capable of interpreting dialogues, the method was employed to explore strategies and phenomena in interpretation [Krause, 1997]. Bederson et al. also employ WOZ experiments to explore the idea of an iterative translation process supported by a combination of machine translation with monolingual human speakers [Bederson et al., 2010].

Materials: The Wizard of Oz experiment was conducted with the help of an early version of the *WebWoz* Wizard of Oz prototyping tool [Schlögl et al., 2010b] which allowed for speech input and text output and produced time stamped logs for every system utterance. Text output

was chosen to ensure that the subjects were not influenced by the particular voice of the system (either synthetic or human recordings). The user interface included a 'source button' which, when clicked, lead to the English source of the current system utterance.

Two systems, one with the *Systran* and one with the *Google* translations were implemented. Questionnaires were used to capture demographic data of the participants (cf. Figure C.2 in Appendix C) and to assess the quality of the translations (cf. Figure C.5 in Appendix C). In the latter questionnaires a five-point Likert scale was used, ranging from 1 (strongly agree) to 5 (strongly disagree). Questionnaires, instructions etc. were presented to participants on paper. Interactions with the system and interviews with the participants were audio recorded.

Two scenarios were used in the study:

- Imagine you stay in Dublin for the period of 6 months. You are looking for an Internet connection which you can use at home as well as on campus of your University. The offer should be as cheap as possible.

- Imagine you are looking for an Internet connection at your home which is as fast as possible. You are also looking for the highest possible download allowance. The price of the connection, therefore, does not matter to you.

Subjects: Eight German ERASMUS students participated in the study. They were reasonably fluent in English, had moderate to low computer skills and only little experience with dialogue systems. Their average age was 22, ranging from 20 to 24.

Procedure: After filling out a questionnaire for demographic data, participants were asked to read the introduction to the study which explained that the study was meant to try out two versions of a dialogue system that provides information on Internet offers. Subjects were also told that the dialogue systems had recently been automatically translated from English to German, such that they accept German speech as input and produce German text as output, the latter due to problems with the synthesizer. In addition, a 'source button' was introduced which enables the user to see the source of the translation in the case of problems with the system output. Next, participants were asked to read one of the two scenarios, to find a solution with one of the two systems, and to fill out a questionnaire about the interaction. This sequence was repeated for the remaining scenario and system. Scenarios and systems were equally shuffled between participants. After finishing these tasks, the experimenter explained that the participants had been interacting with a human wizard instead of a system. An interview about the system outputs they had received using the system logs followed.

Results: In the extrinsic evaluation the performance of the two dialogue systems was analysed in terms of efficiency and quality similar to those used by the *Paradies* framework [Walker et al., 2000].

Efficiency: Efficiency results, are presented in Table 4.21. The elapsed time, the time between the first and the last system utterances in the dialogue, favours the *Google* over the *Systran* system, not only when summed over the participants (*Google* 250 seconds per interaction on average (stdev 42.42), *Systran* 285 seconds per interaction on average (stdev 77.72)), but also dependent on whether the *Google* system was used before or after participants interacted with the *Systran* system. The total number of system and user turns is higher in interactions with the *Google* system. With respect to the user turns, this results from the fact that in the cases were the participant started with the *Google* system, the number of user turns is almost doubled.

	Google		
	user turns	system turns	elapsed time
1st	75	79	17:28
2nd	38	65	15:45
sum	113	144	33:23
	Systran		
	user turns	system turns	elapsed time
1st	42	69	20:05
2nd	43	65	17:32
sum	85	134	37:37

Table 4.21 – Number of user turns, system turns and elapsed time in minutes.

Quality: Dialogue quality was measured with the use of the source button which was clicked 25 times in all interactions. In the interviews participants stated that they did not understand the system's output in five of these cases (two times *Google*; three times *Systran*). Especially, the *Systran* compound *'Überlandleitungverbindung'* caused confusion. In seven cases participants did not trust the translation (three times *Google*; four times *Systran*). Participants stated that they used the button because they hoped for more information in the English source text and they believed that sometimes information had gone missing in the translation. In the other 13 cases the button was used out of curiosity because the participant wanted to see what happens, or to understand where an awkward translation originated from (five times for *Google* and eight times for *Systran*).

To assess the dialogue quality, it was also examined how often the break off utterance and the three utterances that signal an interaction problem were used in the dialogues. The break off utterance *'Sorry, but I have no information on that topic.'* was used twice (one time *Google*; one time *Systran*). The negative feedback *'Sorry, I did not understand you. Could you say that again?'* was used in seven cases (five times *Google*; two times *Systran*). The system could

Q		*GOOGLE*	*SYSTRAN*
Q1	I always knew what the system was asking me for.	2.5; 3	3; 3
Q2	I had serious problems understanding the German texts.	4; 4	4; (4, 5)
Q3	I would rate the German utterances as excellent.	3.5; (3, 5)	4 (3, 5)
Q4	There were awkward words and phrases in the German dialogue.	2.5; 2	2; 2
Q5	The German utterances were fluent.	3; 3	3.5 (3, 4)
Q6	I would rather use the English original.	2; 2	2; 1
Q7	I would rate the German utterances as incomprehensible.	4; (3, 5)	4.5; 5

Table 4.22 – Median and mode for questionnaire items. A five-point Likert scale was used, ranging from 1 (strongly agree) to 5 (strongly disagree).

also offer the participant to go back a step which happened once with the *Google* system. The system utterance *'Do you want to start over again?'* did not occur in the dialogues.

Perception of MT Output: In addition to the interaction analysis seven statements were included into the questionnaire to obtain a more detailed insight in how participants perceived the MT output. Notably, the statement *'I had serious problems understanding the German texts.'* was rated negatively for both systems (*Google* median of 4, mode of 4 and *Systran* median of 4 and mode of 4 and 5). Responses to 27 out of the 56 ratings (= 8 participants x 7 items) did not show any differences between the two systems. Table 4.22 presents the results for these items (Q1-Q7). *Google* performed better in the case of questions Q1, Q3, Q4 and Q5, whereas *Systran* performed better for Q6 and Q7.

Insights from Interviews: Interviews with participants P1 to P8 that were carried out after they had worked through the two scenarios, reveal some background on the participants' judgments of the MT output. For instance, wrong word orders were recognized by all participants during the interactions, but were not perceived as big obstacles for comprehension, they only 'spoil the overall impression' (P3). The word *'sorry'* was not recognised as non-German by P1, P5, P6, but P3, P4 and P7 said they perceived it as 'impolite in the context of an information system'. P4 and P6 only noticed that the *Google* utterance (*'Have a good day'*) was not translated to German when they were asked about it in the interview. A possible explanation for this lies in what P7 said: 'as an ERASMUS student switching between the two languages is in the daily routine'. P2 remarked that the translation of *'options'* into *'Wahlen'* by *Systran* did not really influence the comprehensibility but 'disturbs the fluency'. In principle, *'Wahl'* is an acceptable translation for *'option'* in the sense of *'choice'*. However, the plural (*'Wahlen'*) can only be used for the German word *'Wahl'* in the sense of *'election'*. The spelling mistake in the translation for *internet connection* by *Systran* (*'Internetanschlusse'*) was not recognized in the interaction by any of the subjects. A considerable number of participants also only realised the absurdity of the *Systran* compound *'Kilobyteantriebskraftgeschwindigkeit'*, the translation for *'kilobyte upload speed'* during the interview. A possible reason for this may be that the word appeared in a lengthy description which was only scanned by participants to find relevant

values (P7). Another reason could be the participants' low computer literacy. For example, P1 said 'I read it and thought that this is just another of these computer terms' and P7 'saw Kilobyte and did not really read the rest of the word'. Similar responses were observed with the translations of *'download allowance'*. Both systems failed with the translation (*'download Zertifikat'* (*Google*) and *'Downloadzulagen'* (*Systran*)). P1 said he thought that it was not 'of much importance' to understand the word properly, because it was 'not essential to carry on with the conversation'.

4.3.3 Discussion and Conclusion

An intrinsic and extrinsic evaluation of MT output was performed and it was expected that the best translation would receive the best ratings in the intrinsic evaluation, and results in the smoothest interaction in the extrinsic evaluation. However, results present a different picture. The intrinsic evaluation showed a clear preference for the *Google* translation. A similar result is reported in [Kit and Wong, 2008] where the BLEU and NIST scores of *Google* are slightly better than those of *Systran* for most language pairs.[13] Annotation of the MT output indicated that the preference of the judges was most likely caused by the fact that *Systran* included more erroneous words in its output due to ambiguous source words, spelling mistakes and the generation of non-existent compounds. However, when the MT outputs were used in a dialogue setting, the better performance of the *Google* system in terms of intrinsic evaluation was not reflected in the interactive context.

In terms of efficiency, it took participants slightly longer to finish their tasks with the *Systran* system, but in carrying out their first task, participants needed more dialogue turns when using the *Google* system. The interaction quality analysis suggest that the *Google* system performed less well than the *Systran* system, since the former resulted in dialogues containing more utterances which indicates interaction problems. From the use of the source button a difference between the systems could not be concluded. The assessment of the system utterances indicated that participants had no difficulties in understanding the system and only in a few cases preferred one system to the other. In summary, the extrinsic evaluation did not render a preference for the *Google* system as was expected from the outcome of the intrinsic evaluation. In the interviews it was discovered that most translation errors, especially those of the *Systran* system, were only recognised by a small number of participants during the interaction or not at all. This may have resulted from the computer literacy of the participants but the experiment task may also have had an effect.

Despite the fact that the extrinsic evaluation is a small-scale study in a limited domain

[13] This study considered the domain of legal texts and compared translations from various languages into English.

involving a single language pair, results show a difference to the evaluation of the MT output in isolation. Depending on the context of use, the decontextualised intrinsic methods may not provide us with the most accurate results. As future work it could be interesting to add extra errors into the MT output, in order to investigate the issue of user acceptance versus error rate in greater depth.

In summary, this experiment confirmed that evaluation efforts need to keep the user and application context in mind if they aim to produce meaningful results for real-life situations. In setting up the WOZ study it was also discovered that existing standardised questionnaires do not capture the system performance in terms of the interaction factors necessary to determine the quality of dialogue systems. Finally, very valuable information was gained through subject interviews as an additional measurement method.

4.4 Text to Speech

As aforementioned in the *Related Work* Chapter, the evaluation of TTS system is an especially difficult task because many subjective factors influence the perception of a TTS output and evaluation of TTS systems is therefore still dependent on human judges. There are no universally agreed evaluation criteria for TTS systems and no evaluation method did achieve wide acceptance. The perhaps best known evaluation metric that has been used is the mean opinion score (MOS) which has been discussed in detail in the *Methods* Chapter in Section 3.2.3. MOS and most of the other methods that were discribed in the section on TTS evaluation (Section 2.6.2) have in common that they fail to capture the overall quality of the interaction with a TTS system and to predict how the output might affect task performance. Evaluation efforts that take into account the task at hand and the context of use are mostly restricted to the evaluation of spoken dialogue systems.

This section describes an experiment to compare intrinsic and extrinsic TTS evaluation methods. The assessment of the employed TTS system was combined with an experiment to evaluate the combination of MT and TTS. Both aspects under observation, the pure TTS evaluation and the evaluation of combined MT and TTS output, were two-fold, combining an intrinsic and extrinsic evaluation part. In the intrinsic evaluation a modified mean opinion score was applied. For the extrinsic evaluation a Wizard of Oz study was conducted, based on the extrinsic MT experiment (cf. Section 4.3.2). The scenario was re-used in order to take advantage of the already established system utterances and system set-up, as well as the knowledge that has been gained while carrying out the MT experiment.

Several hypotheses were under test. In the intrinsic assessment it was expected that the mistakes in the machine translation would have an influence on the perception of the synthesis.

Therefore the synthesised utterances from gold standard translation are expected to be rated better than the synthesis of the same utterances which were machine translated, although the same synthesiser was used to synthesise both sets of utterances. Similarly, in the extrinsic evaluation it was expected that the interaction in the control group which used the system with the gold standard translations, would be smoother and faster than the interactions in the group with the machine translation. Since the extrinsic evaluation also included a comparison between the text and the speech mode another hypothesis was that in the control group the interaction in the speech mode would be smoother and faster than in the text mode. A fourth hypothisis is that the interaction in the machine translation group will be more troubled in speech mode than in text mode. This is expected due to the fact that mistakes in the machine translation are very likely which will, due to the transient nature of speech, probably have a bigger influence in the speech modality than in the interactions with text output.

Material: The set of English utterances that was used in the extrinsic MT experiment was reviewed and supplemented with three longer description texts. The ten utterances with slots were multiplied and filled with the respective slot fillers because this simplified the implementation of the WOZ experiment with speech output. The resulting set comprised of 72 English utterances. This set was translated by an independent human translator, constituting in a set of 72 gold standard translations. The English set was also translated using the *Systran* translation service[14] to form the MT set of utterances. These three sets of utterances were synthesised into speech output using the *MUSE* text-to-speech system [Cahill and Carson-Berndsen, 2010]. The synthesised utterances varied in length between less than a second and 49 seconds, with a majority of very short utterances (between 1 and 3 seconds).

4.4.1 Intrinsic Evaluation

The intrinsic evaluation of the TTS system is based on the mean opinion score that was introduced in Chapter 2.

Subjects: Eight native speakers of the German language took part in the experiment, who were unfamiliar with the dialogues that have been used in the experiment. An overview of their demographics and their experience with TTS systems can be found in Table 4.23. Their age ranged from 26 to 34 years with an average age of 30. Five of the participants were male and three were female. Participants stated that they had no or only minor familiarity with TTS systems.

Material: A selection of 25 utterances was the basis for the intrinsic evaluation. These utterances were a refined set of the 72 utterances that were discussed in the previous section.

[14] `http://www.systranet.com/` All translations were generated on 16/01/2012.

number	age	gender	experience with TTS systems
01	34	male	minor
02	32	male	minor
03	32	female	none
04	31	female	none
05	28	male	minor
06	29	male	minor
07	26	male	none
08	28	female	minor

Table 4.23 – Overview of the participants of the intrinsic TTS evaluation.

Very short utterances like *ok*, *yes* and *no* were omitted and only two of the slot filler result-utterances (one for a landline and one for a mobile Internet recommendation) that were found in different versions in the original set were chosen. Two sets of these 25 utterances were used in the evaluation, one was a set of gold standard translations (GS), the other set comprised the *Systran* machine translations (MT), resulting in a total of 50 utterances to be assessed.

Procedure: A listening test similar to the mean opinion score (MOS) (cf. Section 2.6.2) was carried out. To implement the experiment an open source survey application called *LimeSurvey*[15] was utilized which enabled web-accessibility for the participants. Several modifications have been made to the original ITU Recommendation [ITU, 1994] with reference to criticism on MOS. Firstly Lewis proposed to use seven-point scales with labels at the end of the scale [Lewis, 2001] which was put into practice in the experiment described here. Another point of critique is that it is problematic to mix global items such as the overall sound quality with items that are more specific such as articulation [Viswanathan and Viswanathan, 2005]. It has been shown that in MOS the global item on overall sound quality is independent of the other items of the MOS scale [Viswanathan and Viswanathan, 2005]. Therefore in the setting of the intrinsic TTS evaluation that is subject of this section the listeners had to evaluate the overall speech quality in general for all 50 (25 machine translated and 25 gold standard) utterances in a first round. The scale ranged from bad (0) to excellent (6). In order to prevent ordering effects, the presentation of the utterances was randomised between participants.

In a second round the listeners had to evaluate the 50 utterances again, according to the following four attributes:

- speed (too fast ... too slow)
- distinctness (very clear ... very unclear)
- stress (not annoying ... very annoying)
- naturalness (very natural ... very unnatural)

[15] http://www.limesurvey.org/

For these four attributes a seven-point scale with labels at the end was also employed. The respective labels are shown next to the criteria in the itemisation above. The presentation of the four criteria was randomised between participants and the order of the utterances was shuffled within the criteria evaluation, also in order to avoid order effects in the experiment. Altogether the participants had to rate the two sets (MT and GS) of 25 utterances in five different categories which resulted in a total of 250 values for each of the participants. Since the evaluation of 250 utterances is very demanding the participants were able to pause the experiment, save the results and return to the evaluation at a later time.

Results: For each of the 25 utterances the experiment resulted in eight ratings (one per participant) for each of the five rating criteria, once for the pure TTS mode and once for the combination of MT and TTS. In the following the results of this evaluation will be presented ordered by the rating criteria that were judged. The description will start with the overall quality judgements, this will be followed by an outline of the rating results for speed, accentuation, distinctness and naturalness.

First of all, the participants had to judge the overall quality for the 25 MT utterances and the 25 gold standard utterances which were presented to them in random order. They had to judge the overall quality on a seven-point Likert scale from bad (0) to excellent (6). For each utterance the median and mode of all eight ratings were determined (cf. Table D.1 in the Appendix D). A plot of the mdeian for all utterances comparing the gold standard and the machine translation can be found in Figure 4.5. It can be seen that the overall quality of the utterances was rated to be rather bad in both cases. For the gold standard the median of all ratings for all utterances and participants was 3 (mode = 3) and for the machine translation it was 1 (mode = 1). For most of the utterances the gold standard ratings are well above the machine translation ratings. Only utterance 19 received better ratings in the machine translation (median of 1 in GS and 1.5 in MT). The differences in the ratings are very small and looking at the utterance (cf. Appendix D) it can be seen that the gold standard and machine translation are almost similar for the utterance. The linear correlation between the ratings for gold standard and machine translation are only medium with a Pearson coefficient of $r = 0.32$ which can also be concluded by looking at the two graphs in Figure 4.5.

To see what factor might have influenced the overall quality rating the four attributes, speed, distinctness, stress, and naturalness were rated by the 8 judges. The order of the utterances and the order of the attributes were randomised to prevent order effects.

Speed was judged on a seven-point Likert scale from too fast (0) to too slow (6). Again median and mode of the ratings of the eight judges were determined (cf. Table D.1 in Appendix D) and a plot of the median values for all utterances comparing the gold standard and the

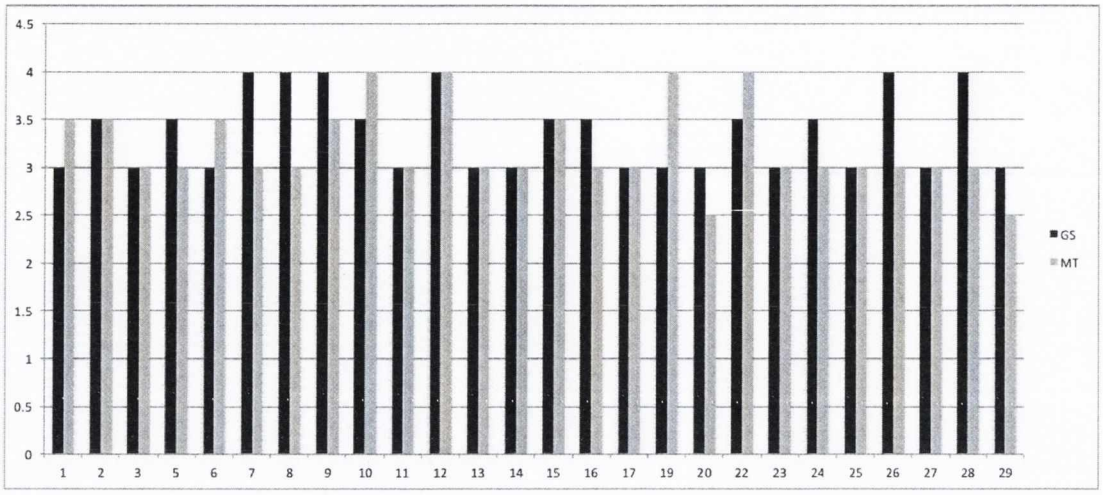


Figure 4.5 – Comparison of the median values of gold standard (GS) and machine translation (MT) ratings for overall quality for each of the 25 utterances. The seven-point Likert scale ranged from bad (0) to excellent (6)

machine translation can be found in Figure 4.6. In theory if the speech is full of mistakes and therefore the listeners needs more time to process the information it will be perceived as being to fast. Overall it can be said that both gold standard and machine translation were perceived to be almost at the right pace (median and mode of 3 in for GS and MT) but looking at the average over all ratings it can be assumed that it was rather too slow (GS avg=3.39, stdev= 0.92 and MT avg= 3.29, stdev=0.97). A value of exact three in the rating would have meant that the utterance was perceived to be uttered in the exact right speed. Looking at the averaged ratings this value was assigned to ten MT utterances and only to two GS utterances. In the single ratings, however, out of the 200 ratings overall (eight judges rated 25 utterances in each mode), the value three was assigned in 92 cases to the GS utterances and in 94 cases to the MT utterances. Therefore, it must be deduced that the GS utterances were simply judged to be slower in more cases than the MT utterances which could lead to the conclusion that the pace of the synthesiser could be increased. The medians of the ratings show a low variance (corrected variance GS=0.25 MT=0.20). Two of the utterances (20 and 29), in both cases MT utterances, were rated as being too fast. The automatic translations in these two utterances are strikingly error prone.

Accentuation was rated on a seven-point scale from annoying (0) to not annoying (6). The median values of the ratings of the eight judges can be found in a Table D.1 in Appendix D. The gathered data was plotted comparing the gold standard and the machine translation. The plot can be found in Figure 4.7. Overall the gold standard and the machine translation are perceived to be rather moderately accentuated, with the gold standard performing only

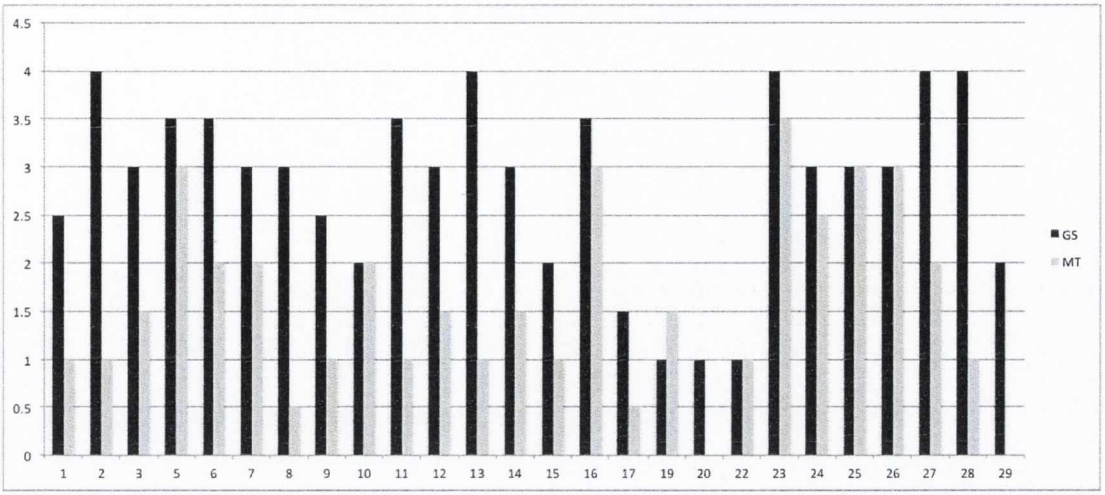


Figure 4.6 – Plot of the median values of gold standard (GS) and machine translation (MT) speed ratings for each of the 25 utterances. The seven-point Likert scale ranged from too fast (0) to too slow (6).

slightly better on average over all participants and ratings (GS median and mode of 3; MT median and mode of 2). The correlation between the ratings for the gold standard utterances and the machine translation utterances is only small ($r = 0.29$), the correlation between the overall quality ratings of the GS with the accentuation of the GS utterances and the correlation between the overall quality ratings of the MT with the accentuation ratings of the MT are on the other hand strong ($r = 0.85$ and $r = 0.76$ respectively). Three of the utterances had a better rating for the machine translation than for the gold standard translation. The median for the GS version of utterance 5 is 3 and therefore one rating scale point lower than the median of the MT utterance. The translations of utterance 25 and 26 exactly the same for gold standard and machine translation, however their median differs by 0.5 scale points.

Distinctness was rated on a seven-point scale from very clear (0) to not very unclear (6). The median values of the ratings of the eight judges can be found in Table D.1 in Appendix D. The ratings were plotted to enable a comparison between the gold standard and the machine translation. This plot can be found in Figure 4.8. The averaged distinctness ratings show a very high variance (corrected variance GS=1.42 MT=1.73). Overall the rating for the GS are considerable better than the ratings for the MT utterances, with a median of 2 (mode = 2) for GS and a median of 4 (mode = 5) for MT. Three of the utterances got a better rating for the machine translation (5, 19, and 23). These difference between GS and MT ratings were, however, very small except for utterances 5. The distinctness ratings correlate strongly with the overall quality ratings for GS ($r = 0.82$) and for MT ($r = 0.80$). The correlation between the MT ratings and the GS ratings is much lower ($r = 0.58$).

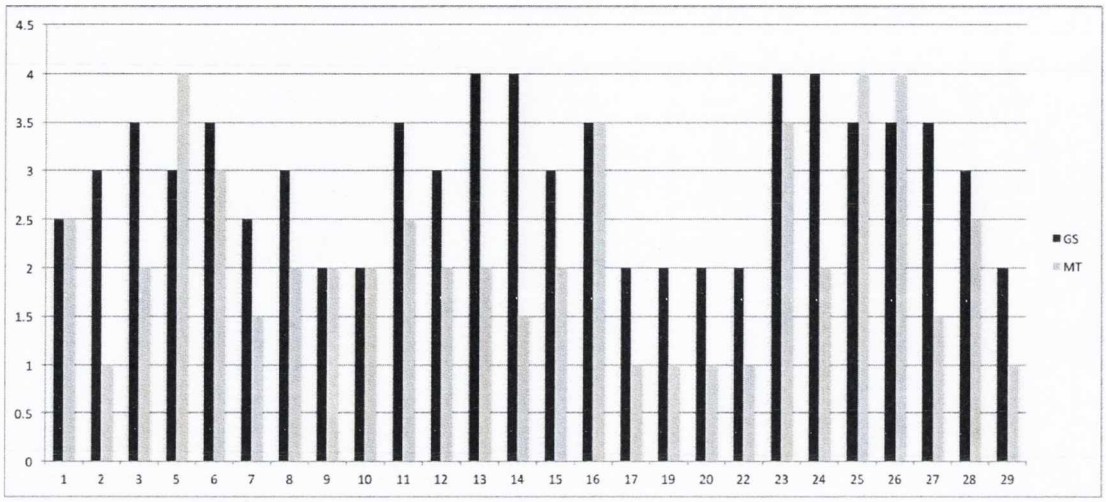


Figure 4.7 – Plot of the median values of gold standard (GS) and machine translation (MT) ratings for accentuation. The seven-point scale from annoying (0) to not annoying (6).

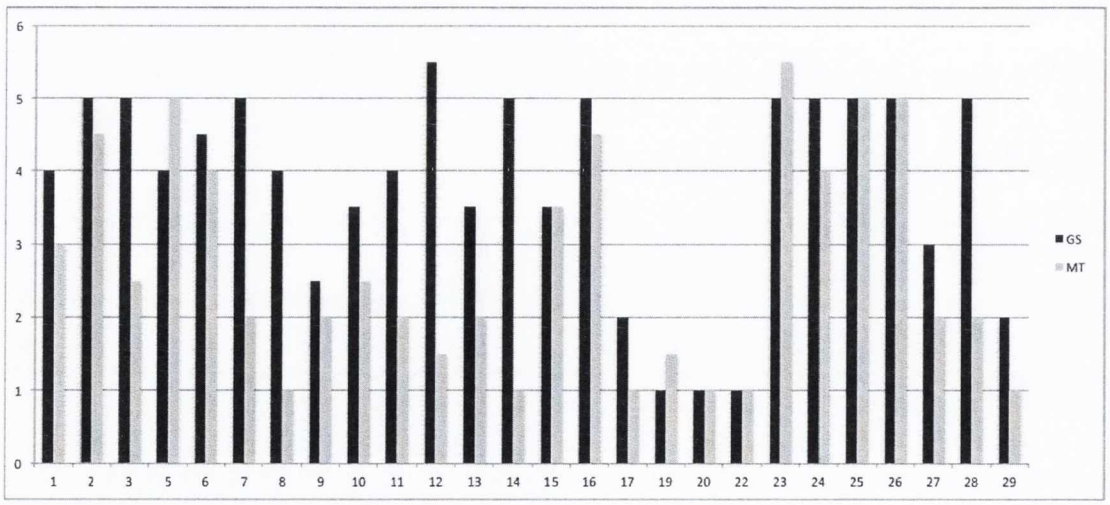


Figure 4.8 – Plot of the median values of gold standard (GS) and machine translation (MT) ratings for distinctness. The seven-point scale from very clear (0) to not very unclear (6).

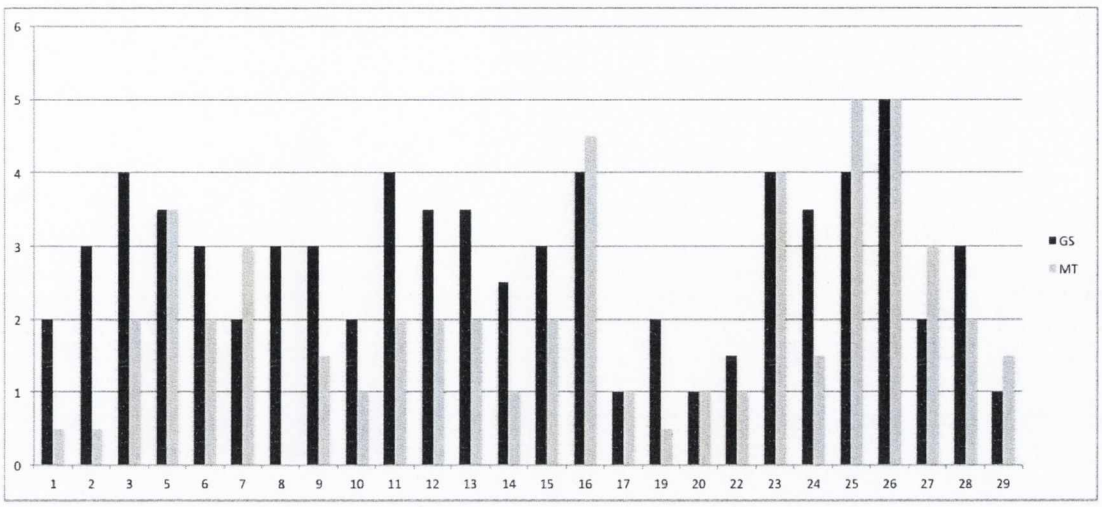


Figure 4.9 – Plot of the median values of gold standard (GS) and machine translation (MT) ratings for naturalness. The seven-point scale from very natural (0) to not very unnatural (6).

The perceived naturalness for the utterances had to be rated by the eight listeners on a seven-point scale from very natural (0) to not very unnatural (6). The averaged ratings of the judges can be found in Table D.1 in Appendix D. A plot of the median values for all utterances comparing the gold standard and the machine translation can be found in Figure 4.9. The ratings also show a high variance (corrected variance GS=0.69 MT=1.06). They correlate linearly with the overall quality ratings (GS $r = 0.64$ MT $r = 0.77$) and with a Pearson coefficient of $r = 0.69$ the MT and the GS ratings seem to correlate with each other. For the five utterances number 7, 16, 25, 27 and 29 the averaged ratings for the machine translations are better than the ratings for the gold standard.

4.4.2 Extrinsic Evaluation

The extrinsic evaluation of the TTS system was based on the setting of the Wizard of Oz experiment which was carried out in the extrinsic MT experiment (cf. Section 4.3.2). The Internet search scenario was re-used in order to be able to fall back on the already established system utterances and to profit from the knowledge that has been gained while carrying out the earlier MT experiment. The refined set of utterances was synthesised and implemented into a newer version of the Wizard of Oz prototyping tool *WebWoz*[16].

Subjects: Sixteen native German speakers took part in the experiment. They were unfamiliar with the utterances that were used in the experiment. Their age ranged from 21 to 31

[16]The iterative development process of *WebWoz* was driven by the two WOZ experiments that are described in this chapter. For a better insight on how the experiments influenced the development process please refer to [Schlögl et al., 2010a, 2011]

years with an average age of 24.94. Six participants were male and 10 were female. These sixteen subjects were divided in two groups of eight participants. One group interacted with a system using the gold standard translation, the other dealt with the machine translated utterances. An overview of the demographics of the 16 participants can be found in Table 4.24. The age in the gold standard group ranged from 21 to 26 years, with an average of 23 years. The age in the machine translation group ranged from 22 to 31 years, with an average of 26.875 years. The participants were, furthermore, asked to state their expertise with dialogue systems, the English language, and computers in general. Considering the general computer skills and the experience with dialogue systems the gold standard group did show a better familiarity on average than the MT group. However, all participants only had moderate to no experience at all with dialogue systems and stated good knowledge with the English language on average.

number	age	gender	computers	experience with dialogue systems	English
machine translations					
01	23	male	expert	minor	moderate
02	24	female	moderate	moderate	good
03	26	female	moderate	none	expert
04	21	female	moderate	none	good
05	22	female	moderate	moderate	expert
06	23	male	good	moderate	moderate
07	23	female	moderate	minor	good
08	22	female	moderate	none	good
gold standard					
09	22	female	moderate	minor	good
10	23	female	moderate	minor	moderate
11	24	male	good	minor	good
12	27	male	expert	moderate	expert
13	31	male	good	moderate	good
14	30	female	expert	none	good
15	29	female	expert	moderate	expert
16	29	male	good	moderate	good

Table 4.24 – Overview of the demographics of the 16 participants that took part the extrinsic TTS evaluation. Participant 1 - 8 interacted system which was implemented with the machine translated utterances. Participant 9 -16 used the gold standard version.

Participants in the machine translation group were on average older than those in the gold standard group. There is also a difference in the distribution of sexes between the two groups, while the gender distribution in the gold standard group was balanced, there were six women and only two men in the machine translation group. It is possible that the experience with computers in general and dialogue systems could have an impact on the results. In both cases the gold standard group seemed to have better experiences. Moreover, in the gold standard group six of the participants stated they had been facing the problem of getting an Internet

connection in Ireland already and might therefore have more context knowledge, whereas in the MT group only three participants have dealt with this subject prior to the experiment.

Procedure: A Wizard of Oz study was carried out similar to the experiment described in Section 4.3.2. Two systems were implemented, one with the gold standard translation, one with the machine translated utterances. In both cases the wizard was able to choose between speech and text for the output mode. A between subject evaluation was carried out with one group of eight participants interacting with the gold standard system and the other group of eight participants interacting with the MT system. Each participant had to fulfil two tasks with the system, for one scenario the system would answer with speech output and in the other scenario with text output. The two scenarios that were used in the study were:

- Imagine you stay in Dublin for the period of 6 months. You are looking for an Internet connection which you can use at home as well as on campus of your University. The offer should be as cheap as possible.

- Imagine you are looking for an Internet connection at your home which is as fast as possible. You are also looking for the highest possible download allowance. The price of the connection, therefore, does not matter to you.

The order of the tasks and the order of output modes was equally shuffled between participants in order to prevent ordering effects. First, participants were asked to fill out a questionnaire to obtain their demographic data. The following introduction to the study explained that the study was meant to try out two output modes (i.e. text and speech output) in a dialogue system that provides information on Internet offers and that the participants, therefore, had to carry out two tasks with the system, once with speech and once with text output. Subjects were also told that the dialogue systems had recently been automatically translated from English to German. This introduction was also given to the group of participants who dealt with the human translations. In addition, the 'source button' was introduced as a means to trace the source of the translation in the case of problems with the system output. Next, participants were asked to read one of the two scenarios, to find a solution with one of the two versions of the systems and to fill out a questionnaire about the interaction. Questionnaires for the interaction with the text output mode comprised of 23 items. Questionnaires in the speech mode had three additional items (the questionnaires can be found in Appendix D). A five-point Likert scale was used in the questionnaires, ranging from 1 (strongly agree) to 5 (strongly disagree). Questionnaires, instructions etc. were presented to participants on paper. Interactions with the system and interviews with the participants were audio recorded. This sequence was repeated for the remaining scenario and output mode. After finishing these tasks, the experimenter

explained that the participants had been interacting with a human wizard instead of a system and interviewed them about the system outputs on basis of the system logs and a set of prepared questions.

Results: In the extrinsic evaluation the performance was analysed in terms of task success, quality of the speech synthesis, translation quality, number of dialogue turns and reaction time.

Task success: Each participant had to carry out two tasks in which they were asked to find out the suitable offer according to a given scenario. For each of the two offers the participants had to protocol the company name that offered the Internet connection, the maximum download speed and the download allowance per month, as well as the monthly price. In seven cases participants returned the wrong offer. This happened in four cases in the gold standard group, and in three cases in the machine translation group, whereas three of the wrong results in the gold standard group were gathered in speech mode. In the MT group two of the wrong results were returned in the text mode. Also several results were returned with missing information about the offer (e.g. company name). Participant 4 for example explicitly stated on the result sheet that the company name was incomprehensible. Three such cases, where single information was left out could be found. Twice in the MT group, once in the speech and once in the text mode, and one time in text mode in the gold standard group.

Another indicator on how efficiently the task was fulfilled is the number of dialogue turns. The number of dialogue turns in text and speech mode was identified for both groups of participants. A participant's dialogue turn here is considered to include all words that are uttered between two system turns. In theory, the lesser turns were needed the smoother the interaction. An overview of the number of dialogue turns for each participant (text and speech mode accounted separately) can be found in Table 4.25. The minimum number of dialogue turns to complete the task was achieved by participant 01 and 04, with only 18 turns in the text mode. Both interacted with the MT version of the system. The maximum number of turns can be found in the control group in the speech mode (participant 10). With 41 turns this participant needed by far the most turns followed by participant 11 with 31 turns, also in the GS speech mode. It is also noticeable that the difference in the number of turns in the MT group is not bigger than six, whereas the differences between the number of turns in speech and text mode are rather high (e.g. 28 for participant 10, ten for participant 12 and eight for participant 15). The average number of turns for MT in speech mode is 23.375 and for text mode it is 23. In the control group the average number of turns in speech mode is 27 and 22.125 in text mode.

From Table 4.26 it can be seen that the number of turns needed by the control group with the gold standard translations (393) exceeds the number of turns that was counted in the MT group (371). Furthermore, there are considerably more turns in the speech mode (403) than in

participant	speech	text
01	23	18
02	27	27
03	21	21
04	19	18
05	28	30
06	26	21
07	20	26
08	23	23
09	19	20
10	47	19
11	31	24
12	29	19
13	25	19
14	21	21
15	21	29
16	23	20

Table 4.25 – Number of dialogue turns for each participant.

the text mode (361). The big difference can mainly be accounted for by the high number of turns in the gold standard group interacting with speech mode (216).

	GS	MT	sum
speech	216	187	403
text	177	187	361
sum	393	371	

Table 4.26 – Overview of the number of dialogue turns summed over the machine translation (MT) and the control group (GS) for speech and text mode.

Quality of the speech synthesis: Similar to the extrinsic evaluation of MT output which was described in the previous section (Section 4.3.2), after each interaction the participants had to complete a questionnaire regarding several factors. The questionnaire that was handed out to the participants after interacting with the system in speech mode included three additional items about the perception of the synthesis. A Likert scale ranging from 1 (always) to 5 (never) was used for Q3. Q1 and Q2 had to be rated on the Likert-scale ranging from 1 (strongly agree) to 5 (strongly disagree). The aggregated results can be found in Table 4.27. The ratings show an average around 3 and a high variation, resulting in a high standard deviation. Therefore the results cannot count as significant.

In the post experiment interview the control group was asked to indicate abnormalities they noticed in the synthesised utterances. Almost all experienced a wrong accentuation or that words were swallowed. Three participants noticed echo effects. Others observed situations where the system made pauses at wrong or awkward times and left pauses out where there should have been one. Participant 11 further experienced clicking noises in the synthesis. An

	gold standard	machine translation
Q1 I did not like the voice of the system	3.125 (1.4577)	3.25 (1.3887)
Q2 The quality of the speech output was good.	2.375 (1.0607)	3 (0.9258)
Q3 The speech output was comprehensible.	2.5 (0.5345)	2.625 (0.9161)

Table 4.27 – Overview of the three items about speech synthesis. These were unique to the questionnaire that was handed out after the interaction with speech output. The results are aggregated over all eight participants in the gold standard and the machine translation group. Standard deviation in brackets.

overview of their answers can be found in Table 4.28.

participant	09	10	11	12	13	14	15	16
wrong accentuation	✓	✓	✓	✓	-	✓	✓	✓
swallowing of words	✓	-	-	✓	✓	✓	✓	✓
problems with pauses	✓	-	✓	✓	-	-	-	-
echo effects	-	✓	-	-	-	-	✓	✓
clicking noises	-	✓	-	-	-	-	-	-

Table 4.28 – Overview of perceived problems of the synthesised utterances.

Translation quality: Six items in the questionnaire were targeted to assess the translation quality. A five-point Likert scale from strongly agree (1) to strongly disagree (5) was employed. Both groups answered the questions although the control group was interacting with a system that was operating on the human gold standard translations. In the instruction the control group was also advised that the system had been translated automatically. The aim behind this was that both groups should interact with the system under exact the same condition.

		gold standard		machine translation	
		text	speech	text	speech
Q1	I always knew what the system was asking me for.	1.625 (0.488)	2 (0.690)	2.375 (1.272)	3.375 (0.690)
Q2	I would rate the German utterances as excellent.	1.875 (0.577)	2.875 (0.899)	4.375 (0.755)	4.75 (0.488)
Q3	There were awkward words and phrases in the German dialogue.	4.375 (1.112)	3.125 (1.215)	1.375 (0.534)	1.375 (0.534)
Q4	The German utterances were fluent.	1.75 (0.899)	3.375 (0.951)	3.75 (0.951)	4.125 (0.577)
Q5	I would rather use the English original.	3.375 (1.272)	3.125 (1.215)	2 (0.690)	1.875 (0.816)
Q6	I would rate the German utterances as incomprehensible.	4.75 (0.488)	3.5 (1.272)	3.375 (0.975)	2.875 (1.154)

Table 4.29 – Overview of the six items that assess translation quality aggregated over all eight participants in the gold standard and the machine translation group. Standard deviation in brakets.

Use of the source button: Comparing the usage of the source button between the group that had the gold standard translations and the group that used the machine translations it can be seen that the button was hit 18 times in both cases. Looking at the use of the source button for the machine translation group it can be seen that the usage was considerably higher in the

participant	text mode	speech mode
machine translation		
01	1	1 (results)
02	0	3 (results 2 times)
03	1	6 (results 4 times)
04	0	0
05	1 (results)	0
06	1 (results)	3 (results)
07	0	1 (results)
08	0	0
sum	4	14
control group		
09	0	2 (results)
10	0	8 (results)
11	1 (results)	1 (results)
12	0	3 (results)
13	0	3 (results)
14	0	0
15	0	0
16	0	0
sum	1	17

Table 4.30 – Comparison of the number of source button uses for text and speech mode.

speech mode (14 times) than in the text mode (4 times). The button was used to review results, twice in the speech mode, and eleven times in the text mode. Three other utterances caused the use of the source button. Once in text mode participant 01 was troubled to understand the machine translation of *'Sorry, I did not understand you. Could you say that again?'* (*'Traurig, verstand ich Sie nicht. Konnten Sie das wieder sagen?'*). This utterance includes several mistakes. The first sentence has a wrong word order. Another mistake that can be found in this utterance is the wrong translation of *sorry* into the word for sad rather than the excuse. Participant 03 was troubled with the machine translation of the utterance *'You have the option to sign on to 'pay as you go' where you do not have to sign a contract. Usually this option is more expensive per unit, but you have the option to stop using the services whenever you like.'*, and therefore wanted to have a look at the English original sentence (cf. C.3 in Appendix C). Especially the first sentence in the translation of the utterance comprises of many translation mistakes and is therefore rather incomprehensible. In the speech mode two different utterances caused the participants to hit the source button. One was a rather long explanation which includes several mistakes among them wrong word order and compounding of words that do not belong together (cf. utterance 29 D.4). Two participants (02 and 03) stated in the interview that they hit the source button since the utterance was incomprehensible to them. Participant 03 was further troubled by the utterance *'Do you have any other requirements? Do you, for example, prefer a lower price to the download speed? Or is price not an issue?'*, but stated

in the interview after the interaction that it was an issue with the content and not the translation. In the control group the source button was hit only to review results. This happened once in text mode (participant 03) and 17 times in the speech mode. Participant 10 in the control group showed insecure behaviour when fulfilling the first task in speech mode, reviewing all the results several times. Whenever an offer was presented, participant 10 hit the source button which triggered the text to appear on screen. This behaviour leads to a very high usage of the source button in the control group for the speech mode.

Reaction times: From the time logs the reaction times of the participants, that is the time span from one system utterance to another, could be gathered. The average reaction time in the speech mode was with 14.875 seconds slower than the average reaction time in text mode (12.813 seconds). Utilising the overall 764 reaction time values for all four modes a one-tailed t-test was performed. An overview of the t-test results can be found in Table 4.31. Running the t-test on the gold standard group, text mode against speech mode, it can be seen that the reaction time is significantly faster in text mode than for the speech mode. This significance could not be confirmed by the one-tailed t-test that was run on the MT group samples (text vs. speech). A possible explanation for this observation is that in the control group with gold standard translations the participants are reading faster than listening to the text, while this effect is annihilated in the MT group through the time that the participants need to interpret and correct mistakes. A t-test was also run on the data to compare the reaction times for GS and MT in text mode and in speech mode. The results show that reaction times in the control group were significantly faster than reaction times in the MT group for the text version of the system. This significant difference between MT and GS could not be confirmed in the speech mode. This is possible due to the system speed which is the same and dependent on the utterance length the participants might have had plenty of time to interpret and correct the meaning while they waited until the system finished speaking. This assumption can be backed by the results of the intrinsic evaluation which showed that the pace of the synthesiser was perceived to be rather too slow.

	t-test result	interpretation
GS (speech vs. text)	0.000162	significant
MT (speech vs. text)	0.126662	not siginificant
Text (GS vs. MT)	0.000892	significant
Speech (GS vs. MT)	0.196814	not significant

Table 4.31 – Overview of the p-value results of the one-tailed t-test testing for significant difference in reaction times for control group with gold standard translations (GS) and machine translation group (MT).

Based on the observation that the machine translation compounds words that do not be-

long together which was made in the extrinsic MT evaluation, a deeper investigation into the issue was executed. The machine translations included six compounds that were 'invented' by the machine translation system. In the post experiment interviews with the participants it was therefore established whether the participants noticed these mistakes. An overview of the results can be found in Table 4.32.

compound	01	02	03	04	05	06	07	08
USB-Stock (USB-stick)	✓	-	✓	✓	✓	✓	-	✓
land line (Telefonlinie)	-	-	✓	✓	-	-	-	✓
landline connection (Überlandleitungverbindung)	✓	✓	✓	✓	✓	✓	✓	✓
kilobyte upload speed (Kb.antriebskraftgeschwindigkeit)	✓	✓	-	✓	✓	✓	✓	✓
experience shows (Erfahrungsshows)	-	✓	-	✓	-	✓	-	-
download allowances (Downloadzulagen)	-	-	-	-	-	-	-	-

Table 4.32 – Overview of the compounding mistakes. Participants were asked whether they noticed these mistakes. If so, a ✓can be found in the respective cell of this table.

Related to the compounding mistakes, another interesting observation was made. One of the questions asked by the system required the participant to decide between one of two options. The original utterance was *'Are you looking for mobile Internet or a landline connection?'*. Option A, *mobile Internet*, and Option B, *landline connection*, where mistakenly translated to *bewegliches Internet* and *Überlandleitungsverbindung* respectively. The first mistake stems from the ambiguity of the word mobile which in this context should have been translated to *mobil*, The machine translation system, however, interpreted mobile in the sense of *agile* or *flexible*. The second mistake stems from the rule based nature of the machine translation software that was used. The system binds the translations for *landline* (*Überlandleitung*) and *connection* (*Verbindung*) into a compound that does not exist in the German language. The correct translation would have been *Festnetzanschluss*.

These mistakes are obvious and would have been easily noticed by fluent (native) speakers of the German language. Therefore, it was surprising to find that the participants in the experiment (all native speakers of German) repeated the wrong words in their answers. Except for participant 5 all answered by exactly repeating the word *Überlandleitungsverbindung*. Participant 5 abbreviated the word in his answer to *Landleitung*. None of the participants attempted a correction. In the post-experiment interview, the participants were inquired about this behaviour. Most of them stated that they knew that the words were incorrect and they were aware of the intended correct meaning due to the context, but they concluded the system would understand these words since it came up with them in the first place. They even explained that they assumed that the system would only understand these mistakes. This observation lead to the experiment that will be described in the following section on speaker alignment in synthesised machine translated communication.

4.4.3 Discussion and Conclusion

The intrinsic evaluation was expected to confirm the hypothesis that the mistakes in the machine translation will have an influence on the perception of the synthesis. Therefore the intrinsic evaluation aimed to prove that the ratings for the synthesised gold standard translations are better than the ratings for the same utterances which were machine translated and synthesised, although the same synthesiser was used for both sets of utterances.

The overall quality ratings for the GS are considerably better than the ratings for the MT utterances, with an average of 3.7 (stdev= 1.596) for GS and an average of 2.77 (stdev= 1.765) for MT. The linear correlation between the ratings for gold standard and machine translation (overall quality) are only medium with a Pearson coefficient of $r = 0.32$ which further indicates that the differences in the ratings are not due to the synthesiser. The naturalness ratings for MT and GS correlate with each other. A probable explanation for this correlation is that naturalness is a factor that depends on the synthesiser and is independent of the mistakes in the MT. However, on average the naturalness ratings were also higher in the GS group than in the MT group, as were the averaged ratings for accentuation and distinctness. Speed was rated to be to slow in many cases for the gold standard, whereas it was judged to be just right for most of the MT utterances. The conclusion that can be drawn from these results is that the pace of the TTS system can be enhanced for regular sentences that are error free and grammatically correct, whereas the pace might be well suited in cases where the input text can be expected to include mistakes (e.g. from a machine translation).

These results seem to back the hypothesis that the synthesised gold standard translations were perceived to be better than the synthesised machine translations, although the same synthesiser was used.

In the extrinsic evaluation it was expected that the interaction in the control group with the gold standard translation would be smoother than in the MT group. This would comply with the findings in the intrinsic evaluation which showed that the quality of the gold standard utterances was perceived to be better than the quality of the machine translated utterances, although both sets of utterances were synthesised with the same synthesiser. The extrinsic results for task completion do not confirm that more mistakes would result from an interaction with the machine translated version of the system. Interestingly enough, if wrong choices of offers are considered, the gold standard group showed the worst performance in task success, especially in the speech mode. Another indicator for the interaction was the use of the source button which enabled the participant to see the text of the utterance that had just been uttered. Comparing the usage of the source button in speech mode between the two groups it can be seen that the button was hit more often in the gold standard group (17 times) than in the MT

group (14 times). However, the button was used in the GS group only to review the results. Therefore, the reason behind the usage was not only that the utterance was perceived to be incomprehensible. The result utterance contained the necessary information to fulfil the task, hence, it was useful for the participants to have the utterance in writing in front of them, to be able to filter and copy the important information instead of asking the system to repeat the result. The use of the source button can, thus, be seen as a strategy to fulfil the task more efficiently, rather than as indicator for incomprehensible material. It has to be stated though, that in the post experiment interview several participants indicated problems in understanding the names of service providers which had to be inserted into the answer sheet. The number of dialogue turns was higher for the control group than in the MT group, contradicting the hypothesis as well. The t-test did show significantly faster reaction time for the control group, however, only in text mode.

Another hypothesis of the extrinsic evaluation was that in the control group the interaction in the speech mode would be smoother and faster than in the text mode. However, the number of turns in speech mode exceeded the number of turns in the text mode by far. The source button was also hit more often in the speech mode than in the text mode, although in both modes only to review results. As already hypothesised this might have been a strategy in the speech mode to enable the participants to quickly retrieve the necessary information for their answer sheets. The two tailed t-test which compared the reaction times of speech and text mode in the control group found that the reaction time was significantly faster in the text mode which further contradicts the hypothesis. A possible explanation is that the participants were simply much faster in reading for themselves than the synthesiser pace in reading out. The participants might even have skimmed over some of the utterances rather than reading every utterance word by word.

Contrary to this it was expected that the interaction in the machine translation group would be more troubled in speech mode than in text mode. This hypothesis was based on the premise that mistakes in the machine translation are to be expected and since speech is transient, it was hypothesised that the mistakes have a bigger influence in the speech modality than in the interactions with text output. The results of the extrinsic experiment show that the number of dialogue turns was about the same for text (184) and speech mode (187). The analysis of the use of the source button shows that it was used less often in text mode (4 times) than in the speech mode (14 times) which is a slight indication for the hypothesis to hold true. However the results of the t-test comparing the reaction times for speech and text mode in the MT group do not show a significant result, therefore contradicting the hypothesis.

The extrinsic evaluation gives no clear indication that participants in the control group

experienced smoother interaction with the system than participants in the MT group, when looking at task completion, number of turns, the usage of the source button. The t-test also only confirmed faster reaction times in the text mode. Therefore the intrinsic results could not be reflected by the extrinsic results, rejecting the hypothesis that the utterances that were rated better in the intrinsic evaluation would also render a smoother interaction in the extrinsic evaluation.

4.5 Speaker Alignment in MT mediated Communication

Motivated by the observation of the Wizard of Oz experiments, in which it was observed that participants regularly repeated the erroneous machine translation output when answering the system's questions a targeted study with thirty participants was performed. This aimed at finding out whether people would align their answers to the wording presented in the questions or if they would correct the given alternatives or engage in some sort of attempted repair. This section is largely based on [Schneider and Luz, 2011].[17]

Understanding user expectations and behaviour when communicating with machines has long been regarded as an important element in the design of spoken language dialogue systems, being seen by many as having an impact comparable to that of good speech recognition and synthesis technology on the success of such systems [Dybkjaer and Bernsen, 2001]. Human factors research into the way people communicate in everyday dialogues has therefore drawn on linguistics and cognitive science theories (e.g. Grices' maxims, speech act theory) and empirical findings (e.g. memory constraints, lexical and syntactic alignment) to create system design guidelines. Similarly, applications such as speech-to-speech translation which employ speech technology to mediate human communication, stand to gain from the results of human factors research.

In (same-language) human communication, dialogue partners often align their linguistic behaviour to one another in terms of lexical and grammatical choices. Successful communication is more likely to occur when they become well aligned [Pickering and Garrod, 2004]. This phenomenon[18] has been extensively studied by linguists and psychologists. Findings of these studies have been taken up by language technology researchers and designers of spoken language dialogue systems, who exploited the fact that people can be shaped by the system's output in their lexical choices and phrase structures [Zoltan-Ford, 1991; Brennan, 1996; Gustafson et al., 1997]. Techniques developed along these lines have been used for language

[17]The study described in this paper was planned and conducted by myself under the supervision of my supervisor Saturnino Luz who also kindly helped with the refinement of my first draft of the paper.

[18]Different terms (e.g. entrainment, priming, lexical convergence) are used to refer to this phenomenon in the literature. For the remainder of this thesis, however, only the term 'alignment' will be used.

modelling and the design of repair strategies. Given this background, one might expect that other forms of computer mediated dialogues such as communication in speech-to-speech translation could also benefit from the phenomenon of alignment. Contrary to that expectation, this study suggests that even though the phenomenon of alignment can be exploited to good effect in traditional spoken language dialogue systems, the same phenomenon can be detrimental to the performance of S2S translation systems. In other words, in certain situations, alignment is something designers of S2S translation systems will need to guard against. It seems that the fact that communication using a S2S system is a two way process that involves humans is often overlooked. This study indicates that it is also necessary to take into account the influence of the S2S output on the ASR component. As S2S translation applications become more widespread, alignment issues need to be revisited so that their implications for the design of such interactive systems can be better understood. In particular, designers need a greater understanding of how to preserve the positive effects of alignment currently exploited in dialogue systems while avoiding its pitfalls.

4.5.1 Related work

The phenomenon of linguistic alignment has been researched in the fields of linguistics (especially in the area of corpus linguistics) and cognitive science. Alignment in human-human communication can range from the word level to sentence level. Brennan and Clark use the term *lexical entrainment* to refer to the fact that when two people repeatedly discuss the same object, their lexical choices tend to converge. In doing so people achieve conceptual pacts, or shared conceptualisations which they mark by using the same terms [Brennan and Clark, 1996]. Evidence of alignment on a sentence level can also be found. In such cases the term *priming* is usually used which refers to a process that influences linguistic decision-making, where a linguistic choice (prime) of a speaker influences the recipient to make the same decision, i.e. re-use the structure, at a later choice-point [Reitter et al., 2006a].

In human-computer conversation, data have also been collected which confirm alignment. Brennan reports on a Wizard-of-Oz experiment using a database query task in which *lexical convergence* with computers is discovered [Brennan, 1996]. Results of an experiment by Branigan et al. show that alignment occurs whether participants believe they are interacting with another human participant or with a computer [Branigan et al., 2003]. These studies traditionally involve grammatically correct and error free language. However, Bortfeld and Brennan [Bortfeld and Brennan, 1997] observed a degree of *conceptual entrainment* in human-human dialogues, where native English speakers produce non-idiomatic referring expressions in order to ratify a mutually achieved perspective with non native speakers. Although this phenomenon

is somehow related to the participant behaviour that was observed in this study, I am are not aware of any other studies where alignment is investigated in the context of computer-mediated spoken language translation.

As mentioned above, the research on alignment, especially the observation that people can be shaped in their lexical choices so as to use vocabulary and phrase structures aligned to certain system outputs, has influenced work in the area of spoken language dialogue systems. In the early nineties a study tested whether people can be shaped to use vocabulary and phrase structure of a program's output [Zoltan-Ford, 1991]. The study found that users of natural language programs will model the program's output, that there is no difference in modelling or shaping effects between spoken and typed inputs, and that it is easier for people to model both the length and the vocabulary of a terse computer output than of a conversational computer output. Approaching the vocabulary problem, i.e. the potential for enormous variability in dialogue, Brennan summarises the results of a series of experiments, where visually separated pairs of people had to line up identical sets of picture cards in the same order, and two Wizard of Oz experiments using a database query task. The results of these studies are discussed with respect to their implications to modelling and constraining lexical variability in spontaneous human-computer dialogue [Brennan, 1996]. The vocabulary problem is also targeted by Gustafson et al. who report on Wizard of Oz experiments which show that people mostly adapt their lexical choices to system questions, and investigate the possibility to add an adaptive language model to the speech recognisers in a spoken dialogue systems [Gustafson et al., 1997].

4.5.2 Targeted question-answering study

In order to investigate this issue further, a smaller scale study with a simple targeted question-answering set-up was conducted. Participants were asked to answer questions similar to the one in the preliminary WOZ experiment in which the initial observation was made. The questions were set up so as to enable to assessment of the extent to which the alignment behaviour observed in the WOZ study would persist.

Procedure: Ten dual choice questions were prompted to the participants, who were asked to choose one out of two options (e.g. *Would you rather travel to Rome or to Athens?*). The questions were translated from English into German using the online machine translation service of *Systran*[19]. The aim was to investigate how people would react to obviously wrong translations, in particular whether they would correct the mistakes or adopt them in their answers. Therefore, the set-up had to force the participants to choose between incorrect options.

[19] `http://www.systranet.com/` (last accessed 21/08/2012)

In order to do so, the questions were adjusted as follows: in five of the sentences the two options were adjusted, so that they would both contain mistakes, and in the other five questions the two options were corrected so that they contained no errors. This setting ensured that in five cases participants had to select an erroneous option. The five sentences containing error-free options served as fillers and masked the intent of the study.

The machine translations were kept and only single words (the options) were changed. The rest of the sentence was not altered. Therefore, mistakes of the translation system were preserved. In the ten cases (=5 sentences $\times$ 2 options), where the options were adjusted mistakes were emulated that were seen before, since predicting errors of the machine translation system is non trivial. The options that were changed were simple nouns. The aim was that the introduced mistakes resemble the mistakes observed in the two WOZ studies, especially the one where the system produced a word that cannot be found in a German dictionary. It had to be sure that the mistakes would be identifiable when uttered by the participants for the analysis. However, another aim was to keep the wrong words comprehensible. Therefore, only additional letters or syllables were introduced into the words or an umlaut was changed. It was also ensured that there were similar mistakes for both options within one question.

Following the above described alterations, the resulting ten questions were synthesised using the *MUSE* text-to-speech system [Cahill and Carson-Berndsen, 2010] and recorded. An overview of the German questions and their original English version can be found in Table 4.33.

The experiment was set up as a question-answering session. Twelve presentation slides were prepared. The first and last slide contained an introduction and a debriefing respectively. One of the ten questions was on each of the remaining ten slides. The order in which the questions were presented was randomized, so as to control for order effects. The participants were instructed that they would get one of ten machine translated and synthesised questions at a time which they had to answer aloud before they could proceed to the next question. The transition to the next question automatically triggered the playback of the synthesised question. All answers were audio recorded and later transcribed.

Based on the results of the initial experiment, it was hypothesised that the participants would repeat the erroneous options.

Participants: The experiment was conducted with thirty native speakers of the German language, all aged between 25 and 42 (avg. 27.7). Thirteen of the participants were female, seventeen were male.

Results: In total, each participant had to choose ten out of twenty options. Five times, they were forced to choose between options described by sentences containing incorrect words as

	Sentences with correct options
1	Würden Sie mögen eher nach Rom oder nach Athen reisen? *(Would you rather like to travel to Rome or to Athens?)*
2	Finden Sie das Schach oder Fussball aufregender? *(Do you find chess or football more exciting?)*
3	Würden Sie eher Albert Einstein oder Marie Curie treffen vorziehen? *(Would you rather want to meet Albert Einstein or Marie Curie?)*
4	Mögen Sie eher zum Rockmusik oder Popmusik hören? *(Do you prefer to listen to Rock or Pop music?)*
5	Mögen Sie eher Früchte oder Gemüse? *(Do you prefer fruits or vegetables?)*
	Sentences with incorrect options
6	Finden Sie **olivengrün** oder **goldengelb** schöneres Farbe? *(Do you like olive green or golden yellow more?)*
7	Essen Sie eher mögen **Blauenbeerpfannkuchen** oder **Schweinerhaxe**? *(Would you rather eat blueberry pancakes or knuckle of pork?)*
8	Lesen Sie eher **Liebe-Geschichten** oder **Verbrechen-Bücher**? *(Do you prefer to read love stories or detective stories?)*
9	Konnten dir Sie eher vorstellen, um **Frösche-Schenkel** zu essen, oder zu essen brodelnde **Fledermause**? *(Could you rather imagine to eat frog legs or boiled bats?)*
10	Tun Sie mögen **Montagen** oder **Donnerstagen** besser? *(Do you prefer Mondays or Thursdays?)*

Table 4.33 – The ten sentences used in the experiment. The erroneous options in the last five sentences were highlighted. The English version is shown in brackets after each question.

the main descriptor. In the other five cases, they had to choose from correct options. Some of the options were picked only by few participants (e.g. *Fledermausen*, the incorrect translation of bat) whereas others were popular (e.g. *Blauenbeerpfannenkuchen*, the incorrect version of blueberry pancake). Figure 4.10 shows an overview how often the options in the five sentences with incorrect options were chosen.

In what follows, I will concentrate on the responses to sentences containing incorrect options. In about half the cases, participants corrected the mistakes in the wrong options. In 70 out of the 150 (= five sentences × thirty participants) cases, in which participants were facing a wrong translation, they aligned their answer to the word used in the wrong option. Only six participants corrected all five wrong options, three of the participants, on the other hand, aligned to the options completely. Participant 13 intended to correct *Verbrechen-Bücher* but failed by uttering the word *Verbrecher-Bücher* which also does not exist. The correct translation would have been *Kriminalroman* or *Krimis*. Participant 6 (who answered in whole phrases) not only repeated the mistakes in four out of five cases but also copied wrong phrases from the questions with the right options. In a question rendered by the system in the wrong word order, for example, he repeated the option and the verb phrase. The option was lexically correct

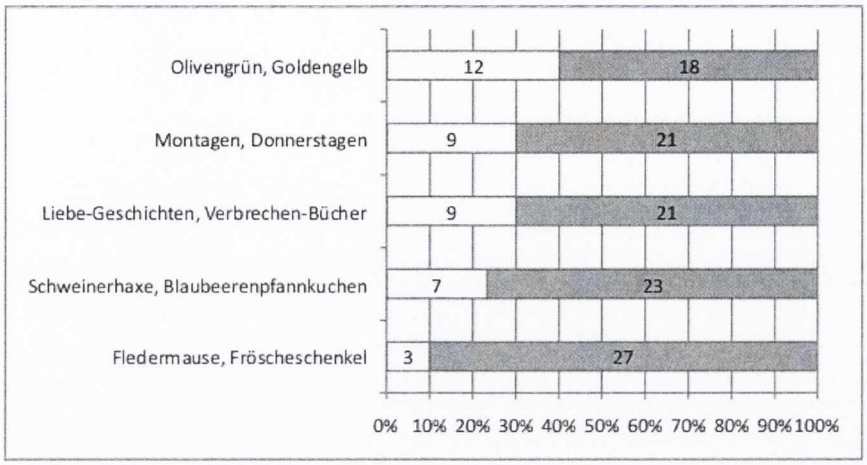


Figure 4.10 – Distribution of incorrect options as selected by the participants.

but the participant aligned to the wrong word order (e.g. question: *Would you Marie Curie or Albert Einstein rather meet?* participant's answer: *Marie Curie rather meet.*). Participant 30 repeated the mistakes first and later corrected them (e.g. *'Fröscheschenkel ... Froschschenkel'*). A similar observation was made when participant 26, who corrected all options, started to utter *Verbrechen...*, then paused and corrected the answer to *Krimis*. These occurrences were counted as corrections.

Wrong word	alignment	cor.	false cor.
Verbrechen-Bücher	17	3	1
Olivengrün	11	1	0
Goldengelb	10	8	0
Frösche-Schenkel	9	18	0
Blauenbeerpfannkuchen	7	15	1
Donnerstagen	6	15	0
Montagen	4	5	0
Liebe-Geschichten	3	6	0
Fledermause	2	1	0
Schweinerhaxe	1	6	0
sum	70	78	2

Table 4.34 – Overview of the ten erroneous options sorted by the number of participants who aligned with them (first column). The second column shows how often the participants corrected (cor.) the option, and the last column shows the number of false corrections (i.e. failed attempts by the participant to correct a wrong word).

Certain words were singled out for alignment more often than others. The option *olive green (olivengrün)*, for example, was chosen twelve times out of thirty and only corrected once. An overview of how often participants aligned to the options or corrected them can be found in Table 4.34. Due to the varying frequency of choice, the alignment rate was calculated

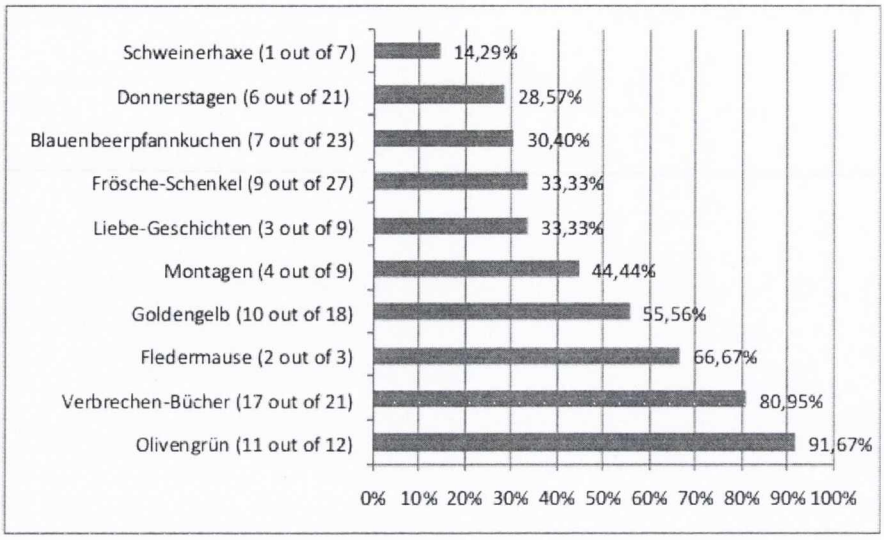


Figure 4.11 – Alignment values relative to the number of times that the participants chose the option.

relative to the number of times that every option was chosen. For an overview of how often an option was chosen and how often participants aligned to the chosen option, as well as the relative alignment to the number of times that the option was chosen, please refer to Figure 4.11.

On the other hand, it is also interesting to notice the cases in which participants corrected the mistakes most often. For example *knuckle of pork (Schweinerhaxe)* which was chosen seven times, was corrected in six cases. Table 4.35 shows an overview of the options that were corrected in more than 50% of the cases in which the respective answers were picked.

Wrong option	picked	correction (rel. value)
Schweinerhaxe	7	85.71%
Donnerstagen	21	71.43%
Blauenbeerpfannkuchen	23	69.57%
Frösche-Schenkel	27	66.67%
Liebe-Geschichten	5	66.67%
Montagen	9	55.56%

Table 4.35 – Overview of options that were corrected in more than half of the cases in which they were picked. The second column shows how often the participants picked the option. The third column shows the relative value of correction.

4.5.3 Discussion and Conclusion

The results of the targeted question-answering experiment support the initial hypothesis that people would align their behaviour and repeat incorrect words (and occasionally incorrect syn-

tax) when presented with machine translated and synthesised text.

In 47% of the cases, where participants had to choose between two options described by incorrect words they simply used the wrong terms instead of attempting a repair. One participant went as far as copying the wrong word order rendered by the MT system in the questions. Only 20% of the participants attempted repairs in all cases, with one failing to correct an option. The false correction of a word could also be observed in another case.

The results also show that some words are more likely to be corrected than others. It is speculated that this is due to the specific type of error introduced in the option. The translation of *blueberry pancake*, for example, was corrected in most of the cases. This might be due to the introduction of an extra syllable at the end of the translation of the word *blue* (*blau*). This syllable increases the effort needed to utter the word and disrupts its fluency. In the other cases, however, the fluency of the option was preserved. Since the experiment employed only a small number of questions to be answered in an otherwise de-contextualised situation, generalisation would be premature. A different experimental setting and further data would be needed in order to better investigate this hypothesis.

Since user awareness of the occurrence of miscommunication is often crucial to the success of computer mediated communication, from a practical system's design perspective, a deeper investigation into error types and their likelihood to trigger alignment in a fully contextualised situation (e.g. a task-oriented dialogue mediated by MT) would also be desirable. A complementary study could also assess the effect of the output modality (text versus speech) on user alignment to incorrect words.

Although details on the nature and effects of linguistic alignment to translated content still need elucidation, general implications of the findings reported above to the design of S2S translation systems can be pointed out. It would seem necessary, for instance, to better integrate speech recognition and machine translation modules so that the latter is prevented from producing output that might lead to recognition errors further down in the interaction sequence. Recent work on MT-ASR integration might help address this issue, at least in part. Jiang et al. for instance, propose a method for tight coupling between ASR and MT to enhance speech translation performance [Jiang et al., 2011]. The strength of their proposed strategy is that the MT system can recover from ASR errors before an ASR error is translated into an incorrect word. However, their approach carries information only one way, from ASR to MT. The results of this study indicate that there is also a need for a back channel from the machine translation to the ASR component.

Interaction design might also play an important role in avoiding problems due to alignment. Similarly to Brennan's recommendation for spoken dialogue systems, that the language

system should present only terms in the output that the system can process as input [Brennan, 1996], I suggest that in the first instance output of the MT component should be adjusted to the capabilities of the ASR component. Where this is not entirely possible, the system should implement robust mechanisms for meaning negotiation between the conservational partners. Such mechanisms could then perhaps feed information back to the machine translation and speech components in order to mitigate further problems.

Machine translation is still far from perfect, and it is likely that in a S2S translation scenario mistakes in the MT output will occur. The findings suggest that in such cases there is a good chance that people will adopt such mistakes as lexical items.

The goal of an ASR system is to map from an acoustic signal to a string of words. In order to perform this matching, the acoustic signal is matched to an entry in a dictionary which contains the word pronunciations, or the signal is split into phonemes (i.e. smallest segmental units of sound) and these are matched to graphemes (i.e. fundamental units in a written language). In situations where one communication partner repeats MT output that can not be found in the dictionary, the ASR performance will degrade. In spoken language translation the ASR error rate should generally be lower than for other applications that involve ASR such as information retrieval tasks for example, because the ASR output constitutes the input for the machine translation [Waibel and Fügen, 2008]. However, even if recognisers were able to match the phonemes to the correct string of graphemes this would only pass the problem on to the machine translation module which would produce an incorrect back translation. If the mistranslation of *landline connection (Überlandleitungsverbindung)*, for instance, was back translated by the same system the result would be *overhead power line connection* which would obviously cause semantic problems to the conversational partners.

Brennan [2000] surveys aspects of interactive language use, among them the entrainment phenomena, and discusses implications for computational linguistics and human computer interaction. She suggests that future dialogue interfaces should include resources to enable users to negotiate meanings, model context and recognise which referring expressions are likely to index a particular conceptualisation. The findings of this experiment show that there is a flipside to this argument, namely that potential problems due to alignment might arise when the machine assumes a mediating role.

The aim of study that is described in this section is to confirm the occurrence of alignment effects in a MT mediated setting. In order to be able to do that, the data employed in this experimental setting had to be constrained to a few clear-cut and comparable sentences. However, as mentioned above the small linguistic sample and the de-contextualised nature of the task that participants were asked to perform prevents from making broad generalisations. While this

sort of trade-off between coverage and testability is quite common in studies and evaluations of applications of language technology, I acknowledge that considerably more investigation into human factors is needed in this area.

Therefore, it is necessary to further study issues concerning linguistic alignment in contexts involving computer mediation through machine translation, speech recognition and speech synthesis. A more immediate plan is to investigate in which cases people are more likely to align to the mistakes. Results of this investigation could help to avoid such error types in the output. It will also be useful to examine if the conclusions of this study also apply to other languages and cultural settings. The next logical step is to repeat the study with English speakers. Finally, to explore the effects of presentation modality (speech, text, graphics) on the user's tendency to align to incorrect output is desirable.

4.6 Summary

Four experiments form the foundation of this research and although they all are stand-alone experiments with their own hypotheses, the common topic is the comparison of intrinsic and extrinsic evaluation methods. This section aims to summarise the outcomes of the results as fundament for the discussion in the next chapter of how these results interlink and contribute to the answer the research question.

The ASR component was evaluated intrinsically, using the traditional word error rate, and extrinsically by performing a map task experiment. The experiment examined the correlation between the extrinsic measures (i.e. time to task completion, direction scores and word ratio) and the WER values of sixteen ASR mediated dialogues. It has been shown that the extrinsic measures correlate with each other, but that no or only minor linear correlation could be demonstrated comparing the extrinsic measures with the word error rate analysis. A qualitative analysis of the errors in the ASR transcripts demonstrated that many errors only involve single letters per word (e.g. compounding and decomposition mistakes). The way the word error rate is calculated allows for no mistakes and therefore does not take into account the ability of a human listener to compensate small errors. Hence, on base of the assumption that the evaluation on word level might not be fine grained enough to reflect the ASR accuracy, five variants of error rates have been implemented and their correlation with the extrinsic measures have been analysed. The results were two-sided, one of the extrinsic measures, the word ratio, had a higher correlation with the WER but no correlation with the new metrics, but the other extrinsic measures (i.e. time to task completion, number of words uttered, direction scores) did show no or only a minor correlation with WER but achieve (borderline) medium correlation with the new intrinsic measures. This held especially true for the SoundER variant without

boundaries which matched homophones to the same encoding. However, the correlation between the extrinsic methods with the alternative intrinsic methods still only showed a small to medium correlation.

The MT experiment compared an intrinsic evaluation, in which the quality of translated sentences was assessed in isolation by three human judges, first using preferences ratings, followed up by the application of an error annotation scheme and an extrinsic evaluation which incorporated the translations into a Wizard of Oz system. The WOZ experiment was run with eight German subjects. Two machine translation systems were under test and it was hypothesised that the translation with the best ratings in the intrinsic evaluation would produce the smoothest interaction in the extrinsic evaluation. The results, however, did not support this assumption. While the intrinsic evaluation showed a clear preference for one of the two translation systems under test which was supported by other intrinsic evaluations reported in the literature, this better performance could not be shown by the extrinsic evaluation. When the MT outputs were used in a WOZ dialogue setting, the better performance of the system in terms of intrinsic evaluation was not reflected.

In the TTS experiment, a modified version of the traditional intrinsic assessment method called mean opinion score was employed. In the extrinsic evaluation, the same utterances were incorporated in a dialogue scenario implemented into a WOZ experiment. This experiment was combined with an experiment to evaluate the combination of MT and TTS. Therefore, the extrinsic and intrinsic evaluation was performed on a set of synthesised human gold standard translations and a set of synthesised machine translations. The intrinsic evaluation backed the assumption that the synthesised gold standard translations were perceived to be better than the synthesised machine translations, although the same synthesiser was used. Therefore, it was expected that in the extrinsic evaluation the interaction in the control group with the gold standard translation would be smoother than in the MT group. Again the extrinsic evaluation did not show a significant difference between results from the gold standard group compared to the results from the machine translation group, and therefore the hypothesis could not be confirmed.

The experiment in which the machine translation component was combined with the TTS component did lead to an interesting observation. It was observed that participants regularly repeated the erroneous machine translation output when answering the system's questions. They would go as far as using words in their communication which cannot be found in a lexicon. To examine this phenomenon a small scale question answering experiment with 30 participants was carried out which aimed at investigating whether people would align their answers to the wording presented in the questions or if they would correct the given alternatives

or engage in some sort of attempted repair. The results indicated that when presented with synthesised machine translation output people are indeed likely to adopt the mistakes they hear as part of their own speech and repeat these mistakes in their answers. This in turn probably has a considerable influence on the performance of the ASR system which is the component in the speech-to-speech translation system that will have to process the erroneous answers.

This chapter outlined the experiments which were carried out to investigate the research question. The next chapter will discuss the results and conclude on their implications for this thesis.

Chapter 5

Conclusion

5.1 Introduction

The evolution of speech technology can be compared to the evolution of the personal computer and graphical interfaces. At first computers were tools to carry out special computational tasks, used by a small group of people who were highly trained in their use. With the evolution of the computer to an every day appliance the group of users changed to every-day people. As a consequence the software which offered the most intuitive and usable interfaces became successful. The user and the task thus became the centre of attention of software developers. Similarly, speech technologies such as ASR, MT, and TTS have now gained a maturity where they are employed in all kinds of every-day scenarios and applications. Traditionally, the evaluation of speech technology has been mainly diagnostic and developer-driven.

A central part of this thesis is the examination of intrinsic and extrinsic NLP evaluation methods. The discussion of intrinsic and extrinsic evaluation methods in the literature has been reviewed in the Related Work Section. Intrinsic evaluation is based on measuring the performance of an isolated NLP system which is mainly characterised by the performance with respect to a gold standard result. Extrinsic evaluation considers the NLP system in a more complex setting, i.e. as integrated part of a comprehensive system or as precise function for a human user. The performance of the NLP technology is then rated with regards to its utility in the complex system or for the human user. Thus, extrinsic evaluation aims to assess the (often indirect) effect of a module on task- and context-dependent variables such as user performance.

5.2 Research Question Revisited

Since intrinsic evaluations are generalisable, repeatable and comparable from one evaluation to the other, they are the most common in NLP evaluation efforts. But these automatic metrics

seem to have reached their limits. For some applications they are simply not discriminative enough anymore. According to Jurafsky and Martin [2008], only extrinsic evaluation can indicate an improvement in performance on a real task. ASR, MT, and TTS have in common that they are traditionally evaluated intrinsically. However, all state a need for alternative (viz. extrinsic) evaluations. While ASR seems to be the component of S2S translation that is the least in need of improvements on its traditional evaluation metric (WER), MT researchers are still actively looking for an alternative to their state-of-the-art method BLEU. TTS is probably the most troublesome component to evaluate because the factors that influence the perception of speech, such as the voice for example, are highly subjective. No widely used automatic evaluation metric for TTS assessment exists, and van Santen even argues that, because of practical and fundamental reasons (e.g. the conceptual multidimensionality), speech quality can not be assessed formally [van Santen, 1997, page 242]. A serious drawback of extrinsic evaluation is the effort associated with this kind of experiments. In order to get statistically significant results a very large number of participants has to be tested, rendering extrinsic evaluation especially time and cost intensive. Since they are limited to the context of the application under test they may further not be generalisable to other applications.

Therefore, this thesis examined the potential benefits of extrinsic evaluation methods to ASR, MT, and TTS. In order to investigate this question, a series of experiments was carried out in which extrinsic and intrinsic evaluation methods were compared to each other. As a case study the three components of speech-to-speech translation were examined. It was investigated whether extrinsic evaluation results correspond to intrinsic evaluation results. The experiments have shown that this is not the case and that there are situations or setting in which intrinsic evaluation falls short. As will be argued in the following discussion, ASR, MT, and TTS can profit from extrinsic evaluations but not without several caveats. Since it has been argued that extrinsic evaluation is beneficial for NLP evaluation, this thesis further examines how such extrinsic evaluation methods can be implemented.

5.3 Summary of the Experiments

The experiment concerning the ASR component examined the correlation between extrinsic measures (i.e. time to task completion, direction scores and word ratio) and the corresponding word error rate values of sixteen ASR mediated dialogues. It has been shown that the extrinsic measures only show a minor or no correlation with the word error rate. Based on the assumption that the evaluation on word level might not be fine grained enough to reflect ASR accuracy, five variants of error rates have been implemented and their correlations with the extrinsic measures have been analysed. The results indicate that in certain situations the

correlation results improve and therefore the new proposed variants might be better suited than the word error rate. While the results only improved marginally, the improvements illustrate the fact that extrinsic evaluation can be useful in fine tuning intrinsic evaluation metrics. In addition to this general implication, the experiment revealed problems of the ASR system to recognise compounds correctly. In some cases the system decomposed compounds into their stems. In other cases words were compounded although the composition did not result in a valid compound.

Similar results were exposed by the MT experiment which employed a WOZ experiment as extrinsic evaluation method. This experiment also confirmed that depending on the context of use, intrinsic methods may not provide the most accurate results. The extrinsic evaluation also identified that the MT system runs into trouble with the translation of German compounds. The system under test did compound translations of several words into new words which are nonsensical.

One aspect of the TTS experiment was an evaluation of the combination of MT and TTS. The results showed that the synthesised gold standard translations were perceived to be better than the synthesised machine translations, although the same synthesiser was used. This experiment lead to an interesting observation. Participants regularly repeated erroneous machine translation output when answering the system's questions. They would go as far as using words in their communication which cannot be found in a lexicon.

This phenomenon was further examined in a small scale question answering experiment with 30 participants which indicated that when presented with synthesised machine translation output people are indeed likely to adopt the mistakes they hear as part of their own speech and repeat these mistakes in their answers. This behaviour in turn is likely to have a negative impact on the processing of the user response, by inducing speech recognition errors and their cascading effects on the machine translation module. The experiment therefore accentuate the necessity to consider the combination of single components in the evaluation of complex systems such as speech-to-speech translation which are traditionally assessed based on the single components or as a whole (end-to-end evaluation).

5.4 Implications

In summary, the experiments confirmed that evaluation efforts need to keep the user and application context in mind if they aim to produce meaningful results for real-life situations. The results of the three experiments comparing extrinsic and intrinsic evaluation methods were similar in that it was shown that extrinsic results are less dramatic than intrinsic results which goes hand in hand with an observation made by Molla [Molla and Hutchinson, 2003]. The human

ability to process languages is still unmatched by technology. Human conversation is full of errors, ill-formed sentences, false starts and hesitations. These complicate matters for NLP components. However, if humans are faced with such qualitative and quantitative errors they are well able to filter the relevant content out, sometimes even without noticing the mistakes. Intrinsic measures seem to be unforgiving of small mistakes and therefore often do not reflect human judgment. Intrinsic methods are mostly automatic and therefore can be considered to generate cheaper and quicker results, often being the first choice for developers. However, intrinsic evaluations seldom offer the full picture, often disregarding the user and the task to which the system is set, for instance. It has been shown that for all three components the intrinsic evaluations typically do not reflect the human perception of the performance and therefore might only be of limited value. Extrinsic evaluation has the main advantage that it offers a better insight into why a system failed which cannot be inferred from the results of intrinsic evaluation. The issues of dealing with compounds that the MT system was facing cannot be inferred from a BLEU score for example.

When gold standards are easy to define or already available, intrinsic methods are cheap and preferred. Such evaluations, however, run the risk for researchers to tune their systems towards good scores with respect to a gold standard rather than towards good system output. The proposed metric to evaluate the performance of an ASR system on basis of sounds rather than words has shown promising results in regards of the correlation with extrinsic measures and still benefits from a simple automatic approach. An important discovery made in the ASR experiment is the implication that extrinsic evaluations can possibly be used to fine-tune intrinsic evaluation metrics.

These points confirm that the evaluation of ASR, MT und TTS can benefit from extrinsic methods but with several caveats. Without doubt extrinsic methods are laborious because they involve human tests. Not only is their execution demanding, but also extrinsic evaluation produces a vast amount of data. Planning how to capture this data and deciding which data is crucial is also very difficile issue. Usually extrinsic evaluations require a large number of repetitions in order to achieve statistically significant results. Furthermore, since the user and tasks can differ greatly, a large variation of customised evaluations is necessary, leading to often isolated efforts which are not comparable, not standardised, and seldom repeated.

Both types of evaluations have their advantages and drawbacks, and the situations in which they are applied have to be considered carefully. It is also a difficult question to generalise an answer to the question when extrinsic evaluation should be applied. In summary, to find the perfect balance between intrinsic evaluation that steers the development and extrinsic evaluation putting the user at the centre of the evaluation is worth to aspire. The latter should not be

lost out of sight since NLP technology tries to imitate and understand human behaviour and is developed as tool or aid for human users.

Since it was shown that extrinsic evaluations can be beneficial for the evaluation of all components of speech-to speech translation, it is also interesting to consider how such evaluation methods can be implemented. In this research the map task and the Wizard of Oz technique were employed in the extrinsic evaluations. Both have proven to be useful tools in the evaluation of ASR, MT and TTS in interactive situations.

The map task offers a controlled environment to collect spontaneous spoken language corpora. The map task has several positive features. To give directions is a mundane task that is easily understood by the participants. Due to the task oriented nature of the map task the dialogues are constrained and therefore more predictable. Further, since the observer in the experiment shares the domain knowledge (about the map) with the participants it is easier to construe the speakers' communication intention, the knowledge about the mismatches enables the observer also to predict which parts of the route should be straightforward and which would be difficult. Since the sets of objects in the map can be freely choosen by the experimenter the map task offers also a perfect opportunity to examine all kinds of language specialties such as the use of referring expressions for instance. Experiments can be re-run with different settings on the same map material. The analysis of the map task data showed, however, that the proposed evaluation metric to assess the performance of the participants on drawing the route, hence on their task performance is impractical and missleading. Therefore, a new assessment method has been introduced that is much simpler while still being discriminative enough to determine the amount of miscommunication.

The Wizard of Oz method is a well established early stage prototyping method and has a long history in the design of speech applications, being a powerful method when the input modality has a high computation/cognition ratio. However, it is seldom employed to evaluate NLP systems. The fundamental advantage of the WOZ technique is that researchers can simulate a functionality of a system that has not been implemented yet and it can even be applied if the proposed system goes beyond what is feasible. It can further be used to deliberately introduce mistakes in order to investigate the effect of these mistakes on the participants. But Wizard of Oz experiments are in many cases materially demanding when it comes to the set-up in order to create the illusion of a running system. The developement of a general WOZ prototyping platform (e.g. *WebWoz* [Schlögl et al., 2011]) will hopefully alleviate this problem in future. An important aspect to ensure meaningful results in Wizard of Oz studies is to ensure that the wizard acts consistently which also is a difficult task.

5.5 Research Outlook

This thesis focussed on the three components of S2S translation. However, the findings of this thesis can also be applied to other interactive, possibly multilingual, speech applications such as spoken language translation or (multilingual) dialogue applications.

Several aspects have been discovered that are worth further investigation but to do so would be beyond the scope of this thesis.

As already discussed, the number of evaluation metrics to choose from can be vast, as can be the ammount of collected data. Therefore, rational decisions have to be made which metrics will be used to analyse the experiment results. The extrinsic experiment around automatic speech recognition for example could probably profit from a quantitative analysis of situations in which the communication between instructor and follower failed. However, this would have required the development of an annotation scheme and the annotation (possibly with more than one annotator) of communication problems in the dialogues. In order to be discriminative this evaluation would also need a larger number of participants. The MT experiment could also be repeated with extra errors added to the output in order to investigate the issue of user acceptance versus error rate.

An interesting observation that was made in all experiments is the difficulty of ASR, MT and TTS to cope with compounds. These difficulties seem to be recognized by the different research fields. In machine translation, for instance, the decomposition of compounds is widely explored (e.g. [Koehn and Knight, 2003] and [Popovic et al., 2006]), whereas less literature can be found on merging words into compounds (c.f. [Stymne, 2009]). Still compounds are an unsolved, yet very interesting problem that is not unique to the German language. There are many other languages that have compounds, for example Spanish, Dutch, Icelandic, Finish, just to name a few.

The experiments that were described in this thesis, except the alignment study were run with a number of participants that was sufficient to test the hypotheses of the experiments presented in Chapter 4. However, for many of the different aspects (e.g. compounding mistakes in ASR and their influence on the perception) that were discovered in these experiments the collected data was to small to enable differences to be tested statistically. To thoroughly proof new hypotheses around such findings, new experiments would have to be run with a much bigger set of participants. The alignment study was carried out with 30 participants and confirmed alignment effects in MT mediated settings which was an interesting aspect within this research. However, it is interesting to investigate in greater detail in which cases people are more likely to align to mistakes. This would require a more comprehensive experiment with a bigger set of sentences under test, to which errors are introduced systematically.

Appendix A

List of Abbrevations

ASR	Automatic Speech Recognition
BLEU	BiLingual Evaluation Understudy
BP	Brevity Penalty
CER	Character Error Rate
CMU	Carnegie Mellon University
CSR	Continuous Speech Recognition
EBMT	Example Based Machine Translation
DARPA	Defense Advanced Research Project Agency
DMOS	Degradation Mean Opinion Score
DCIEM	Defence and Civil Institute of Environmental Medicine
GS	Gold Standard
GALE	Global Autonomous Language Exploitation
HCI	Human-Computer Interaction
HCRC	Human Communication Research Centre
HMM	Hidden Markov Model
IDA	Institute for Defense Analysis
ITU	International Telecommunication Union
IWR	Isolated Word Recognition
IWSLT	International Workshop on Spoken Language Translation
MOS	Mean Opinion Score
MER	Match Error Rate
MRT	Modified Rhyme Test
MT	Machine Translation

NEC	Nippon Electrical Corporation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
PSOLA	Pitch-Synchronous Overlap and Add
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
S2S	Speech-to-Speech
SASSI	Subjective Assessment of Speech System Interfaces
SDS	Spoken Dialogue System
SLT	Spoken Language Translation
SoundER	Sound Error Rate
SPIN	SPeech In Noise
SylER	Syllable Error Rate
TER	Translation Error Rate
TIDES	Translingual Information Detection Extraction, Summarization
TREC	Text REtrieval Conference
TTS	Text-to-Speech
WER	Word Error Rate
WIP	Word Information Preserved
WOZ	Wizard of Oz

Appendix B

Materials of the ASR experiments

landmark	direction features	1	2	3	4	5	6	7	8
glasses	↑→↓	3	3	3	3	3	3	3	0
suitcase	→↑	2	2	2	2	2	2	2	0
stapler	↑←	2	2	2	2	2	2	2	2
suitcase 2	←↑→	3	3	3	2	3	3	3	0
coffee cup	→↑←	3	3	3	3	3	3	3	0
flipchart	←↑→	3	0	3	2	2	3	0	3
calendar	→↑	2	0	2	2	1	0	2	2
finish	↑	1	1	1	1	1	0	1	1
overall	19	19	13	19	17	18	16	16	8
penalty	-	0	6	0	5	0	2	4	10
sum	-	19	7	19	12	18	14	12	-2

Table B.1 – Overview of the direction feature calculations for the office map.

landmark	direction features	1	2	3	4	5	6	7	8
cowboy	←↑→	3	3	3	0	3	3	3	3
indian boy	→↑←	3	3	3	0	3	3	3	3
cactus	←↑→	3	3	3	0	3	3	3	3
hanging tree	→↑	1	2	2	0	2	2	2	2
tepee	↑→	2	2	2	0	2	2	2	2
wagon	→↑←	3	3	3	0	3	3	3	2
oil rig	←↑	2	2	2	2	2	2	2	0
horses	↑←↓	3	3	3	3	3	3	3	3
cowboy hat	↓←↑	3	3	3	3	3	3	3	3
saloon	↑←↓	3	3	3	3	3	3	3	0
finish	↓	1	1	1	1	1	0	0	1
overall	28	27	28	28	12	28	27	27	22
penalty	-	2	2	0	11	0	0	1	3
sum	-	25	26	28	1	28	27	26	19

Table B.2 – Overview of the direction feature calculations for the wild west map.

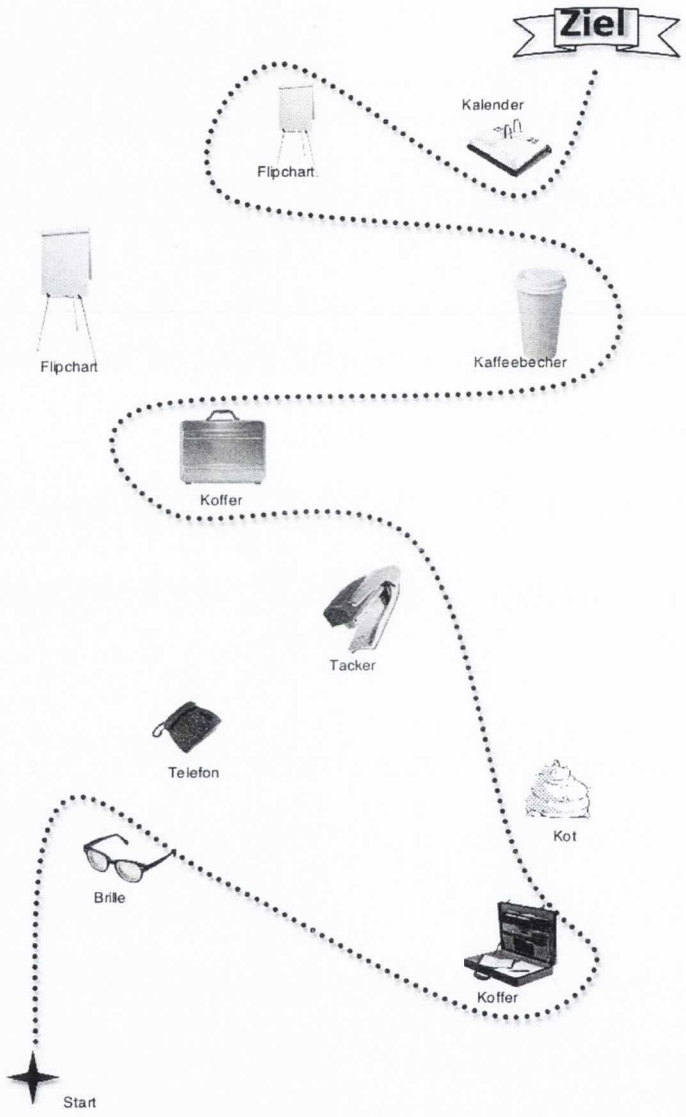


Figure B.1 – The map that was presented to the instruction giver in the office scenario in the extrinsic ASR evaluation (map task).

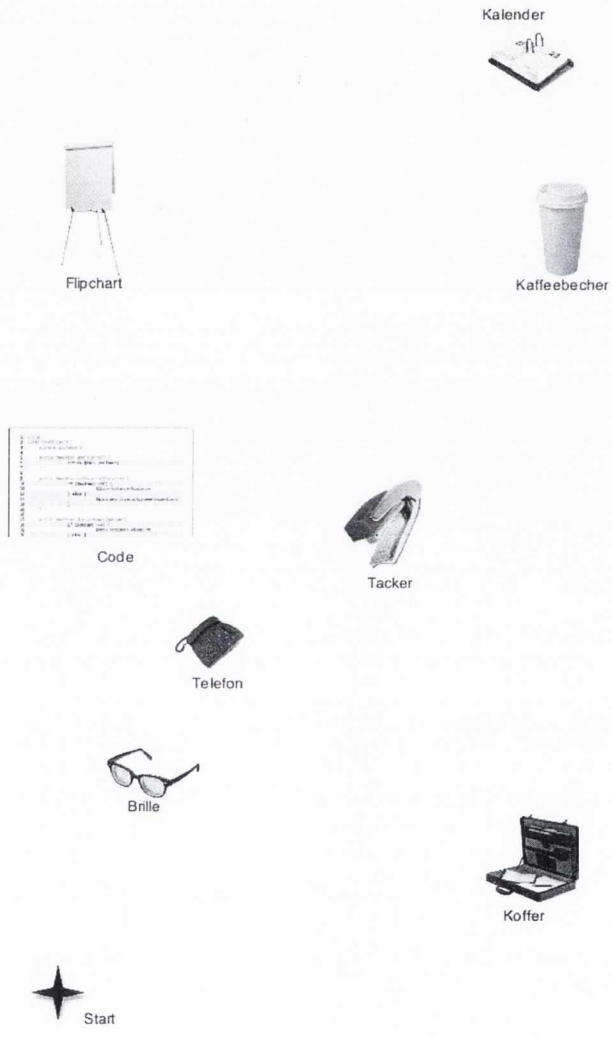


Figure B.2 – The map that was presented to the follower in the office scenario in the extrinsic ASR evaluation (map task).

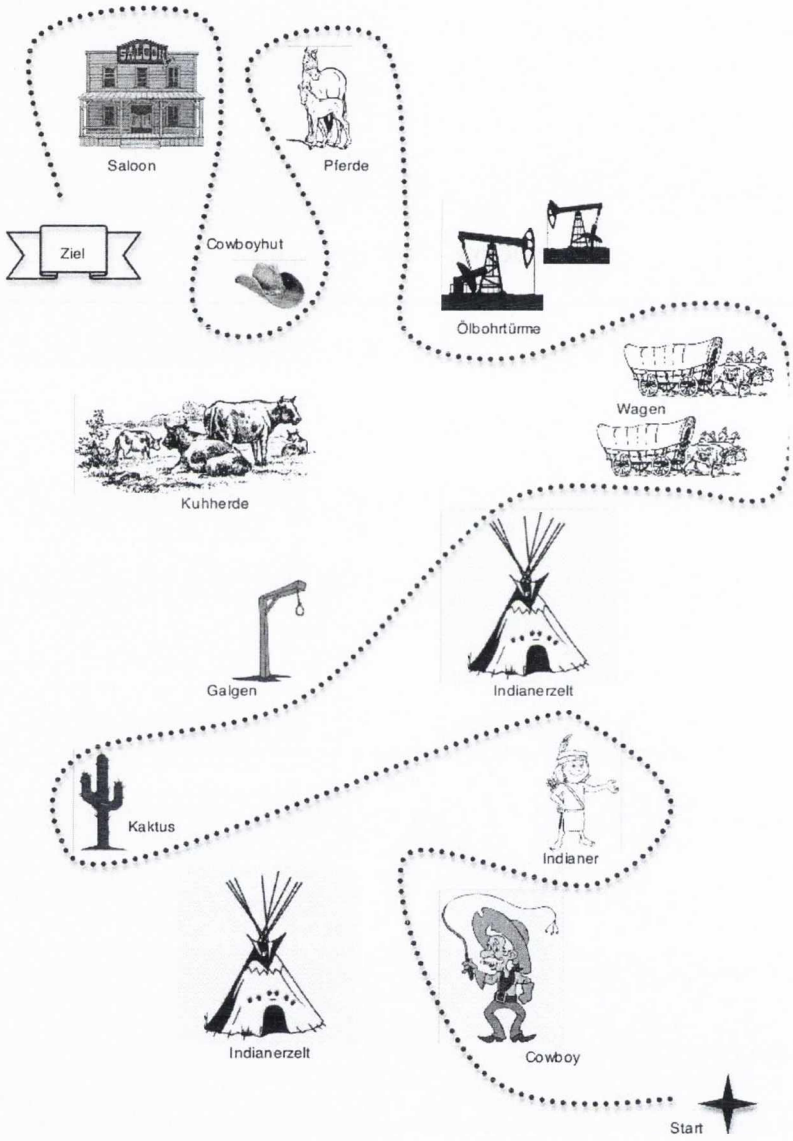


Figure B.3 – The map that was presented to the instruction giver in the wild west scenario in the extrinsic ASR evaluation (map task).

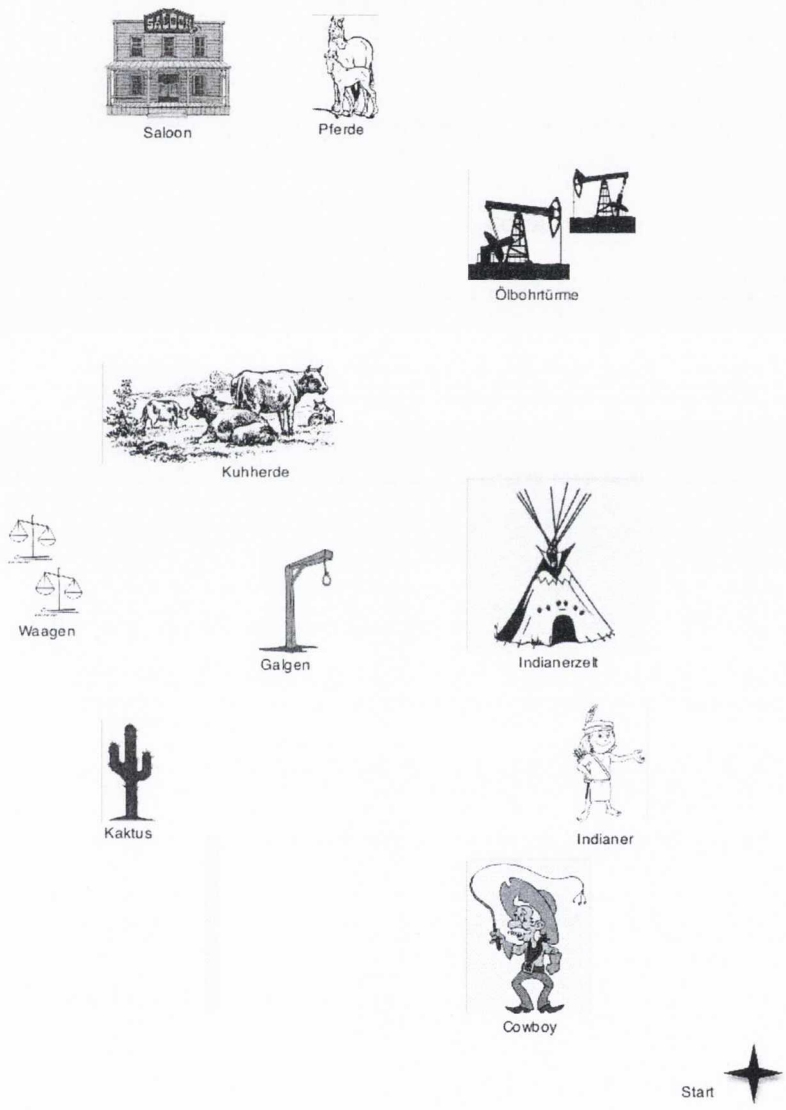


Figure B.4 – The map that was presented to the follower in the wild west scenario in the extrinsic ASR evaluation (map task).

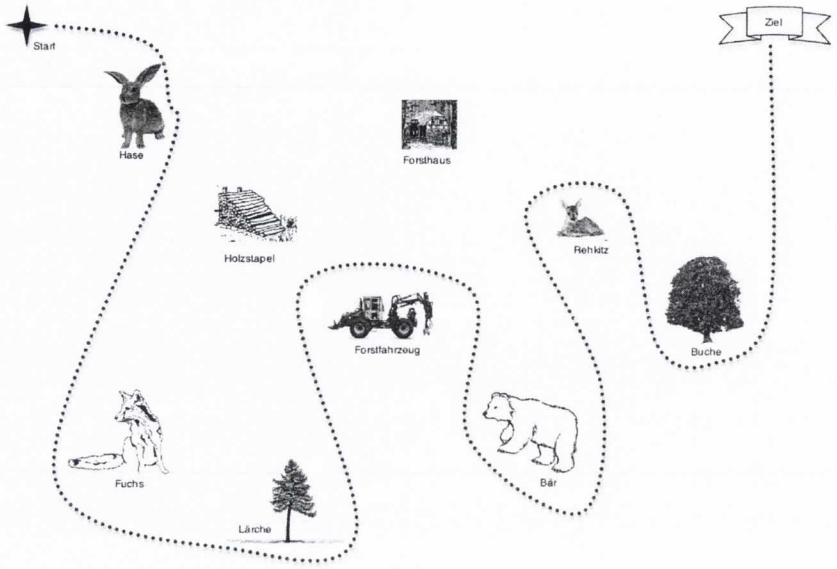


Figure B.5 – The map that was presented to the instruction giver in the forest scenario in the extrinsic ASR evaluation (map task).

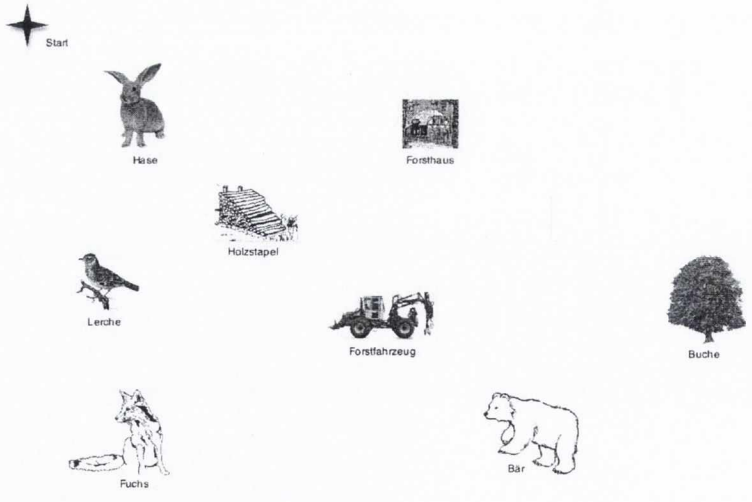


Figure B.6 – The map that was presented to the follower in the forrest scenario in the extrinsic ASR evaluation (map task).

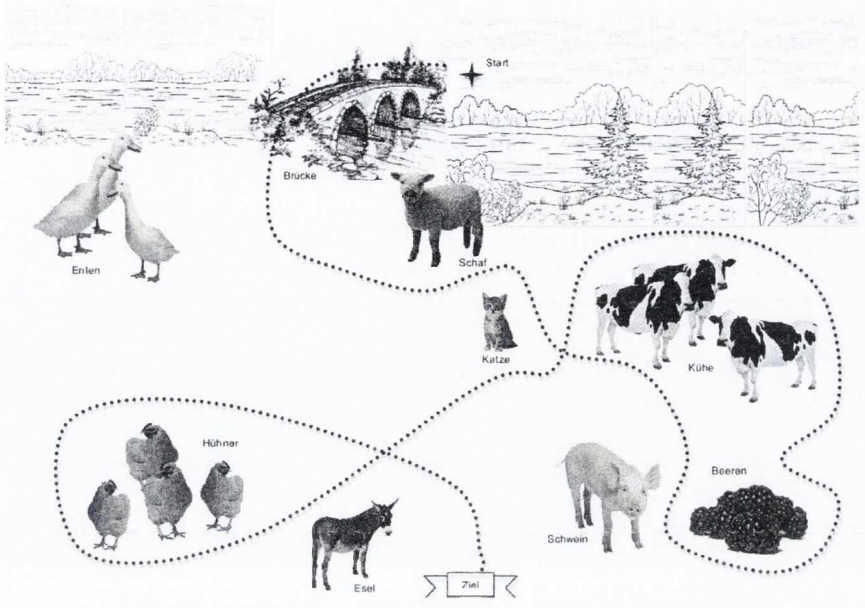


Figure B.7 – The map that was presented to the instruction giver in the river scenario in the extrinsic ASR evaluation (map task).

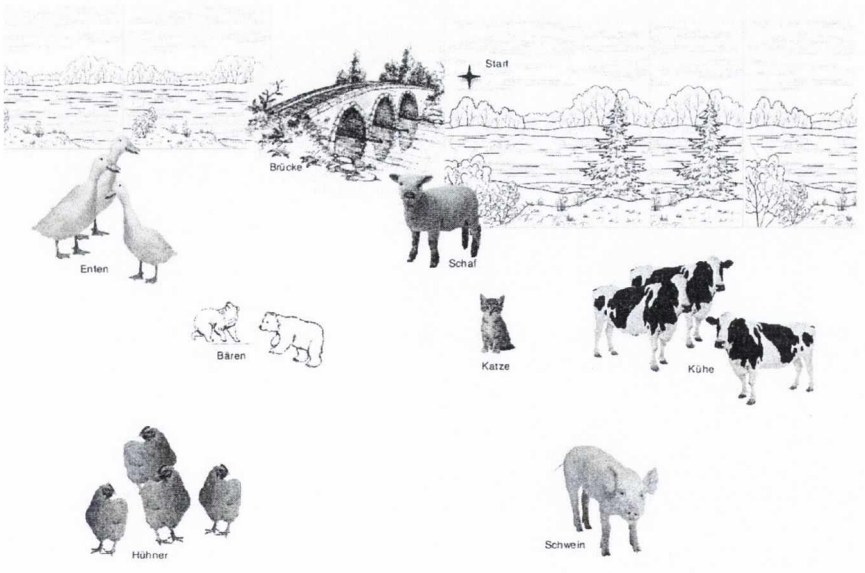


Figure B.8 – The map that was presented to the follower in the river scenario in the extrinsic ASR evaluation (map task).

landmark	direction features	1	2	3	4	5	6	7	8
rabbit	→↓	2	2	2	2	2	2	2	0
fox	↓→	0	2	2	0	2	2	2	2
larch	→↑	2	0	0	2	2	2	2	2
forestry vehicle	↑→↓	0	2	2	2	3	3	3	3
bear	↓→↑	0	3	0	0	3	3	3	3
fawn	↑→↓	3	3	3	3	3	3	3	0
beech tree	↓→↑	3	3	3	3	3	3	3	3
finish	↑	1	1	1	1	1	1	1	1
overall	18	10	16	13	13	18	18	18	14
penalty	-	8	4	1	3	0	0	0	4
sum	-	2	12	12	10	18	18	18	10

Table B.3 – Overview of the direction feature calculations for the forest map.

landmark	direction features	1	2	3	4	5	6	7	8
bridge	←↓	2	2	2	2	2	2	2	2
sheep	→	1	1	1	1	0	1	1	0
cat	→↓	2	2	2	2	2	2	2	0
pig	→↓	2	2	2	2	1	2	1	2
berries	↓→↑	3	3	3	3	3	3	0	0
cows	↑←↓	3	3	3	3	3	3	3	3
cat	←	1	1	1	1	1	0	0	0
donkey	←↓	2	2	2	2	1	0	0	2
chicken	←↑→	3	3	3	3	2	3	3	3
donkey	→	1	1	1	1	1	1	0	0
finish	↓	1	1	1	1	1	1	1	1
overall	21	21	21	21	21	17	18	13	13
penalty	-	0	0	0	0	4	0	0	5
sum	-	21	21	21	21	13	18	13	8

Table B.4 – Overview of the direction feature calculations for the river map.

	1	2	3	4	5	6	7	8
mode	ASR				phone			
office	6:57	18:05	5:18	27:31	5:40	5:40	4:05	2:26
forest	6:19	12:27	13:45	9:52	14:14	2:52	4:27	2:22
mode	phone				ASR			
wild west	4:17	5:27	3:45	5:14	24:13	9:10	11:35	7:12
river	6:17	4:15	2:00	8:25	21:48	9:26	6:53	5:26

Table B.5 – Overview of the task completion times for each of the eight sessions ordered by maps.

	1	2	3	4	sum	5	6	7	8	sum
mode			ASR					phone		
office	417	1085	318	1651	3471	340	340	245	146	1071
forest	379	747	825	592	2543	854	172	267	142	1435
sum overall	796	1832	1143	2243	6014	1194	512	512	288	2506
mode			phone					ASR		
wild west	257	327	225	314	1123	1453	550	695	432	3130
river	377	255	120	505	1257	1308	566	413	326	2613
sum	634	582	345	730	2291	2761	1116	1108	758	5743
sum overall	1430	2414	1488	2973	-	3955	1628	1620	1046	-

Table B.6 – Overview of the task completion times in seconds for each of the eight sessions ordered by maps.

	map	# words instructor	# words follower	# words sum	time	words per sec (instr)	words per sec (follower)	words per second (sum)
1	office	314	374	688	417	0,753	0,897	1,650
	forest	310	315	625	379	0,818	0,831	1,649
	wild west	486	96	582	257	1,891	0,374	2,265
	river	726	231	957	377	1,926	0,613	2,538
2	office	440	623	1063	1085	0,406	0,574	0,980
	forest	302	700	1002	747	0,404	0,937	1,341
	wild west	598	378	976	327	1,829	1,156	2,985
	river	486	167	653	255	1,906	0,655	2,561
3	office	152	298	450	318	0,478	0,937	1,415
	forest	337	763	1100	825	0,408	0,925	1,333
	wild west	302	158	460	225	1,342	0,702	2,044
	river	246	59	305	120	2,050	0,492	2,542
4	office	774	941	1715	1651	0,469	0,570	1,039
	forest	287	254	541	592	0,485	0,429	0,914
	wild west	585	143	728	314	1,863	0,455	2,318
	river	781	421	1202	505	1,547	0,834	2,380
5	office	681	98	779	340	2,003	0,288	2,291
	forest	1080	329	1409	854	1,265	0,385	1,650
	wild west	524	1786	2310	1453	0,361	1,229	1,590
	river	917	498	1415	1308	0,701	0,381	1,082
6	office	499	200	699	340	1,468	0,588	2,056
	forest	235	106	341	172	1,366	0,616	1,983
	wild west	284	589	873	550	0,516	1,071	1,587
	river	148	278	426	566	0,261	0,491	0,753
7	office	434	252	686	245	1,771	1,029	2,800
	forest	452	230	682	267	1,693	0,861	2,554
	wild west	421	619	1040	695	0,606	0,891	1,496
	river	265	212	477	413	0,642	0,513	1,155
8	office	183	89	272	146	1,253	0,610	1,863
	forest	235	80	315	142	1,655	0,563	2,218
	wild west	250	333	583	432	0,579	0,771	1,350
	river	140	197	337	326	0,429	0,604	1,034

Table B.7 – Overview of the number of words uttered and time, and number of words uttered per second.

reference word	ASR output
Haufen Kot	haufenkot (7 times)
geschlossenen Koffer	Geschlossenenkoffer (5 times)
richtung salon	richtungssaloon
richtung katze	richtungskatze
richtung hühner	richtungshühner
richtung kaktus	richtungskaktus
Indianerzelt	Indianer Zelt (7 times)
Linksbogen	links Bogen (3 times)
Vorderbeinen	vorder Beinen (2 times)
Tacker	Acker (8), Tag Herr (3), Packer (3), Targa (3), Tage (2), Tag der (2), Hacker (2), AK (2), Tag, Tag war, Shaker, Trucker, starker, intakter, Tanker Forstfahrzeug vors Fahrzeug (3), befasst Fahrzeug, Forst vorzeigt
nordöstlich	nordwestlich (8)
Start	Staat (4)
Koffer	kupfer, korfu, phosphor
Tackers	ackers, tankers (2)
Hasen	hafen (3), phasen
lärche	lehrreiche (2), lehrreicher (3), lehrbücher
beeren	werden (2), bergen, mehreren
Rehkitz	reh kit, weg jetzt, high kids, seekids, rehkids (2), reh kids
ja	jahr (4)
Bär	bea
Bären	näheren, deren (3), wären
Reh	br (3), reha (3), de
Buche	bucher (5), buchung
östlich	westlich (5)
nördlich	örtlich, natürlich
osten	westen, resten
nordöstlich	nordwestlich (8)
östlich	westlich
südöstlich	südwestlich
nordosten	nordresten
harter	harper (2)
Galgen	geigen (3), geil (2), geilsten, geil gehen
Ölbohrturm	labor römer, bild board form, göteborg toner , Öl bob thoma (2), bill board schon (2)
Turm	tor (3)
Saloon	Iserlohn (2), sein lohn, salon(5)
Saloons	salomons, lohns

Table B.8 – Overview of the most frequent ASR mistakes. The number in brackets indicates the number of occurances of the mistake.

reference word	ASR output
Kuh	coup, kur, Q
Kühe	tür (2), kühle
Hund	rund(2)
Hahn	hand (2), haben
kuhjunge	cool junge (3)
nördlich	nordlicht (5)
Pferden	fair den, serben, verden (2)
jedenfalls du	jedenfallsdu
über der brille	Überbrille
blatt papier	blattpapier
verwirrend und	verwirrendund
osten nach	ostennach
jetzt nach	jetztnach
richtig tropfen	richtigtropfen
dort sechs	dortsechs

Table B.9 – Overview of the most frequent ASR mistakes (part2). The number in brackets indicates the number of occurances of the mistake.

Appendix C

Materials of the MT experiments

English source	German gold standard	*Google* translation	*Systran* translation
Hello, how can I help you?	Hallo, wie kann ich Ihnen helfen?	Hallo, wie kann ich Ihnen helfen?	Hallo, wie kann ich Ihnen helfen?
Sorry, I did not understand you. Could you say that again?	Bitte entschuldigen Sie, ich habe Sie nicht verstanden. Könnten Sie das noch einmal sagen?	Sorry, ich habe Sie nicht verstanden. Können Sie sagen, das wieder?	Traurig, verstand ich Sie nicht. Konnten Sie das wieder sagen?
I get the impression that you do not understand me. Should we go back a step?	Ich habe den Eindruck, dass Sie mich nicht verstehen. Sollten wir einen Schritt zurück gehen?	Ich habe den Eindruck, dass Sie mich nicht verstehen. Sollen wir gehen einen Schritt zurück?	Ich erhalte den Eindruck, dass Sie mich nicht verstehen. Sollten wir zurück gehen ein Schritt?
Do you want to start over again?	Möchten Sie noch einmal von vorne beginnen?	Wollen Sie wieder von vorn anfangen?	Möchten Sie vorbei wieder beginnen?
You want information on how to connect to the Internet in Ireland.	Sie wollen Informationen darüber, wie man an das Internet in Irland angeschlossen wird.	Sie möchten Informationen über die Verbindung zum Internet in Irland.	Sie wünschen Informationen über, wie man an das Internet in Irland anschließt.
Are you looking for mobile Internet or for a landline connection?	Suchen Sie nach einer mobilen Internetverbindung oder nach einer Festnetzverbindung?	Suchen Sie das mobile Internet oder für einen Festnetzanschluss?	Suchen Sie nach beweglichem Internet oder für eine Überlandleitungverbindung?
You are looking for a mobile Internet connection.	Sie suchen nach einer mobilen Internetverbindung.	Sie suchen eine mobile Internetverbindung.	Sie suchen nach einem beweglichen Internetanschlusse.
You are looking for an Internet connection at your home.	Sie suchen nach einer Internetverbindung bei Ihnen zu Hause.	Sie sind auf der Suche nach einer Internet-Verbindung bei Ihnen zu Hause.	Sie suchen nach einem Internetanschlusse an Ihrem Haus.
Is a twelve month contract an option for you or do you want to be more flexible?	Ist ein Zwölf-Monats-Vertrag eine Option für Sie oder möchten Sie flexibler sein?	Ist eine zwölfmonatige Vertrag eine Option für Sie oder wollen Sie mehr Flexibilität?	Ist ein zwölf-Monats-Vertrag eine Wahl für Sie, oder möchten Sie flexibler sein?
You are looking for a mobile Internet connection without a twelve month contract.	Sie suchen nach einer mobilen Internetverbindung ohne einen Zwölf-Monats-Vertrag.	Sie suchen eine mobile Internet-Verbindung ohne eine zwölfmonatige Vertrag.	Sie suchen nach einem beweglichen Internetanschlusse ohne einen zwölf-Monats-Vertrag.

Table C.1 – The first part of the system utterances that were the basis for the evaluation in the MT experiment series. First the English source, than the German gold standard translation, followed by the *Google* and *Systran* translations. The § symbole denotes places which have to be filled with slot fillers like numbers.

English source	German gold standard	*Google* translation	*Systran* translation
You are looking for a mobile Internet connection and a twelve month contract is an option for you.	Sie suchen nach einer mobilen Internetverbindung und ein Zwölf-Monats-Vertrag ist eine Option für Sie.	Sie suchen eine mobile Internetverbindung und ein zwölf Monats-Vertrag ist eine Option für Sie.	Sie suchen nach einem beweglichen Internetanschlusse und ein zwölf-Monats-Vertrag ist eine Wahl für Sie.
There are § different offers with varying download speeds, download allowances and prices.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten, Download-Grenzen und Preisen.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten, download Zertifikate und Preise.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-geschwindigkeiten, Downloadzulagen und Preisen.
There are § different offers with varying download speeds and prices.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten und Preisen.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten und Preise.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-geschwindigkeiten und -preisen.
There are § different offers with varying download allowances and prices.	Es gibt § verschiedene Angebote mit unterschiedlichen Download-Grenzen und Preisen.	Es gibt § verschiedene Angebote mit unterschiedlichen Download Zertifikate und Preise.	Es gibt § verschiedene Angebote mit unterschiedlichen Downloadzulagen und -preisen.
There are § different offers with varying prices.	Es gibt § verschiedene Angebote mit unterschiedlichen Preisen.	Es gibt § verschiedene Angebote mit unterschiedlichen Preisen.	Es gibt § verschiedene Angebote mit unterschiedlichen Preisen.
Do you have any other requirements?	Haben Sie irgendwelche anderen Anforderungen?	Haben Sie andere Wünsche?	Haben Sie irgendwelche anderen Anforderungen?
These are the options you have:	Dies sind die Optionen, die Sie haben:	Dies sind die Optionen, die Sie haben:	Diese sind die Wahlen, die Sie haben:
A twelve month contract from § with up to § Megabyte download speed and up to § Kilobyte upload speed would be a suitable option for you.	Ein Zwölf-Monats-Vertrag von § mit bis zu § Megabyte Download-Geschwindigkeit und bis zu § Kilobyte Upload-Geschwindigkeit wäre eine passende Option für Sie.	Ein zwölf Monats-Vertrag von § mit bis zu § Megabyte Download-Geschwindigkeit und bis zu § Kilobyte Upload-Geschwindigkeit wäre ein geeigneter Weg für Sie.	Ein zwölf-Monats-Vertrag vom § mit bis zu § Megabyte-Downloadgeschwindigkeit und bis zu # Kilobyteantriebskraft-geschwindigkeit würde eine passende Wahl für Sie sein.

Table C.2 – Second part of the system utterances that were the basis for the evaluation in the MT experiment series. Ffirst the English source, than the German gold standard translation, followed by the *Google* and *Systran* translations. The § symbole denotes places which have to be filled with slot fillers like numbers.

English source	German gold standard	*Google* translation	*Systran* translation
A pay as you go option from § with up to § Megabyte download speed and up to § Kilobyte upload speed is what you are looking for.	Eine pre-paid Option von § mit bis zu § Megabyte Download-Geschwindigkeit und bis zu § Kilobyte Upload-Geschwindigkeit ist wonach Sie suchen.	Ein Pay as you go Option aus § mit bis zu § Megabyte Download-Geschwindigkeit und bis zu § Kilobyte Upload-Geschwindigkeit ist das, was Sie suchen.	Ein Lohn, da Sie Wahl von § mit bis zu § Megabytedownloadgeschwindigkeit gehen und Antriebskraftgeschwindigkeit von bis § Kilobyte ist, nach was Sie suchen.
There is a § Gigabyte download limit.	Es gibt dabei eine § Gigabyte Download-Grenze.	Es ist ein § Gigabyte Download begrenzen.	Es gibt eine § -Gigabyte-Downloadgrenze.
There is no download limit.	Es gibt dabei keine Download-Grenze.	Es ist kein Download begrenzen.	Es gibt keine Downloadgrenze.
This offer will cost § Euro for the year with a modem included.	Dieses Angebot kostet § Euro für das Jahr mit einem Modem inklusive.	Dieses Angebot kostet § Euro für das Jahr mit einem Modem enthalten.	Dieses Angebot kostet § Euro für das Jahr mit einem eingeschlossenen Modem.
This offer will cost § Euro for the year with a modem and line rental included.	Dieses Angebot kostet § Euro für das Jahr mit einem Modem und einem Festnetzanschluss inklusive.	Dieses Angebot kostet § Euro für das Jahr mit einem Modem und Miete inbegriffen.	Dieses Angebot kostet § Euro für das Jahr mit einem Modem und einer Linie die eingeschlossene Miete.
This offer will cost § Euro for six months with a modem included.	Dieses Angebot kostet § Euro für sechs Monate mit einem Modem inklusive.	Dieses Angebot wird § Euro für sechs Monate mit einem Modem Kosten enthalten.	Dieses Angebot kostet § Euro für sechs Monate mit einem eingeschlossenen Modem.
Sorry, I need some time to find this information.	Bitte entschuldigen Sie, ich brauche etwas Zeit diese Information zu finden.	Sorry, ich brauche etwas Zeit, um diese Informationen zu finden.	Traurig, benötige ich einige Zeit, diese Informationen zu finden.

Table C.3 – Third part of the system utterances that were the basis for the evaluation in the MT experiment series. First the English source, than the German gold standard translation, followed by the *Google* and *Systran* translations. The § symbole denotes places which have to be filled with slot fillers like numbers.

English source	German gold standard	*Google* translation	*Systran* translation
Sorry, but I have no information on that topic.	Bitte entschuldigen Sie, aber ich habe keine Informationen zu diesem Thema	Sorry, aber ich habe keine Informationen zu diesem Thema.	Traurig, aber ich haben Sie keine Informationen über dieses Thema.
And finally:	Und schließlich:	Und schließlich:	Und schließlich:
Is there anything else I can do for you?	Gibt es noch etwas, dass ich für Sie tun kann?	Gibt es etwas, sonst kann ich für Sie tun?	Gibt es noch etwas, die ich für Sie tun kann?
I hope the information I have given you was helpful. Have a good day.	Ich hoffe, dass die Information, die ich Ihnen gegeben habe war hilfreich. Haben Sie einen guten Tag.	Ich hoffe, die Informationen, die ich habe Ihnen hilfreich war. Have a good day.	Ich hoffe, dass die Informationen, die ich Sie gegeben habe, hilfreich waren. Haben Sie einen guten Tag.
You have the option to sign on to pay as you go where you do not have to sign a contract. Usually this option is more expensive per unit, but you have the option to stop using the services whenever you like.	Sie haben die Option sich für Prepaid anzumelden, wobei Sie keinen Vertrag unterschreiben müssen. Gewöhnlich ist diese Option pro Einheit teurer, aber Sie haben die Möglichkeit jederzeit damit aufzuhören diesen Service zu nutzen.	Sie haben die Möglichkeit, auf Zeichen zu zahlen, wie Sie dorthin, wo Sie nicht auf einen Vertrag zu unterzeichnen. Normalerweise ist diese Option teurer pro Stück, aber Sie haben die Möglichkeit, die Nutzung der Services, wann immer Sie möchten.	Sie haben die Wahl, zum an zu unterzeichnen, um zu zahlen, während Sie gehen, wohin Sie einen Vertrag nicht unterzeichnen müssen. Normalerweise ist diese Wahl pro Einheit teurer, aber Sie haben die Wahl, zum die, Dienstleistungen zu verwenden zu stoppen, wann immer Sie mögen.
Do you, for example, prefer a lower price to the download speed? Or is price not an issue?	Bevorzugen Sie beispielsweise einen geringeren Preis gegenüber der Download-Geschwindigkeit? Oder ist der Preis kein Thema?	Haben Sie zum Beispiel lieber einen niedrigeren Preis, um die Download-Geschwindigkeit? Oder ist der Preis kein Thema?	Bevorzugen Sie zum Beispiel einen niedrigeren Preis gegenüber der Downloadgeschwindigkeit? Oder ist Preis nicht eine Frage?

Table C.4 – Fourth part of the system utterances that were the basis for the evaluation in the MT experiment series. First the English source, than the German gold standard translation, followed by the *Google* and *Systran* translations. The § symbole denotes places which have to be filled with slot fillers like numbers.

Alter: _____

Geschlecht:
[ ] männlich
[ ] weiblich

Bildungsgrad:
[ ] Student im ___Jahr
[ ] PhD/Doktorand im ___Jahr
[ ] sonstiges: _____

Art des Studiums:
[ ] Kunst, Geistes-, Sozialwissenschaften
[ ] Naturwissenschaften (Mathematik, Biologie, ..)
[ ] Gesundheitswissenschaften (Medizin, Zahnmedizin, ..)
[ ] sonstige: _____

Muttersprache:
[ ] Deutsch
[ ] Englisch
[ ] sonstige: _____

Bitte beurteilen Sie ihre Computerkenntnisse:
[ ] schlecht
[ ] unterdurchschnittlich
[ ] durchschnittlich
[ ] gut
[ ] exzellent

Bitte umblättern

Bitte beurteilen Sie ihre Erfahrungen mit Dialogsystemen:
[ ] kaum
[ ] wenig
[ ] durchschnittlich
[ ] viel
[ ] sehr viel

Bitte beurteilen Sie Ihre Englischkenntnisse:
[ ] schlecht
[ ] unterdurchschnittlich
[ ] durchschnittlich
[ ] gut
[ ] exzellent

Welche Hilfsmittel nutzen Sie (wenn überhaupt) um von Deutsch nach Englisch (und umgekehrt) zu übersetzen? *(Bitte ordnen Sie diese nach Häufigkeit von 1 (sehr häufig) bis 5 (sehr selten). Mehrfachantworten sind möglich.)*
[ ] Wörterbücher (aus Papier)
[ ] Online-Wörterbücher (e.g., `http://dict.leo.org/`)
[ ] Internet-Übersetzungssysteme
[ ] ich frage bilinguale Personen
[ ] sonstiges: _____

Ich habe mich bereits mit der Problematik in Irland einen Internetanschluss zu bekommen auseinander gesetzt.
[ ] Ja
[ ] Nein

Figure C.1 – The demographic questionnaire that was presented to participants at the beginning of the extrinsic MT evaluation.

Age: _____

Gender:
[ ] male
[ ] female

Level of education:
[ ] student in the ___ year
[ ] PhD in the ___ year
[ ] other: _____

Type of education:
[ ] arts, humanities, social sciences
[ ] engineering, mathematics, computer science, natural sciences
[ ] health sciences
[ ] other: _____

native language:
[ ] German
[ ] English
[ ] other: _____

Please judge your familiarity with Computers:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

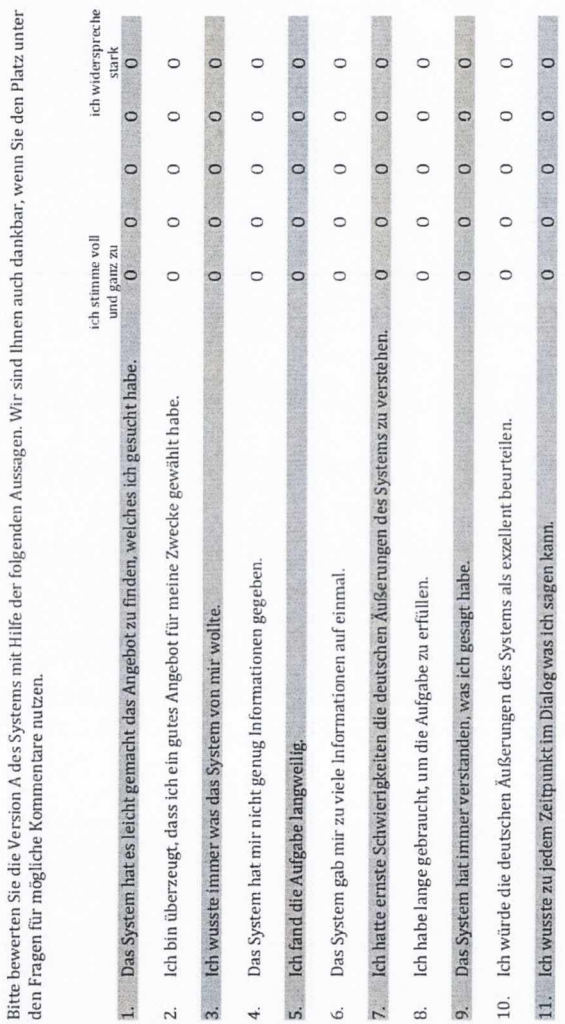
Please judge your familiarity with Spoken Dialogue systems:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

Please judge your English capabilities:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

What sort of language aids (if any) do you use for translations from English into German and vice versa? (Rank the options according to frequency of use. 1 means very often and 5 seldom. Several things can be ranked the same)
[ ] paper dictionaries
[ ] online dictionaries
[ ] internet machine translators
[ ] ask a bilingual person
[ ] other _____

I was already concerned with the topic of getting an Internet connection in Ireland before I took part in this study.
[ ] yes
[ ] no

Figure C.2 – The English version of the demographic questionnaire.



Bitte bewerten Sie die Version A des Systems mit Hilfe der folgenden Aussagen. Wir sind Ihnen auch dankbar, wenn Sie den Platz unter den Fragen für mögliche Kommentare nutzen.

	ich stimme voll und ganz zu				ich widerspreche stark
1. Das System hat es leicht gemacht das Angebot zu finden, welches ich gesucht habe.	O	O	O	O	O
2. Ich bin überzeugt, dass ich ein gutes Angebot für meine Zwecke gewählt habe.	O	O	O	O	O
3. Ich wusste immer was das System von mir wollte.	O	O	O	O	O
4. Das System hat mir nicht genug Informationen gegeben.	O	O	O	O	O
5. Ich fand die Aufgabe langweilig.	O	O	O	O	O
6. Das System gab mir zu viele Informationen auf einmal.	O	O	O	O	O
7. Ich hatte ernste Schwierigkeiten die deutschen Äußerungen des Systems zu verstehen.	O	O	O	O	O
8. Ich habe lange gebraucht, um die Aufgabe zu erfüllen.	O	O	O	O	O
9. Das System hat immer verstanden, was ich gesagt habe.	O	O	O	O	O
10. Ich würde die deutschen Äußerungen des Systems als exzellent beurteilen.	O	O	O	O	O
11. Ich wusste zu jedem Zeitpunkt im Dialog was ich sagen kann.	O	O	O	O	O
12. Die Antworten des Systems stimmten mit meiner Erwartung überein.	O	O	O	O	O

Figure C.3 – The German version (page one) of the questionnaire about the interaction which had to be filled out by the participants after each task was fulfilled. Extra space was left below the questions for comments.

	ich stimme voll und ganz zu				ich widerspreche stark
13. Im deutschen Text gab es merkwürdige Worte und Formulierungen.	O	O	O	O	O
14. Ich habe das System gerne benutzt.	O	O	O	O	O
15. Ich konnte schnell finden wonach ich gesucht habe.	O	O	O	O	O
16. Die deutschen Äußerungen des Systems waren fließend.	O	O	O	O	O
17. Das System gab mir viele unnütze Informationen.	O	O	O	O	O
18. Die Antworten des Systems waren angemessen.	O	O	O	O	O
19. Das System war sehr freundlich.	O	O	O	O	O
20. Ich würde das englische Original bevorzugen.	O	O	O	O	O
21. Ich würde ein ähnliches System nutzen um ein gutes Breitband-Internet-Angebot zu finden.	O	O	O	O	O
22. Mit dem System umzugehen war wirklich nervig.	O	O	O	O	O
23. Ich würde die deutschen Äußerungen des Systems als unverständlich bewerten.	O	O	O	O	O

Haben Sie allgemeine Kommentare zu dem System? Bitte nehmen Sie sich die Zeit uns diese hier zu erörtern:

Figure C.4 – The German version (page two) of the questionnaire about the interaction which had to be filled out by the participants after each task was fulfilled.

Questionnaire

	strongly disagree				strongly agree
1. The system made it easy to find the offer that I was looking for.	O	O	O	O	O
2. I feel confident that I selected a good offer for my purposes.	O	O	O	O	O
3. I always knew what the system was asking me for.	O	O	O	O	O
4. The system did not give me enough information.	O	O	O	O	O
5. I found the task boring.	O	O	O	O	O
6. The system gave me too much information in one go.	O	O	O	O	O
7. I had serious problems understanding the German texts.	O	O	O	O	O
8. It took me a long time to complete the task.	O	O	O	O	O
9. The system did always understand what I said.	O	O	O	O	O
10. I would rate the German utterances as excellent.	O	O	O	O	O
11. I knew what I could say at each point of the dialogue.	O	O	O	O	O
12. The system responses agreed with my expectation.	O	O	O	O	O
13. There were awkward words and phrases in the German dialogue.	O	O	O	O	O
14. I liked using the system.	O	O	O	O	O
15. I could quickly find what I was looking for.	O	O	O	O	O
16. The German utterances were fluent.	O	O	O	O	O
17. The system gave me a lot of unnecessary information.	O	O	O	O	O
18. The system's responses were appropriate.	O	O	O	O	O
19. The system was very friendly.	O	O	O	O	O
20. I would rather use the English original.	O	O	O	O	O
21. I would consider using a similar system to find a good offer on broadband Internet.	O	O	O	O	O
22. Interacting with the system was really annoying.	O	O	O	O	O
23. I would rate the German utterances as incomprehensible.	O	O	O	O	O

Figure C.5 – The English translation of the questionnaire about the interaction which had to be filled out by the participants after each task was fulfilled.

Appendix D

Materials of the TTS experiments

	quality		accentuation		distinctness		naturalness	
	GS	MT	GS	MT	GS	MT	GS	MT
1	2,625	1,250	2,625	2,750	3,875	2,875	2,250	1,125
2	4,000	0,875	3,125	1,625	4,500	3,375	3,000	1,375
3	3,250	1,750	3,375	2,875	4,500	2,875	3,250	2,375
5	3,500	3,000	3,500	2,875	3,750	4,500	3,375	3,500
6	3,500	2,125	3,500	3,250	4,375	4,125	3,250	2,250
7	2,750	2,125	3,250	2,125	4,625	2,875	2,750	2,500
8	3,375	0,750	3,125	2,250	4,000	1,625	3,000	1,000
9	2,750	1,250	3,000	2,375	3,000	2,125	3,250	1,875
10	1,875	1,625	1,875	2,625	3,125	2,625	2,375	1,250
11	3,625	1,125	3,500	3,000	4,125	2,250	3,625	2,500
12	3,500	1,875	3,000	2,500	5,000	1,250	3,125	2,125
13	3,750	1,000	3,625	1,750	3,250	2,500	3,375	2,625
14	3,500	1,875	4,000	2,000	4,750	1,500	2,875	1,875
15	2,250	1,125	3,375	2,000	3,250	3,375	2,000	2,125
16	3,125	2,750	3,125	3,375	4,750	4,375	3,875	3,875
17	1,250	0,500	2,125	1,250	1,750	1,000	1,375	1,875
19	1,125	1,500	1,750	1,750	1,500	1,625	2,375	0,750
20	1,375	0,875	2,125	1,375	1,000	1,125	1,125	1,750
22	1,250	1,250	2,125	1,500	1,500	0,875	2,500	1,125
23	3,500	3,750	3,750	3,500	4,375	5,375	3,750	3,750
24	3,250	2,500	3,250	2,625	4,625	3,750	3,500	2,125
25	3,000	2,750	3,375	3,375	4,625	4,875	4,250	4,375
26	3,000	3,250	3,500	3,500	5,250	4,875	4,750	4,625
27	3,625	1,750	3,250	1,875	3,375	2,500	2,500	2,750
28	3,750	1,125	3,625	2,875	4,750	2,125	3,250	1,750
29	2,375	0,500	2,000	1,250	2,750	1,625	2,000	1,750
max	4,000	3,750	4,000	3,500	5,250	5,375	4,750	4,625
min	1,125	0,500	1,750	1,250	1,000	0,875	1,125	0,750

Table D.1 – Overview of the ratings for the overall quality, accentuation, distinctness, and naturalness.

#	English source	German gold standard	*Systran* translation
1	Hello. I am a dialogue system with the aim to help you finding a good offer for a broadband Internet connection. What exactly are you looking for?	Hallo. Ich bin ein Dialogsystem das darauf abzielt Ihnen zu helfen ein gutes Angebotes für einen Breitband-Internetanschluss zu finden. Wonach genau suchen Sie?	Hallo. Ich bin ein Dialogsystem mit dem Ziel, zum Sie zu helfen ein gutes Angebot für einen Breitbandinternetanschluss finden. Nach was genau suchen Sie?
2	Sorry, I did not understand you. Could you say that again?	Bitte entschuldigen Sie, ich habe Sie nicht verstanden. Könnten Sie das bitte noch einmal sagen?	Traurig, verstand ich Sie nicht. Konnten Sie das wieder sagen?
3	You want information on how to connect to the Internet in Ireland.	Sie wollen Informationen darüber, wie man an das Internet in Irland angeschlossen wird.	Sie wünschen Informationen über, wie man an das Internet in Irland anschließt.
5	You are looking for a mobile Internet connection.	Sie suchen nach einer mobilen Internetverbindung.	Sie suchen nach einem beweglichen Internetanschluss.
6	You are looking for a mobile Internet connection without a twelve month contract.	Sie suchen nach einer mobilen Internetverbindung ohne einen Zwölf- Monats-Vertrag.	Sie suchen nach einem beweglichen Internetanschlusse ohne einen zwölf-Monats-Vertrag.
7	Do you have any other requirements? Do you, for example, prefer a lower price to the download speed? Or is price not an issue?	Haben Sie irgendwelche anderen Anforderungen? Bevorzugen Sie beispielsweise einen geringeren Preis gegenüber der Download-Geschwindigkeit? Oder ist der Preis kein Thema für Sie?	Haben Sie irgendwelche anderen Anforderungen? Bevorzugen Sie zum Beispiel einen niedrigeren Preis gegenüber der Downloadgeschwindigkeit? Oder ist Preis nicht eine Frage?
8	Sorry, but I have no information on that topic.	Bitte entschuldigen Sie, aber ich habe keine Informationen zu diesem Thema.	Traurig, aber ich haben Sie keine Informationen über dieses Thema.
9	Is there anything else I can do for you?	Gibt es noch etwas, dass ich für Sie tun kann?	Gibt es noch etwas, die ich für Sie tun kann?
10	I hope the information I have given you was helpful. Have a good day.	Ich hoffe, dass die Information, die ich Ihnen gegeben habe war hilfreich. Haben Sie einen guten Tag.	Ich hoffe, dass die Informationen, die ich Sie gegeben habe, hilfreich waren. Haben Sie einen guten Tag.

Table D.2 – The first part of the system utterances that were the basis for the evaluation in the TTS experiment series. First the English source, than the German gold standard translation, followed by the *Systran* machine translation.

#	English source	German gold standard	*Systran* translation
11	There are eighteen different offers with varying download speeds, download allowances and prices.	Es gibt achtzehn verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten, Download-Grenzen und Preisen.	Es gibt achtzehn verschiedene Angebote mit unterschiedlichen Download-geschwindigkeiten, Downloadzulagen und Preisen.
12	There are twelve different offers with varying download speeds, download allowances and prices.	Es gibt zwölf verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten, Download-Grenzen und Preisen.	Es gibt zwölf verschiedene Angebote mit unterschiedlichen Downloadgeschwindigkeiten, Downloadzulagen und Preisen.
13	There are five different offers with varying download speeds, download allowances and prices.	Es gibt fünf verschiedene Angebote mit unterschiedlichen Download-Geschwindigkeiten, Download-Grenzen und Preisen.	Es gibt fünf verschiedene Angebote mit unterschiedlichen Downloadgeschwindigkeiten, Downloadzulagen und Preisen.
14	There are ten different offers with varying download allowances and prices.	Es gibt fünfzehn verschiedene Angebote mit unterschiedlichen Download-Grenzen und Preisen.	Es gibt fünfzehn verschiedene Angebote mit unterschiedlichen Downloadzulagen und -preisen.
16	And finally:	Und schließlich:	Und schließlich:
17	A Pay as you go thirty days from meteor with up to seven point two Megabyte download speed would be a suitable option for you. There is a five Gigabyte download limit. This offer will cost nineteen Euro ninety-nine for the month.	Dreißig Tage Prepaid Internet von meteor mit bis zu sieben Komma zwei Megabyte Download-Geschwindigkeit wäre eine passende Option für Sie. Es gibt dabei eine fünf Gigabyte Download-Grenze. Dieses Angebot kostet neunzehn Euro neunundneunzig im Monat.	Ein Lohn, da Sie Wahl vom meteor mit bis zu sieben Komma zwei Megabyte-Downloadgeschwindigkeit würde eine passende Wahl für Sie sein. Es gibt eine fünf -Gigabyte-Downloadgrenze. Dieses Angebot kostet neunzehn Euro neunundneunzig für das Monat.

Table D.3 – The second part of the system utterances that were the basis for the evaluation in the TTS experiment series. First the English source, than the German gold standard translation, followed by the *Systran* machine translation.

#	English source	German gold standard	*Systran* translation
19	A twelve month contract from BT with up to nine point nine Megabyte download speed would be a suitable option for you. There is a ten Gigabyte download limit. This offer will cost twenty-two Euro fifty-nine for the month.	Ein Zwölf-Monats-Vertrag von BT mit bis zu neun Komma neun Megabyte Download-Geschwindigkeit wäre eine passende Option für Sie. Es gibt dabei eine zehn Gigabyte Download-Grenze. Dieses Angebot kostet zweiundzwanzig Euro neunundfünfzig im Monat.	Ein zwölf-Monats-Vertrag vom BT mit bis zu neun Komma neun Megabyte-Downloadgeschwindigkeit würde eine passende Wahl für Sie sein. Es gibt eine zehn -Gigabyte-Downloadgrenze. Dieses Angebot kostet zweiundzwanzig Euro neunundfünfzig für das Monat.
20	A Pay as you go thirty days from Three-G with up to seven point two Megabyte download speed and up to two-thousand-forty-eight Kilobyte upload speed would be a suitable option for you. There is a ten Gigabyte download limit. This offer will cost twenty-five Euro for the month.	Dreißig Tage Prepaid Internet von Drei-G mit bis zu sieben Komma zwei Megabyte Download-Geschwindigkeit und bis zu zweitausend-achtundvierzig Kilobyte Upload-Geschwindigkeit wäre eine passende Option für Sie. Es gibt dabei eine zehn Gigabyte Download-Grenze. Dieses Angebot kostet fünfundzwanzig Euro im Monat .	Ein Lohn, da Sie Wahl vom Drei-G mit bis zu sieben Komma zwei Megabyte-Downloadgeschwindigkeit und bis zu zweitausend-achtundvierzig Kilobyteantrieb-skraftgeschwindigkeit würde eine passende Wahl für Sie sein. Es gibt eine zehn -Gigabyte-Downloadgrenze. Dieses Angebot kostet fünfundzwanzig Euro für das Monat.
22	A twelve month contract from BT with up to nine point nine Megabyte download speed would be a suitable option for you. There is a thirty Gigabyte download limit. This offer will cost thirty-two Euro for the month.	Ein Zwölf-Monats-Vertrag von BT mit bis zu neun Komma neun Megabyte Download-Geschwindigkeit wäre eine passende Option für Sie. Es gibt dabei eine dreißig Gigabyte Download-Grenze. Dieses Angebot kostet zweiunddreißig Euro im Monat.	Ein zwölf-Monats-Vertrag vom BT mit bis zu neun Komma neun Megabyte-Downloadgeschwindigkeit würde eine passende Wahl für Sie sein. Es gibt eine dreißig -Gigabyte-Downloadgrenze. Dieses Angebot kostet zweiunddreißig Euro für das Monat.

Table D.4 – The third part of the system utterances that were the basis for the evaluation in the TTS experiment series. First the English source, than the German gold standard translation, followed by the *Systran* machine translation.

#	English source	German gold standard	*Systran* translation
23	Do you have any other requirements?	Haben Sie irgendwelche anderen Anforderungen?	Haben Sie irgendwelche anderen Anforderungen?
24	These are the three cheapest results. They will be presented separately. Say "next" if you wish to see the following offer.	Dies sind die drei billigsten Ergebnisse. Sie werden einzeln aufgeführt. Sagen Sie "next" wenn sie das nächste Angebot sehen möchten.	Diese sind die drei billigsten Ergebnisse. Sie werden separat dargestellt. Sagen Sie „next“ wenn Sie das folgende Angebot sehen möchten.
25	The first offer is:	Das erste Angebot ist:	Das erste Angebot ist:
26	The second offer is:	Das zweite Angebot ist:	Das zweite Angebot ist:
27	Are you looking for mobile Internet or for a landline connection?	Suchen Sie nach einer mobilen Internetverbindung oder nach einer Festnetzverbindung?	Suchen Sie nach beweglichem Internet oder für eine Überlandleitungverbindung?
28	Sorry, but I need some time to process this information.	Entschuldigen Sie, aber ich brauche ein wenig Zeit diese Informationen zu verarbeiten.	Traurig, aber ich benötigen Sie einige Zeit, diese Informationen zu verarbeiten.

Table D.5 – The fourth part of the system utterances that were the basis for the evaluation in the TTS experiment series. First the English source, than the German gold standard translation, followed by the *Systran* machine translation.

#	English source	German gold standard	*Systran* translation
29	Are you looking for mobile Internet or for a landline connection? Let me explain the differences to you. If you decide for mobile Internet you will be able to access the Internet from almost anywhere in Ireland. In order to do so you will need to plug a modem, usually a USB stick, into your computer. Experience shows considering mobile Internet is wise when you are travelling a lot. A landline connection on the other hand restricts your Internet access to your home. However this connection type is usually more stable and offers a bigger bandwidth. To set up landline broadband Internet a phoneline at your home is required.	Suchen Sie nach einer mobilen Internetverbindung oder nach einer Festnetzverbindung? Lassen Sie mich Ihnen die Unterschiede erklären. Wenn Sie sich für mobiles Internet entscheiden werden Sie in der Lage sein von fast überall in Irland auf das Internet zuzugreifen. Um dies zu tun müssen Sie ein Modem, normalerweise ein USB-Stick, in Ihren Computer einstecken. Die Preise für den USB Stick sind in unserer Berechnung bereits berücksichtigt. Die Erfahrung zeigt, dass Sie mobiles Internet in betracht ziehen sollten, wenn Sie viel reisen. Ein Festnetzanschluss beschränkt Ihren Internet-Zugang auf Ihr Zuhause. Jedoch ist diese Verbindungsart gewöhnlich stabiler und bietet eine größere Bandbreite. Um Festnetz Breitband-Internet einzurichten benötigen Sie eine Telefonleitung bei Ihnen zu Hause. Der Preis für eine Telefonleitung ist in unserer Berechnung bereits berücksichtigt.	Suchen Sie nach beweglichem Internet oder für eine Überlandleitungverbindung? Lassen Sie mich die Unterschiede Ihnen erklären. Wenn Sie für bewegliches Internet entscheiden, sind Sie in der Lage, auf das Internet von zuzugreifen fast überall in Irland. Zwecks Sie so zu tun muss ein Modem, normalerweise einen USB-Stock, in Ihren Computer verstopfen. Die Preise für den USB-Stock sind in unseren Berechnungen eingeschlossen. Die Erfahrungsshows, die bewegliches Internet betrachten, ist klug, wenn Sie viel reisen. Eine Überlandleitungsverbindung schränkt andererseits Ihren Internet-Zugang zu Ihrem Haus ein. Jedoch ist diese Verbindungsart normalerweise stabiler und bietet eine größere Bandbreite an. Um Überlandleitung Breitbandinternet wird eine Telefonlinie an Ihrem Haus zu gründen angefordert. Der Preis für die Telefonlinie ist bereits in unserer Berechnung eingeschlossen.

Table D.6 – The fifth part of the system utterances that were the basis for the evaluation in the TTS experiment series. First the English source, than the German gold standard translation, followed by the *Systran* machine translation.

	criteria avg.	
	GS	MT
1	2,844	2,000
2	3,656	1,813
3	3,594	2,469
5	3,531	3,469
6	3,656	2,938
7	3,344	2,406
8	3,375	1,406
9	3,000	1,906
10	2,313	2,031
11	3,719	2,219
12	3,656	1,938
13	3,500	1,969
14	3,781	1,813
15	2,719	2,156
16	3,719	3,594
17	1,625	1,156
19	1,688	1,406
20	1,406	1,281
22	1,844	1,188
23	3,844	4,094
24	3,656	2,750
25	3,813	3,844
26	4,125	4,063
27	3,188	2,219
28	3,844	1,969
29	2,281	1,281

Table D.7 – Overview of average over all ratings for accentuation, distinctness, and naturalness.

Alter: _____

Geschlecht: [ ] männlich [ ] weiblich

Bildungsgrad:
[ ] Student im ___ Jahr
[ ] PhD /Doktorand im ___ Jahr
[ ] sonstiges: _____

Art des Studiums:
[ ] Kunst, Geistes-, Sozialwissenschaften
[ ] Naturwissenschaften (Mathematik, Informatik, Biologie, Chemie...)
[ ] Gesundheitswissenschaften (Medizin, Zahnmedizin, ...)
[ ] sonstige: _____

Muttersprache:
[ ] Deutsch
[ ] Englisch
[ ] sonstige: _____

Bitte beurteilen Sie ihre Computerkenntnisse:
[ ] schlecht
[ ] unterdurchschnittlich
[ ] durchschnittlich
[ ] gut
[ ] exzellent

Bitte beurteilen Sie ihre Erfahrungen mit Dialogsystemen:
[ ] kaum
[ ] wenig
[ ] durchschnittlich
[ ] viel
[ ] sehr viel

Bitte beurteilen Sie Ihre Englischkenntnisse:
[ ] schlecht
[ ] unterdurchschnittlich
[ ] durchschnittlich
[ ] gut
[ ] exzellent

Ich habe mich bereits mit der Problematik in Irland einen Internetanschluss zu bekommen auseinander gesetzt.
[ ] Ja
[ ] Nein

Figure D.1 – The demographic questionnaire that was presented to participants at the beginning of the extrinsic TTS evaluation.

Age: _____

Gender:
[ ] male
[ ] female

Level of education:
[ ] student in the ___ year
[ ] PhD in the ___ year
[ ] other: _____

Type of education:
[ ] arts, humanities, social sciences
[ ] engineering, mathematics, computer science, natural sciences
[ ] health sciences
[ ] other: _____

native language:
[ ] German
[ ] English
[ ] other: _____

Please judge your familiarity with Computers:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

Please judge your familiarity with Spoken Dialogue systems:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

Please judge your English capabilities:
[ ] poor
[ ] below average
[ ] average
[ ] good
[ ] excellent

I was already concerned with the topic of getting an Internet connection in Ireland before I took part in this study.
[ ] yes
[ ] no

Figure D.2 – The demographic questionnaire that was presented to participants at the beginning of the extrinsic MT evaluation (english translation).

Participant number:

Bitte machen Sie die folgenden Angaben zu dem von Ihnen gefundenem Angebot.

Antwort Aufgabe A

Anbieter: _____

Download Geschwindigkeit: _____

Maximaler Download im Monat: _____

Preis pro Monat: _____

Antwort Aufgabe B

Anbieter: _____

Download Geschwindigkeit: _____

Maximaler Download im Monat: _____

Preis pro Monat: _____

Figure D.3 – The answersheet that was given out to the participants on which they had to note the Internet provider, maximum download speed, maximum download allowance and price per month.

Questionnaire

	strongly disagree				strongly agree
1. The system made it easy to find the offer that I was looking for.	0	0	0	0	0
2. I feel confident that I selected a good offer for my purposes.	0	0	0	0	0
3. I always knew what the system was asking me for.	0	0	0	0	0
4. The system did not give me enough information.	0	0	0	0	0
5. I found the task boring.	0	0	0	0	0
6. The system gave me too much information in one go.	0	0	0	0	0
7. I had serious problems understanding the German texts.	0	0	0	0	0
8. It took me a long time to complete the task.	0	0	0	0	0
9. The system did always understand what I said.	0	0	0	0	0
10. I would rate the German utterances as excellent.	0	0	0	0	0
11. I knew what I could say at each point of the dialogue.	0	0	0	0	0
12. The system responses agreed with my expectation.	0	0	0	0	0
13. There were awkward words and phrases in the German dialogue.	0	0	0	0	0
14. I liked using the system.	0	0	0	0	0
15. I could quickly find what I was looking for.	0	0	0	0	0
16. The German utterances were fluent.	0	0	0	0	0
17. The system gave me a lot of unnecessary information.	0	0	0	0	0
18. The system's responses were appropriate.	0	0	0	0	0
19. The system was very friendly.	0	0	0	0	0
20. I would rather use the English original.	0	0	0	0	0
21. I would consider using a similar system to find a good offer on broadband Internet.	0	0	0	0	0
22. Interacting with the system was really annoying.	0	0	0	0	0
23. I would rate the German utterances as incomprehensible.	0	0	0	0	0

Figure D.4 – The English translation of the questionnaire about the interaction which had to be filled out by the participants after each task was fulfilled.

	speed	
	GS	MT
1	3,333	4,000
2	3,667	2,667
3	3,000	3,000
5	3,667	3,667
6	3,667	3,000
7	3,667	3,000
8	3,333	3,000
9	3,333	3,667
10	4,000	3,667
11	3,667	3,333
12	4,333	4,333
13	3,333	4,000
14	3,000	3,333
15	3,333	3,667
16	2,333	3,000
17	3,333	3,333
19	4,333	3,667
20	3,667	3,667
22	3,333	4,000
23	3,667	3,000
24	2,667	3,000
25	3,333	2,667
26	2,333	3,000
27	3,333	3,667
28	3,667	3,000
29	4,000	3,000

Table D.8 – Overview of the ratings for speed.

Bibliography

(1994). Telephone transmission quality- a method for subjective performance assessment of the quality of speech voice output devices. ITU-T Recommendation P.85.

Adell, J., Escudero, D., and Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54(3):459–476.

Agirre, E., Màrquez, L., and Wicentowski, R., editors (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Ailomaa, M., Melichar, M., Rajman, M., Lisowska, A., and Armstrong, S. (2006). Archivus: A multimodal system for multimedia meeting browsing and retrieval. In *Proceedings of the Proceedings of the Twenty-First International Conference on Computational Linguistics and Forty-Fourth Annual Meeting of the Association for Computational Linguistics (COLING/ACL) – Interactive Presentation Sessions*, pages 49–52.

Albrecht, J an Hwa, R. (2007). Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the Fortyfifth Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 296–303,.

Alvarez, Y. and Huckvale, M. (2002). The reliability of the ITU-T P. 85 standard for the evaluation of text-to-speech systems. In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP 2002 - INTERSPEECH 2002)*, pages 329–332, Denver, Colorado, USA.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4):351–366.

Anderson, A., Brown, G., Shillcock, R., and Yule, G. (1984). *Teaching Talk: Strategies for Production and Assessment*. Cambridge University Press.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Baeza-Yates, R. and Ribeiro-Neto, B. (2010). *Modern Information Retrieval*. Addison Wesley, second edition.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65 – 72.

Bard, E., Sotillo, C., Anderson, A., Thompson, H., and Taylor, M. (1996). The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, 20(1–2):71–84.

Baum, F. (2008). *The Wizard of Oz*. Puffin Classics.

Bederson, B. B., Hu, C., and Resnik, P. (2010). Translation by iterative collaboration between monolingual users. In *Proceedings of Graphical Interfaces (GI) 2010*, pages 39–46.

Belz, A. (2009). That's nice...What can you do with it? *Computational Linguistics*, 35(1):111–118.

Benoît, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381– 392.

Benzmüller, C., Fiedler, A., Gabsdil, M., Horacek, H., Kruijff-Korbayov, I., Pinkal, M., Siekman, J., Tsovaltzi, D., Vo, B., and Wolska, M. (2003). A Wizard-of-Oz experiment for tutorial dialogues in mathematics. In Aleven, V., Hoppe, U., Kay, J., Mizoguchi, R., Pain, H., Verdejo, F., and Yacef, K., editors, *Supplementary Proceedings of the Eleventh International Conference on Artificial Intelligence in Education (Vol. VIII: Advanced Technologies for Mathematics Education - AIED2003)*, pages 471–481.

Berger, A., Brown, P., Della Pietra, S., Della Pietra, V., Gillett, J., Lafferty, J., Mercer, R., Printz, H., and Ureš, L. (1994). The Candide system for machine translation. In *Proceedings of the Workshop on Human Language Technology (HLT 1994)*, pages 157–162.

Beyer, H. and Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. Morgan Kaufmann.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Bohus, D., Raux, A., Harris, T., Eskenazi, M., and Rudnicky, A. (2007). Olympus: An open-source framework for conversational spoken language interface research. In *Proceedings of HLT-NAACL 2007 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*, pages 32–39.

Bortfeld, H. and Brennan, S. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147.

Boyle, E., Anderson, A., and Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1):1–20.

Bradley, J., Benyon, D., Mival, O., and Webb, N. (2010). Wizard of Oz experiments and companion dialogues. In *Proceedings of the Twenty-Fourth British HCI Group Annual Conference on People and Computers: Celebrating People and Technology (HCI 2010)*, pages 117–123.

Bradley, J., Mival, O., and Benyon, D. (2009). Wizard of Oz experiments for companions. In *Proceedings of the Twenty-Third British HCI Group Annual Conference on People and Computers: Celebrating People and Technology (HCI 2009)*, pages 313–317, Camebridge, UK.

Branigan, H., Pickering, M., Pearson, J., Mclean, J., and Nass, C. (2003). Syntactic alignment between computers and people : The role of belief about mental states. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, pages 186–191.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of International Symposium on Spoken Dialogue (ISSD 1996)*, pages 41–44.

Brennan, S. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings of the Thirty-Eighth Annual Meeting on Association for Computational Linguistics (ACL 2000)*, pages 1–11.

Brennan, S. and Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 22(6):1482–1493.

Bub, T. and Schwinn, J. (1999). The Verbmobil prototype system – a software engineering perspective. *Natural Language Engineering*, 5(1):95–112.

Buxton, B. (2007). *Sketching user experience: Getting the design right and the right design*. Morgan Kaufmann.

Cabral, J., Kane, M., Ahmed, Z., Abou-Zleikha, M., Székely, E., Zahra, A., Ogbureke, K., Cahill, P., Carson-Berndsen, J., and Schlögl, S. (2012a). Rapidly testing the interaction model of a pronunciation training system via Wizard-of-Oz. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Cabral, J., Kane, M., Ahmed, Z., Székely, E., Zahra, A., Ogbureke, K., Cahill, P., Carson-Berndsen, J., and Schlögl, S. (2012b). Using the Wizard-of-Oz framework in a pronunciation training system for providing user feedback and instructions. In *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT 2012)*.

Cahill, P. and Carson-Berndsen, J. (2010). MUSE: An open source speech technology research platform. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT 2010)*, pages 115–120.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 249–256.

Campbell, N. (2010). Expressive speech processing and prosody engineering: An illustrated essay on the fragmented nature of real interactive speech. In Chen, F. and Jokinen, K., editors, *Speech Technology*, pages 105–121. Springer US.

Carletta, J. (1992). Planning to fail, not failing to plan: risk-taking and recovery in task-oriented dialogue. In *Proceedings of the Fourteenth Conference on Computational Linguistics (COLING 1992)*, pages 6–10, Nantes, France.

Carletta, J., Caley, R., and Isard, S. (1993). A collection of self-repairs from the map task corpus. Technical Report HCRC/TR-47, Human Communication Research Centre, Edinburgh, Scotland.

Carroll, J. and Aaronson, A. (1988). Learning by doing with simulated intelligent help. *Communications of the ACM*, 31(9):1064–1079.

Carstensen, K., Ebert, C., Endriss, C., Jekat, S., Klabunde, R., and Langer, H., editors (2001). *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Spektrum Akademischer Verlag.

Charniak, E., Knight, K., and Yamada, K. (2003). Syntax-based language models for statistical machine translation. In *Proceedings of the Nineth Machine Translation Summit (MT Summit IX)*, pages 40–46, New Orleans, USA.

Chatzichrisafis, N., Bouillon, P., Rayner, M., Santaholma, M., Starlander, M., and Hockey, B. (2006). Evaluating task performance for a unidirectional controlled language medical speech translation system. In *Proceedings of the First Workshop on Medical Speech Translation at HLT-NAACL 2006*, number 6, pages 5–12, New York.

Christiansen, M., Collins, C., and Edelman, S. (2009). *Language Universals*. Oxford University Press.

Clayes, E. and Anderson, A. (2007). Real faces and robot faces: The effects of representation on computer-mediated communication. *International Journal of Human-Computer Studies*, 65(6):480–496.

Cooke, M., Hershey, J., and Rennie, S. (2010). Monaural speech separation and recognition challenge. *Computer Speech & Language*, 24(1):1 – 15.

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the Thirty-Nineth Annual Meeting on Association for Computational Linguistics (ACL 2001)*, ACL 2001, pages 148–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coutaz, J., Salber, D., Carraux, E., and Portolan, N. (1996). NEIMO, a multiworkstation usability lab for observing and analyzing multimodal interaction. In *Proceedings of the Conference on Human Factors in Computing Systems: Common Ground (CHI 1996)*, pages 402–403. ACM.

Culy, C. and Riehemann, S. (2003). The limits of N-gram translation evaluation metrics. In *Proceedings of the Nineth Machine Translation Summit (MT Summit IX)*, pages 71–78, New Orleans, USA.

Davis, J. (1998). Active help found beneficial in wizard of oz study. *Information and Software Technology*, 40:93–103.

Davis, K., Biddulph, R., and Balashek, S. (1952). automatic recognition of spoken digits. *Journal of the Acoustical Society of America*, 24(6):637–642.

Dillinger, M. and Seligman, M. (2006). Converser: Highly interactive speech-to-speech translation for healthcare. In *Proceedings of the Workshop on Medical Speech Translation at HLT-NAACL 2006*, pages 40–43.

Dix, A., Finlay, J., and Abowd, G. (2004). *Human-computer interaction*. Pearson Education Limited.

Dixon, N. and Maxey, H. (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electroacoustics*, 16(1):40 – 50.

Doddington, G. (2002). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT 2002)*, pages 138–145, San Diego, California.

Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J., and Gandy, M. (2005). Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP 1996)*, volume 3, pages 1393–1396. Ieee.

Dybkjaer, H., Bernsen, N., and Dybkjaer, L. (1993). Wizard-of-Oz and the trade off between naturalness and recognizer constraints. In *Proceedings of EUROSPEECH 1993*.

Dybkjaer, L. and Bernsen, N. (2000). Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6(3& 4):243–271.

Dybkjaer, L. and Bernsen, N. (2001). Usability issues in spoken dialogue systems. *Natural Language Engineering*, 6(3&4):243–271.

Dybkjaer, L., Bernsen, N., and Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43(1-2):33–54.

Erdmann, R. and Neal, A. (1971). Laboratory vs. field experimentation in human factors - An evaluation of an experimental self-service airline ticket vendor. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 13(6):521–531.

Fabbrizio, G., Tur, G., and Hakkani-Tür, D. (2005). Automated Wizard-of-Oz for spoken dialogue systems. In *Proceedings of INTERSPEECH 2005 - EUROSPEECH*.

Fackrell, J. (2004). Improving pronunciation dictionary coverage of names by modelling spelling variation. In *Proceedings of the Fifth ISCA Workshop on Speech Synthesis*, pages 121–126.

Faulkner, X. (2000). *Usability engineering*. Palgrave Macmillan.

Flanagan, M. (1994). Error classification for MT evaluation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 1994)*.

Forbes-Riley, K., Litman, D., Silliman, S., and Tetreault, J. (2006). Comparing synthesized versus pre-recorded tutor speech in an intelligent tutoring spoken dialogue system. In *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2006)*.

Furui, S. (2005). 50 years of progress in speech and speaker recognition research. *ECTI Transaction on Computer and Information Technology*, 1(2):64–74.

Garofolo, J., Voorhees, E. M., Auzanne, C., Stanford, V. M., and Lund, B. (1998). 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of the Seventh Text REtrieval Conference (TREC 7)*, pages 79–90.

Geutner, P., Steffens, F., and Manstetten, D. (2002). Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

Giménez, J. and Màrquez, L. (2008). A smorgasbord of features for automatic MT evaluation. In *In Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, OH.

Gödde, F., Möller, S., Engelbrecht, K., Kühnel, C., Schleicher, R., Naumann, A., and Wolters, M. (2008). Study of a speech-based smart home system with older users. In *Proceedings of Workshop on Intelligent UIs for Ambient Assisted Living*.

Goldstein, J., Lavie, A., Lin, C., and Voss, C., editors (2005). *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. ACL.

Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication*, 16(3):225 – 244.

Goldstein, M., Bretan, I., Sallnas, E., and Bjork, H. (1999). Navigational abilities in audial voice-controlled dialogue structures. *Behaviour & Information Technology*, 18(2):83–95.

Gong, L. and Lai, J. (2003). To mix or not to mix synthetic speech and human speech? contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. *International Journal of Speech Technology*, 6:123–131.

Good, M., Whiteside, J., Wixon, D., and Jones, S. (1984). Building a user-derived interface. *Communications of the ACM*, 27(10):1032–1043.

Gould, J. and Lewis, C. (1983). Designing for usability - Key principles and what designers think. In *Proceedings of Conference on Human Factors in Computing Systems (CHI 1983)*.

Gould, J. and Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM*, 28(3):300–311.

Guhe, M. (2007). Adaptation of the use of colour terms in referring expressions. In *Proceedings of the Eleventh Workshop on the Semantics and Pragmatics of Dialogue (Decalog 2007)*, pages 167–168, Trento, Italy.

Guhe, M. and Gurman Bard, E. (2008a). Adapting referring expressions to the task environment comparison to existing research. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 2404–2409.

Guhe, M. and Gurman Bard, E. (2008b). Adapting the use of attributes to the task environment in joint action: Results and a model. In *Proceedings of the Twelfth Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL 2008)*, pages 91–98, London, UK.

Gustafson, J., Larsson, A., Carlson, R., and Hellman, K. (1997). How do system questions influence lexical choices in user answers. In *Proceedings of EUROSPEECH 1997*.

Halverson, C., Horn, D., Karat, C., and Karat, J. (1999). The beauty of errors: Patterns of error correction in desktop speech systems. In *Proceedings of the Thirteenth International Conference on Human-Computer Interaction (INTERACT 1999)*. IOS Press.

Harman, D. (1993). Overview of TREC-1. In *Proceedings of the Workshop on Human Language Technology (HLT 1993)*, HLT 1993, pages 61–65, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harris, C. (1953). A study of the building blocks in speech. *Journal of the Acoustical Society of America*, 25(5):962–969.

Hauptmann, A. (1989). Speech and gestures for graphic image manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1989)*, pages 241–245.

He, Y. and Way, A. (2009). Learning labelled dependencies in machine translation evaluation. In *Proceedings of the Thirteenth Annual Meeting of the European Association for Machine Translation (EAMT 2009)*.

Hemdal, J. and Hughes, G. (1967). *Models for the Perception of Speech and Visual Form*, chapter A feature based computer recognition program for the modeling of vowel perception. MIT Press.

Hill, W. and Miller, J. (1988). Justified advice: A semi-naturalistic study of advisory strategies. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 185–190. ACM Press.

Hirschman, L. and Thompson, H. (1997). *Survey of the State of the Art in Human Language Technology*, chapter Overview of Evaluation in Speech and Natural Language Processing. Studies in Natural Language Processing. Cambridge University Press.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1.

Hone, K. and Graham, R. (2000). Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering*, 6(3-4):287–303.

House, A., Williams, C., Hecker, M., and Kryter, K. (1965). Articulatory testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37:158–166.

Hovy, E. (1999). Toward finely differentiated evaluation metrics for machine translation. In *Proceedings of the EAGLES Workshop on Standards and Evaluation*, pages 127–133.

Hovy, E., King, M., and Popescu-Belis, A. (2002). Principles of context-based machine translation evaluation. *Machine Translation*, 17:43–75.

Howell, M., Love, S., and Turner, M. (2005). The impact of interface metaphor and context of use on the usability of a speech-based mobile city guide service. *Behaviour & Information Technology*, 24(1):67–78.

Hunt, A. J. and Black, W. (1996). Unit selection in a concatenative speech synthesis system. In *Proceedings of the IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP 1996)*, pages 373–376, Atlanta, Georgia, USA.

Hutchins, W. (2004). The Georgetown-IBM experiment demonstrated in January 1954. In *Proceedings of the Sixth conference of the Association for Machine Translation in the Americas (AMTA 2004)*, number January, pages 102–114, Washington, DC, USA.

Janarthanam, S. and Lemon, O. (2009). A Wizard-of-Oz environment to study referring expression generation in a situated spoken dialogue task. In *Proceedings of the Twelfth European Workshop on Natural Language Generation (ENLG 2009)*, pages 94–97. Association for Computational Linguistics.

Jelinek, F. (2009). The dawn of statistical ASR and MT. *Computational Linguistics*, 35(4):483–494.

Jiang, J., Ahmed, Z., Carson-Berndsen, J., Cahill, P., and Way, A. (2011). Phonetic representation-based speech translation. In *Proceedings of the Thirteenth Machine Translation Summit (MT Summit XIII)*, Xiamen, China.

Jones, D., Gibson, E., and Shen, W. (2005a). Measuring human readability of machine-generated text: Three case studies in speech recognition and machine translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages 1009–1012.

Jones, D., Shen, W., and Weinstein, C. (2005b). New measures of effectiveness for human language technology. *Lincoln Laboratory Journal*, 15(2):341–345.

Jurafsky, D. and Martin, J. (2008). *Speech and Language Processing*. Prentice Hall International, second edition.

Jurafsky, D. and Martin, J. (2009). *Speech and language processing*. Pearson Education, Prentice Hall International, second edition.

Kalikow, D., Stevens, K., and Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61(5):1337–1351.

Karamanis, N., Luz, S., and Doherty, G. (2010). Translation practice in the workplace and machine translation. In *Proceedings of the Fourteenth Annual Meeting of the European Association for Machine Translation (EAMT 2010)*, Saint-Raphaël, France.

Karamanis, N., Schneider, A., van der Sluis, I., Schlögl, S., Doherty, G., and Luz, S. (2009). Do HCI and NLP interact? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*, Boston, MA, USA. ACM Press.

Karpov, A., Ronzhin, A., and Leontyeva, A. (2008). A semi-automatic Wizard of Oz technique for Let'sFly spoken dialogue system. In Sojka, P., Horák, A., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 5246 of *Lecture Notes in Computer Science*, pages 585–592. Springer Berlin / Heidelberg.

Kelley, J. (1983). An empirical methodology fro writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1983)*, pages 193–196.

von Kempelen, W. (1970). *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*. Number Bd. 4 in Grammatica universalis. F. Frommann.

Kit, C. and Wong, T. (2008). Comparative evaluation of online machine translation systems with legal texts. *Law Library Journal*, 100(2):299–321.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995.

Klemmer, S., Sinha, A., Chen, J., Landay, J., Aboobaker, N., and Wang, A. (2000). SUEDE: A Wizard of Oz prototyping tool for speech user interfaces. In *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software and Technology*, pages 1–10.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 187–193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of th Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, pages 48–54, Edmonton.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41(3-4):295– 321.

Krause, D. (1997). Using an interpretation system - Some observations in hidden operator simulations of 'VERBMOBIL'. In *Proceedings of Dialogue Processing in Spoken Language Systems (Workshop on ECAI 1996)*.

Kumar, E. (2011). *Natural Language Processing*. International Publishing House Pvt. Ltd.

Lamel, L. (1998). Spoken language dialog system development and evaluation at LIMSI. In *Proceedings of the International Symposium on Spoken Dialogue (ISSD 1998)*, Sydney, Australia.

Laoudi, J., Tate, C., and Voss, C. (2006). Task-based MT evaluation: From who/when/where extraction to event understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 2048–2053, Genoa, Italy.

Larsen, L. (2003). Assessment of spoken dialogue system usability- What are we really measuring? In *Proceedings of EUROSPEECH 2003*, pages 1945–1948.

Lee, M. and Billinghurst, M. (2008). A Wizard of Oz study for an AR multimodal interface. In *Proceedings of the Tenth International Conference on Multimodal Interfaces (ICMI 2008)*, pages 249–256.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Levin, L., Bartlog, B., Llitjos, A., Gates, D., Lavie, A., Wallace, D., Watanabe, T., and Woszczyna, M. (2000). Lessons learned from a task-based evaluation of speech-to-speech machine translation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

Lewis, J. (2001). The revised Mean Opinion Scale (MOS-R): Preliminary psychometric evaluation. Technical Report TR 29.3414, IBM Voice Systems.

Liang, F. (1983). *Word hy-phen-a-tion by com-puter*. PhD thesis, Stanford University.

Lin, C. (2004). ROUGE: A package for automatic evaluation of summaries. In *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain.

Lin, C. and Och, F. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lippmann, R. (1987). An introduction to computing with neural nets. *ASSP Magazine, IEEE*, 4(2):4 –22.

Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Liu, D. and Gildea, D. (2005). Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan.

Louwerse, M., Jeuniaux, P., Hoque, M., Wu, J., and Lewis, G. (2006). Multimodal communication in computer-mediated map task scenarios. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society (CogSci 2005)*, number 1998, pages 1717–1722.

Mäkelä, K., Salonen, E., Turunen, M., Hakulinen, J., and Raisamo, R. (2001). Conducting a Wizard of Oz experiment on a ubiquitous computing system doorman. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, number 1, pages 9–11.

Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. (2005). On the use of information retrieval measures for speech recognition evaluation - Research report 04-73. Technical report, IDIAP Research Institute.

McInnes, F.R .and Attwater, D. and Edgington, M. (1999). User attitudes to concatenated natural speech and text-to-speech synthesis in an automated information service. In *Proceedings of EUROSPEECH 1999*, pages 831–834.

Melichar, M. and Cenek, P. (2006). From vocal to multimodal dialogue management. In *Proceedings of the Eighth International Conference on Multimodal Interfaces (ICMI 2006)*, pages 59–67.

Mesgarani, N. and Chang, E. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, advance online publication.

Molla, D. and Hutchinson, B. (2003). Intrinsic versus extrinsic evaluations of parsing systems. In *In Proceedings of EACL Workshop on Evaluation Initiatives in Natural Language Processing*, pages 43–50.

Möller, S. (2005). Parameters for quantifying the interaction with spoken dialogue telephone services. In *Proceedings of Sixth SIGDial Workshop on Discourse and Dialogue (SIGDial 2005)*.

Möller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, J., Schuchardt, D., Fakotakis, N., Ganchev, T., and Potamitis, I. (2004). INSPIRE: Evaluation of a smart-home system for infotainment management and device control. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

Möller, S., Krebber, J., and Smeele, P. (2006). Evaluating the speech output component of a smart-home system. *Speech Communication*, 48(1):1 – 27.

Möller, S., Smeele, P., Boland, H., and Krebber, J. (2007). Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech & Language*, 21(1):26–53.

Möller, S. and Ward, N. (2008). A framework for model-based evaluation of spoken dialog systems. In *Proceedings of Workshop on Discourse and Dialogue (SIGDial 2008)*.

Morris, A. (2002). An information theoretic measure of sequence recognition performance. Technical report, IDIAP Research Institute.

Morris, A. C., Maier, V., and Green, P. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition Institute of Phonetics. In *Proceedings of the Eighth International Conference on Spoken Language Processing (ICSLP 2004)*.

Morton, H., Gunson, N., Marshall, D., McInnes, F., Ayres, A., and Jack, M. (2011). Usability assessment of text-to-speech synthesis for additional detail in an automated telephone banking system. *Computer Speech & Language*, 25(2):341 – 362.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453 – 467.

Myers, C. and Rabiner, L. (1981). A level building dynamic time warping algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):284 – 297.

Nagao, M. (1984). A framework of mechanical translation between Japanese and English by analogy principle. In *Artificial and human intelligence: edited review papers presented at the international NATO Symposium*, pages 173–180, Amsterdam, Holland.

Newlands, A., Anderson, A., and Mullin, J. (2003). Adapting communicative strategies to computer-mediated communication: an analysis of task performance and dialogue structure. *Applied Cognitive Psychology*, 17(3):325–348.

Ozkan, N. and Paris, C. (2002). Cross-fertilization between human computer interaction and natural language processing: Why and how. *International Journal of Speech Technology*, 5(2):135–146.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLUE: a method for automatic evaluation of machine translation. In *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Paul, M., Okuma, H., Yamamoto, H., Sumita, E., Matsuda, S., Shimizu, T., and Nakamura, S. (2008). Multilingual mobile-phone translation services for world travelers. In *Proceedings of the Twenty-Second Conference on Computational Linguistics (COLING 2008)*.

Peissner, M., Heidmann, F., and Ziegler, J. (2001). Simulating recognition errors in speech user interface prototyping. In *Proceedings of HCI International*, pages 233–237.

Pickering, M. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, 27(2):169–90.

Polkosky, M. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, 6:161–182.

Pols, L. (1997). *Survey of the state of the art in human language technology*, chapter Speech synthesis evaluation. Studies in Natural Language Processing. Cambridge University Press.

Popovic, M., Stein, D., and Ney, H. (2006). Statistical machine translation of German compound words. In *Proceedings of the Fifth International Conference on Natural Language Processing (FinTAL 2006)*, pages 616–624.

Postel, H. (1969). Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. *IBM-Nachrichten*, 19:925–931.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Rabiner, L. and Juang, B. (1986). An introduction to Hidden Markov Models. *IEE ASSP Magazine*, 3(1):4 –16.

Rabiner, L. and Juang, B. (1993). *Fundamentals of speech recognition*. Prentice Hall Signal Processing Series. PTR Prentice Hall.

Rajman, M., Ailomaa, M., Lisowska, A., Melichar, M., and Armstrong, S. (2006). Extending the Wizard of Oz methodology for language-enabled multimodal systems. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Rayner, M., Bouillon, P., Brotanek, J., Flores, G., Halimi, S., Hockey, B., Isahara, H., Kanzaki, K., Kron, E., Nakao, Y., Santaholma, M., Starlander, M., and Tsourakis, N. (2008a). The

2008 MedSLT system. In *Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications at COLING 2008*, pages 32–35, Manchester, England,.

Rayner, M., Bouillon, P., Flores, G., Ehsani, F., Starlander, M., Hockey, B., Brotanek, J., and Biewald, L. (2008b). A small-vocabulary shared task for medical speech translation. In *Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications at COLING 2008*, number 8, pages 60–63.

Reddy, D. (1966). Approach to computer speech recognition by direct analysis of the speech wave. *Journal of the Acoustical Society of America*, 40(5):1273–1273.

Reiter, E. (2007). The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287.

Reitter, D., Keller, F., and Moore, J. (2006a). Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL 2006*, pages 121–124.

Reitter, D., Moore, J., and Keller, F. (2006b). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society (CogSci 2006)*, Vancouver, Canada.

Resnik, P. (1997). Evaluating multilingual gisting of Web pages. In *Proceedings of the AAAI Symposium on Natural Language Processing for the World Wide Web*, Stanford.

Rodgers, J. and Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):pp. 59–66.

Saint-Aimé, S., Grandgeorge, M., Le-Pévédic, B., and Duhaut, D. (2011). Evaluation of Emi interaction with non-disabled children in nursery school using wizard of oz technique. In *Proceedings of IEEE Robio*, pages 1147–1153, Phucket, Thailand.

Sakoe, H. (1979). Two-level DP-matching – A dynamic programming-based pattern matching algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(6):588 – 595.

Salber, D. and Coutaz, J. (1993). A Wizard of Oz platform for the study of multimodal systems. In *INTERACT 1993 and CHI 1993 conference companion on Human factors in computing systems (HI 1993)*, pages 95–96, Amsterdam, The Netherlands. ACM Press.

van Santen, J. (1993). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech & Language*, 7(1):49 – 100.

van Santen, J. (1997). *Multilingual text-to-speech synthesis: The Bell Labs approach*, chapter 8 - Evaluation. Springer.

Schlögl, S., Doherty, G., Karamanis, N., Schneider, A., and Luz, S. (2010a). Observing the Wizard: In search of a generic interface for Wizard of Oz studies. In *Proceedings of the Irish Human Computer Interaction Conference (iHCI 2010)*, pages 43–50, Dublin, Ireland.

Schlögl, S., Doherty, G., and Luz, S. (2010b). WebWOZ: A Wizard of Oz prototyping framework. In *Engineering Interactive Computing Systems (EICS) 2010.*

Schlögl, S., Schneider, A., Luz, S., and Doherty, G. (2011). Supporting the Wizard: Interface improvements in Wizard of Oz studies. In *Proceedings of the Twenty-Fifth BCS Conference on Human-Computer Interaction*, Newcastle Upon Tyne, UK.

Schneider, A. and Luz, S. (2011). Speaker alignment in synthesised, machine translated communication. In *Proceedings of the Eighth International Workshop on Spoken Language Translation (IWSLT 2011)*, pages 254–260, San Francisco, USA.

Schneider, A., van der Sluis, I., and Luz, S. (2010). Comparing intrinsic and extrinsic evaluation of MT output in a dialogue system. In *Proceedings of the Seventh International Workshop on Spoken Language Translation (IWSLT 2010)*, pages 329–336, Paris, France.

Schütze, C. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

Schütze, C. (2005). *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, chapter Thinking About What We Are Asking Speakers to Do, pages 457–485. Mouton De Gruyter, New York.

Seligman, M. and Dillinger, M. (2006). Usability issues in an interactive speech-to-speech translation system for healthcare. In *Proceedings of the Workshop on Medical Speech Translation at HLT-NAACL 2006*, pages 1–8, New York, New York, USA. Association for Computational Linguistics.

Seligman, M. and Dillinger, M. (2011). Real-time multi-media translation for healthcare: a usability study. In *Proceedings of the Thirteenth Machine Translation Summit (MT Summit XIII)*, pages 595–602, Xiamen, China.

Serrano, M. and Nigay, L. (2010). A wizard of oz component-based approach for rapidly prototyping and testing input multimodal interfaces. *Journal on Multimodal User Interfaces*, 3(3):215–225.

Sharp, H., Rogers, Y., and Preece, J. (2007). *Interaction Design: Beyond Human Computer Interaction*. Wiley.

Shimizu, T., Ashikari, Y., Sumita, E., Zhang, J., and Nakamura, S. (2008). NICT/ATR Chinese-Japanese-English speech-to-speech translation system. *Tsinghua Science & Technology*, 13(4):540–544.

Shneiderman, B. and Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. Addison Wesley, fourth edition.

Skantze, G. (2005). Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325 – 341.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts.

Spärck Jones, K. and Galliers, J. (1995). *Evaluating natural language orocessing systems - An analysis and review*, volume 1083 of *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*. Springer Berlin Heidelberg.

Spiegel, M., Altom, M., Macchi, M., and Wallace, K. (1988). Using a monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech. In *Proceedings of the American Voice Input/Output Society*.

Stanier, A. (1990). How accurate is SOUNDEX matching? *Computers in Genealogy*.

Starlander, M. and Estrella, P. (2011). Looking for the best evaluation method for interlingua-based spoken language translation in the medical domain. In *Proceedings of the Eighth International NLPCS Workshop Human-Machine Interaction in Translation*.

Strauß, P., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Traue, H., and Weidenbacher, U. (2008). The PIT corpus of German multi-party dialogues. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2442–2445, Marrakech, Morocco.

Stymne, S. (2009). A comparison of merging strategies for translation of German compounds. In *Proceedings of the Student Research Workshop on the Twelfth Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 61–69.

Sutton, S., Cole, R., Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, P., Kain, A., Wouters, J., Massaro, D., and Cohen, M. (1998). Universal speech tools: The CSLU Toolkit. In *In Proceedings of the International Conference on Spoken Language Processing*, pages 3221–3224.

Taylor, K. and White, J. (1998). Predicting what MT is good for: User judgments and task performance. In *Proceedings of Association for Machine Translation in the Americas (AMTA 1998)*, pages 364–373. Springer Berlin.

Taylor, P. (2009). *Text-to-Speech synthesis*. Cambridge University Press.

Thouin, B. (1982). The METEO system. In *Proceedins of the Conference on Practical Experience of Machine Translation*, pages 39–44, London, UK.

Turian, J., Shen, L., and Melamed, I. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of the Nineth Machine Translation Summit (MT Summit IX)*, pages 368–393, New Orleans.

Turunen, M. and Hakulinen, J. (2000). Jaspis - A framework for multilingual adaptive speech applications. In *Proceedings of the Sixth International Conference of Spoken Language Processing (ICSLP 2000)*, pages 719–722.

Varges, S. (2005). Spatial descriptions as referring expressions in the MapTask domain. In *Proceedings of the Tenth European Workshop On Natural Language Generation (ENLG 2005)*, Aberdeen, UK.

Viethen, J., Zwarts, S., Dale, R., and Guhe, M. (2010). Dialogue reference in a visual domain. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2457–2462, Valletta, Malta.

Vintsyuk, T. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4:52–57.

Viswanathan, M. and Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*, 19(1):55 – 83.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260 –269.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the Sixteenth Conference on Computational Linguistics (COLING 1996)*, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vogel, S., Schultz, T., Waibel, A., and Yamamoto, S. (2006). Chapter 10 - speech-to-speech translation. In Schultz, T. and Kirchhoff, K., editors, *Multilingual Speech Processing*, pages 317 -- 397. Academic Press.

Voiers, W., Sharpley, A., and Hehmsoth, C. (1972). Research on diagnostic evaluation of speech intelligibility - Research Report AFCRL-72-0694. Technical report, Cambridge Research Laboratories.

Voss, C. and Tate, C. (2006). Task-based evaluation of Machine Translation (MT) engines: Measuring how well people extract who, when, where-type elements in MT output. In *Proceedings of the Eleventh Annual Meeting of the European Association for Machine Translation (EAMT 2006)*, Oslo, Norway.

Waibel, A. and Fügen, C. (2008). Spoken language translation. *IEEE Signal Processing Magazine*, (May):70–79.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(3):328 –339.

Walker, M., Kamm, C., and Litman, D. (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3-4):363–377.

White, J., Doyon, J., and Talbott, S. (2000). Task tolerance of MT output in integrated text processes. In *Proceedings of the ANLP/NAACL Workshop on Embedded Machine Translation Systems (ANLP/NAACL)*, pages 9–16, Seattle, WA.

Whittaker, S., Walker, M., and Moore, J. (2002). Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

Winograd, T. (2006). Shifting viewpoints: Artificial Intelligence and Human–Computer interaction. *Artificial Intelligence*, (18):1256–1258.

Yamashita, N. and Ishida, T. (2006). Effects of machine translation on collaborative work. In *Proceedings of the Proceedings of the Twentieth Anniversary Conference on Computer Supported Cooperative Work (CSCW 2006)*.

Yazgan, A. and Saraclar, M. (2004). Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, volume 1, pages 745–748.

Ye, Y., Zhou, M., and Lin, C. (2007). Sentence level machine translation evaluation as a ranking. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, Prague, Czech Republic. Association for Computational Linguistics.

Yule, G. and Powers, M. (1994). Investigating the communicative outcomes of task-based interaction. *System*, 22(1):81–91.

Zhou, L., Lin, C., Munteanu, D., and Hovy, E. (2006). ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, NY.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34(4):527–547.

Zue, V. and Cole, R. (1997). *Survey of the state of the art in Human Language Technology*, chapter Spoken Language Input: Overview. Studies in Natural Language Processing. Cambridge University Press.