Evaluation of Text and Speech Systems

Text, Speech and Language Technology

VOLUME 37

The titles published in this series are listed on www.springer.com.

Evaluation of Text and Speech Systems

Edited by

Laila Dybkjær
Prolog Development Center A/S
Brøndby
Denmark

Holmer Hemsen
DFKI Language Technology Lab
Berlin
Germany

Wolfgang Minker
Institute for Information Technology
University of Ulm
Germany

Editors

Laila Dybkjær
Prolog Development Center A/S
Brøndby
Denmark
laila@pdc.dk

Dr. Wolfgang Minker
Institute for Information Technology
University of Ulm
Germany
wolfgang.minker@uni-ulm.de

Holmer Hemsen
DFKI Language Technology Lab
Berlin
Germany
holmer.hemsen@dfki.de

ISBN 978-1-4020-5816-5 e-ISBN 978-1-4020-5187-2

Library of Congress Control Number: 2008924781

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Contents

Preface

This book has its point of departure in courses held at the Tenth European Language and Speech Network (ELSNET) Summer School on Language and Speech Communication which took place at NISLab in Odense, Denmark, in July 2002. The topic of the summer school was "Evaluation and Assessment of Text and Speech Systems".

Nine (groups of) lecturers contributed to the summer school with courses on evaluation of a range of important aspects of text and speech systems, including speaker recognition, speech synthesis, talking animated interface agents, part-of-speech tagging and parsing technologies, machine translation, question-answering and information retrieval systems, spoken dialogue systems, language resources, and methods and formats for the representation and annotation of language resources. Eight of these (groups of) lecturers agreed to contribute a chapter to the present book. Since we wanted to keep all the aspects covered by the summer school, an additional author was invited to address the area of speaker recognition and to add speech recognition, which we felt was important to include in the book. Although the point of departure for the book was the ELSNET summer school held in 2002, the decision to make a book was made considerably later. Thus the work on the chapters was only initiated in 2004. First drafts were submitted and reviewed in 2005 and final versions were ready in 2006.

The topic of evaluation has grown from an afterthought into an important part of systems development and a research topic of its own. The choice of evaluation of text and speech systems as the topic for the 2002 summer school was no doubt a timely one, and evaluation has not become less important since then. On the contrary, and probably fuelled by the increasing sophistication of text and speech systems, evaluation has moved to an even more central position. Thus we believe that time is opportune for a book that provides overviews of evaluation in most key areas within text and speech systems. The book targets not only graduate students and Ph.D. students but also academic and industrial researchers and practitioners more generally who are keen on getting an overview of the state of the art and best practice in evaluation in one or more of the aspects dealt with in this book. Since the evaluation area is constantly

developing, it may be difficult, in particular for newcomers to the field, to get an overview of current and best practice. The book may therefore be suitable both as a course book if the purpose is to give graduate students an overview of text and speech systems and their evaluation, and as supplementary reading material for graduate courses on one or more particular areas of text and speech systems.

We would like to thank the many people who contributed one way or another to the ELSNET summer school in 2002, without which this book would not have been written. We are grateful to all those who helped us in preparing the book. In particular, we would like to express our gratitude to the following external reviewers for their valuable comments and criticism on first drafts of the nine chapters: John Aberdeen, Elisabeth André, Walter Daelemans, Christophe d'Alessandro, John Garofolo, Inger Karlsson, Adam Kilgariff, Alon Lavie, Chin-Yew Lin, Susann Luperfoy, Inderjeet Mani, Joseph Mariani, John Mason, Dominic Massaro, Sebastian Möller, Roberto Pieraccini, Alexander Rudnicky, Michael Wagner, and Andrew Wilson. A special thanks is also due to people at the Department of Information Technology in Ulm and at NISLab in Odense for their support in editing the book.

Laila DYBKJÆR

Holmer HEMSEN

Wolfgang MINKER

Contributing Authors

Niels Ole Bernsen is Professor at, and Director of, the Natural Interactive Systems Laboratory, the University of Southern Denmark. His research interests are spoken dialogue systems and natural interactive systems more generally, including embodied conversational agents, systems for learning, teaching, and entertainment, online user modelling, modality theory, systems and component evaluation, as well as usability evaluation, system simulation, corpus creation, coding schemes, and coding tools.

Nick Campbell is currently an Expert Researcher with the National Institute of Information and Communications Technology (NICT) and the ATR Spoken Language Communications Research Labs. He is also Visiting Professor at Kobe University and at the Nara Institute of Science & Technology in Japan. He received his Ph.D. in Experimental Psychology from the University of Sussex in the UK in 1990. His research interests include prosody processing, corpus collection and analysis, concatenative speech synthesis, and expressive speech processing. As co-founder of ISCA's Speech Synthesis Special Interest Group, he has a particular interest in furthering the capabilities and quality of speech synthesis.

Christopher Cieri is the Executive Director of the Linguistic Data Consortium where he has overseen dozens of data collection and annotation projects that have generated speech and text corpora in many languages to support basic research and human language technology research and development. He has also served as principal investigator on several projects in which LDC coordinated linguistic resources for multiple site, common task technology evaluations. His Ph.D. is in Linguistics from the University of Pennsylvania. His research interests revolve around corpus-based language description, especially in phonetics, phonology, and morphology as they interact with non-linguistic phenomena, for example, in studies of language contact and linguistic variation.

Laila Dybkjær is a Professor at NISLab, University of Southern Denmark. She holds a Ph.D. degree in Computer Science from Copenhagen University. Her research interests are topics concerning design, development, and evaluation of user interfaces, including development and evaluation of interactive

speech systems and multimodal systems, design and development of intelligent user interfaces, usability design, dialogue model development, dialogue theory, and corpus analysis.

Sadaoki Furui is a Professor at Tokyo Institute of Technology, Department of Computer Science, Japan. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human–computer interaction.

Björn Granström is Professor of Speech Communication at KTH in Stockholm and the Director of CTT, the Centre for Speech Technology. He has published numerous papers in speech research and technology. His present research interest is focused on multimodal speech communication, including applications in spoken dialogue systems, rehabilitation, and language learning.

David House is an Associate Professor at the Department of Speech, Music and Hearing, School of Computer Science and Communication, KTH, Stockholm, Sweden. He received his Ph.D. in Phonetics at Lund University, Sweden, in 1991, and has held positions at the Department of Linguistics and Phonetics, Lund University; Department of Logopedics and Phoniatrics, Lund University; and the Department of Languages, University of Skövde. His research interests include speech perception, prosody, tone and intonation, speech synthesis, spoken dialogue systems, and multimodal speech perception.

Nancy Ide is Professor of Computer Science at Vassar College and Chair of the Computer Science Department. She has worked in the field of computational linguistics in the areas of word sense disambiguation and discourse analysis, as well as the acquisition and representation of lexical knowledge. She has been involved in work on the representation of language data and its annotations since 1987, when she founded the Text Encoding Initiative. She is the developer of the Corpus Encoding Standard, and currently serves as Project Leader of the ISO TC37/SC4 Working Group to develop a Linguistic Annotation Framework. Currently, she is Technical Manager of the American National Corpus project, and co-edits both the journal *Language Resources and Evaluation* and the Springer book series on Text, Speech, and Language Technology.

Margaret King is an Emeritus Professor of the School of translation and Interpretation in the University of Geneva, where she was Head of the Multilingual Information Processing Department until 2006. She has worked in the field of semantics and in machine translation. For the last 15 years she has had a special interest in the evaluation of human language technology software, concentrating particularly on systems dealing with written text. She was Chair of the European Union EAGLES working group on evaluation, and continued in this role in ISLE, a joint project of the European Union and the US National Science Foundation.

Wolfgang Minker is a Professor at the University of Ulm, Department of Information Technology, Germany. He received his Ph.D. in Engineering Science from the University of Karlsruhe, Germany, in 1997 and his Ph.D. in Computer Science from the University of Paris-Sud, France, in 1998. He was a Researcher at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), France, from 1993 to 1999, and a member of the scientific staff at DaimlerChrysler, Research and Technology, Germany, from 2000 to 2002.

Patrick Paroubek is a research engineer at LIMSI, a laboratory of the Centre National de la Recherche Scientifique (CNRS). He obtained his Ph.D. in Computer Science in 1989 from P&M Curie University (Paris 6) and specialized shortly after in the study of Natural Language Processing (NLP) systems. He was involved in the organization of several large-scale evaluation campaigns, in particular for parsing systems.

Laurent Romary is current Director of Scientific Information (STI) at CNRS. He got his Ph.D. in computational linguistics in 1989 and his Habilitation thesis in 1999. For several years, he has led the Langue et Dialogue research team (http://www.loria.fr/equipes/led/) at Loria laboratory and conducted various projects on human–machine dialogue, multilingual document management, and linguistic engineering. He was the editor of ISO 16642 (TMF – Terminological Markup Framework) under ISO committee TC37/SC3, and is Chair of ISO committee TC37/SC4 on Language Resource Management. He is also a member of the TEI council (Text Encoding Initiative; http:// www.tei-c.org).

Simone Teufel is a Senior Lecturer in the Natural Language and Information Processing Group at Cambridge University's Computer Laboratory. Her research interests include large-scale, applied NLP, information retrieval (IR), summarization evaluation and discourse parsing. During a Postdoc position at Columbia University in 2000–2001, she was also working on medical information retrieval. Her Ph.D. research (1999, Edinburgh University) looked at the application of discourse learning for summarization; her first degree, in Computer Science, is from Stuttgart University.

Introduction

Laila Dybkjær
Prolog Development Center A/S
Brøndby, Denmark
laila@pdc.dk

Holmer Hemsen
DFKI Language Technology Lab
Berlin, Germany
holmer.hemsen@dfki.de

Wolfgang Minker
Institute for Information Technology
University of Ulm, Germany
wolfgang.minker@uni-ulm.de

Evaluation has eventually become an important part of the general software development process and therefore also of text and speech systems development. The ELSNET summer school entitled "Evaluation and Assessment of Text and Speech Systems" in which this book has its origin, took place in 2002 and was seen as a timely event. Since then the interest in evaluation has continued to increase which we believe is a good reason for following up on the summer school with this book.

The field of text and speech systems is a very broad one comprising highly different types of systems and components. In addition, language resources play a central role in enabling the construction of such system and component types. It would take far more than a single two-weeks summer school or a single volume of a book to cover evaluation in the entire field of text and speech systems. We have therefore decided to let the book reflect the areas selected for the summer school that were meant to collectively illustrate the breadth of the field. They encompass both component, system, and data resource evaluation aspects and are among the most important areas in the field of text and speech systems.

A typical illustration of the architecture of a unimodal spoken dialogue system shows that it roughly consists of a speech recogniser, a natural language understanding component, a dialogue manager, a response or natural language

generation component, and a speech output component. Evaluation of speech recognition and spoken output – not least if a synthetic voice is used – is very important because the performance of these two components strongly influences the user's perception of the system in which they are embedded. We have included a chapter on the evaluation of speech recognition (Chapter 1), as well as a chapter on evaluation of speech synthesis (Chapter 2). The chapter on speech recognition evaluation also deals with evaluation of speaker recognition.

Text-based systems do not include any speech components. Natural language understanding may be needed at a more or less sophisticated level as illustrated by, e.g., question-answering versus information retrieval systems, cf. Chapter 6. This is also true for systems with spoken input where non-task-oriented systems and advanced conversational systems may require a much more detailed and complex understanding than, e.g., a simple bank account system. Natural language understanding is in itself a complex task. Chapter 4 addresses two important aspects of natural language understanding and their evaluation, i.e., part of speech tagging and parsing.

Some text and speech systems need a dialogue manager while others do not. For instance, machine translation does not require a dialogue manager because no dialogue interaction is involved. Whenever some understanding of the user's intentions is needed to find an appropriate reply, there is typically a need for a dialogue manager. For example, spoken dialogue systems normally have a dialogue manager and so do chat bots with written interaction. For simple tasks the dialogue manager may be very simple, whereas the more natural we want to make the conversation with the user the more sophisticated dialogue management techniques are needed. There is no separate chapter on evaluation of dialogue management but aspects of such evaluation are addressed in Chapter 7.

Response or natural language generation is a broad category that may vary considerably in complexity. There are response generation components that use simple pre-defined templates and there are those which try to do sophisticated generation of surface language from semantic contents. Response generation also includes, e.g., the return of a set of documents as in information retrieval tasks. There is no chapter explicitly dedicated to evaluation of response generation, but Chapters 6 and 7 contain elements of such evaluation.

Some or all of the components mentioned above, and maybe other components not mentioned here, may be put together to form a large variety of text and speech systems. We have included examples of evaluation of important system types, i.e., information retrieval and question answering systems (but not summarisation) in Chapter 6, and spoken dialogue systems in Chapter 7. Chapter 5 includes examples of evaluation of machine translation and spell-checkers but has a major emphasis on user-oriented aspects in general. Thus

Chapter 5 is closely related to human factors aspects although there is no separate chapter devoted to human factors in general.

If we combine text or speech with other modalities we would need one or more new components for each modality since we need to recognise and understand input modalities and generate information in the respective output modalities, and may also need to handle fusion and fission of input and output in two or more modalities. This book concentrates on text and speech and only takes a small peep at the multimodal world by including a chapter on talking heads and their evaluation (Chapter 3) and by describing evaluation of two concrete multimodal spoken dialogue systems (Chapter 7).

Nearly all the chapters mentioned so far demonstrate a need for language resources, be it for training, development, or test. There is in general a huge demand for corpora when building text and speech systems. Corpora are in general very expensive to create, so if corpora could be easily accessible for reuse this would of course be of great benefit. Chapter 8 addresses the collection, annotation, quality evaluation, and distribution of language resources, while Chapter 9 discusses standardisation of annotation, which would strongly facilitate reuse. There is no separate discussion of annotation tools.

No common template has been applied across chapters, since this was not really considered advisable given the state-of-the-art in the various subfields. Thus, each of the nine chapters follows its own approach to form and contents. Some chapters address component evaluation, others evaluation at system level, while a third group is concerned with data resources, as described in more detail below.

Component Evaluation. The first four chapters address evaluation of components of text and speech systems, i.e., speech and speaker recognition (Chapter 1), speech synthesis (Chapter 2), audio-visual speech via talking heads (Chapter 3), and part of speech tagging and parsers (Chapter 4). Some of these components may actually also, in certain contexts, be considered entire systems themselves, e.g., a speech synthesizer, but they are often embedded as components in larger text or speech systems.

Chapter 1 by Furui covers speech recognition, as well as speaker recognition. The chapter provides an overview of principles of speech recognition and of techniques for evaluation of speech and speaker recognition.

Speech recognition tasks are categorised as overall belonging to four different groups targeting human–human dialogue, e.g., interviews and meeting summarisation, human–human monologue, e.g., broadcast news and lectures, human–computer dialogue, such as information retrieval and reservation dialogue, and human–computer monologue in terms of dictation, respectively. Each category imposes different challenges on speech recognition.

Regarding evaluation of speech recognition the chapter has its focus on objective performance parameters although subjective measures in terms of, e.g., general impression and intuitiveness are briefly mentioned. Some performance measures, such as word error rate or recognition accuracy, are generally used across application types, while others are particular to a certain category of tasks, e.g., dictation speed for dictation applications. To compare the performance of different speech recognition systems, one must evaluate and normalise the difficulty of the task each system is solving.

Speaker recognition tasks are basically either speaker verification or speaker identification tasks, the former being the more frequent. A serious problem for speaker recognition is that the speech signal usually varies over time or across sessions. To overcome problems relating to such variations, different described normalisation and adaptation techniques can be used. Typical performance evaluation measures for speaker verification are described, including, e.g., equal error rate.

Chapter 2 by Campbell provides an introduction to speech synthesis and its evaluation, and to some of the attempts made over the years to produce and evaluate synthetic speech. There are three main stages in generating speech corresponding to three main components, i.e., language processing, prosody processing, and waveform generation. Approaches and challenges related to these three stages are described. Evaluation of speech synthesis is done both component-wise, as well as at entire speech synthesis system level, using subjective and objective measures. Evaluation of the language processing component mainly concerns the correctness of mapping between text and evaluation. Prosody is often evaluated using the Mean Opinion Score on a team of at least 30 listeners to obtain statistically significant results. The Mean Opinion Score can also be used to evaluate the waveform component.

Intelligibility has been the primary measure used to evaluate synthetic speech at the overall level. However, since synthetic speech nowadays normally is intelligible, naturalness and likeability have moved into focus instead. However, despite progress over the years, synthetic voices are still not like human voices. Control of paralinguistic information is a next challenge, i.e., control of non-verbal elements of communication, which humans use to modify meaning and convey emotion.

The chapter furthermore addresses concerted evaluation events and organisations involved in or related to speech synthesis evaluation. At the end of the chapter an extensive literature list is included, which the interested reader may benefit from when looking for further references within the area of speech synthesis.

Chapter 3 by Granström and House combines spoken output with an animated head. Focus is on the use of talking heads in spoken dialogue applications and on the communicative function of the head. Inclusion of an animated

face in a dialogue system affects the way in which users interact with the system. However, metrics for evaluating talking head technology are not yet well-established. In particular the need for a further exploration of the coherence between audio and visual parameters is stressed.

The chapter briefly explains face models and speech synthesis used at KTH and data collection aimed at a data-driven approach to talking heads. The rest of the chapter then describes various approaches to evaluation and a number of evaluation studies.

Speech intelligibility is important and studies show that it can be increased by adding an animated face. Not only lip movements play a role in a virtual face, but also eyebrow and head movements contribute to communication. Evaluation of these visual cues for prominence is described based on studies of their relation and individual importance.

Visual cues together with prosody seem to affect what users judge to be positive or negative feedback, respectively. A study is presented, which was used to evaluate the influence of these parameters on the users' perception of the feedback. A further study is reported in which the influence of visual cues and auditory cues on the perception of whether an utterance is a question or a statement was evaluated. Evaluation of emotion expressions of the animated face is also addressed together with prosody.

Finally, the chapter describes evaluation studies made with animated heads from two implemented dialogue systems. Evaluation among other things concerned socialising utterances, prosodic characteristics, and facial gestures for turntaking and feedback.

Chapter 4 by Paroubek deals with evaluation of part-of-speech (POS) taggers and natural language parsers both of which form part of natural language processing.

POS tagging is the process of annotating each word in the input text with its morpho-syntactic class based on lexical and contextual information. POS tagging is normally fully automated since tagging algorithms achieve nearly the same quality as human taggers do and perform the task much faster. Accuracy is the most frequently used performance measure for POS taggers. If taggers are allowed to propose partially disambiguated taggings, average tagging perplexity may be used as an appropriate measure. Also precision and recall are used for POS tagger evaluation and so is a combination of the two called the f-measure. Several other measures are also discussed, such as average ambiguity and kappa. There are further parameters, including processing speed, portability, and multilingualism that may be of interest in an evaluation, depending on its purpose.

Parsing is a much more complex task than POS tagging. Parsing may be deep or shallow, i.e., partial. For many tasks shallow parsing is entirely sufficient and may also be more robust than deep parsing. Common approaches

to parsing are briefly presented. An overview of how evaluation of parsers has been approached is given together with a description of measures that have been used and possible problems related to them. Examples of performance measures mentioned are percentage of correct sentences and recall. Evaluation campaigns may be used with benefit to comparatively evaluate a number of parsers on the same test suites.

System Evaluation. The following three chapters deal with evaluation at system level, including software evaluation standardisation with several linguistic examples, e.g., from machine translation (Chapter 5), information retrieval systems, and question answering systems (Chapter 6), and spoken dialogue systems (Chapter 7).

Chapter 5 by King presents general principles of user-oriented evaluation based on work done in the EAGLES and ISLE projects and on ISO standards, in particular ISO/IEC 9126 on software evaluation. Focus is not on individual metrics but rather on what needs to be considered or done before it makes sense to decide on particular metrics. A number of examples are given from the broad area of natural language software, e.g., machine translation.

Software quality must be evaluated in terms of whether the users' needs are satisfied. Therefore the users' needs must first be identified and evaluation criteria must be formulated that reflect their needs. User needs may differ widely depending on the task they need to carry out. Furthermore, the kind of evaluation to choose depends on the purpose of the evaluation. The kinds of evaluation mentioned include diagnostic evaluation, comparative evaluation, progress evaluation, and adequacy evaluation.

It may seem that every evaluation task is one of a kind. However, at a more abstract level there are characteristics, which are pertinent across evaluations of software quality. The six software quality characteristics from ISO/IEC 9126-1 are presented. They include functionality, reliability, usability, efficiency, maintainability, and portability. Each of these is further broken down into sub-characteristics, a few of which are explained in more detail. Software quality, as expressed via these (sub-)characteristics, influences quality in use, which in ISO/IEC 9126-1 is expressed in terms of effectiveness, productivity, safety, and satisfaction.

The ISO quality model may be specialised to take into account the particular software to be evaluated and made more concrete by relating it to the needs of the users. This is done by adding further levels of sub-characteristics. Not all sub-characteristics are equally important. The importance depends on the users and the task. Low importance may be reflected in a low user rating still being defined as satisfactory. The terminal nodes of the quality model must have metrics associated.

Chapter 6 by Teufel addresses evaluation of information retrieval (IR) and question answering (QA) in the context of the large-scale annual evaluation conference series TREC (Text REtrieval Conference), where many systems are evaluated on the same test collections. An IR system returns one or more entire documents considered relevant for the query. Perceived relevance is a problem here because it is subjective. A QA system is supposed to output a short piece of text in reply to a question. QA is a fairly new activity as opposed to IR and more difficult in the sense that a more thorough understanding of the query is needed. Nevertheless the best QA systems get close to the possible maximum score, which is far from the case for IR systems. In both cases evaluation is expensive due to the human effort involved.

On the IR side the chapter focuses on ad hoc document retrieval where queries are not refined. A fixed document collection is also assumed. Evaluation of IR systems is typically a performance evaluation, which involves a set of documents, a set of human generated queries, and a gold standard for what is relevant as decided on by a human judge. Real users are not involved. The primary metrics used are precision and recall, while accuracy is not a good measure. Recall poses a problem because it is basically impossible to go through perhaps a million documents to check that no relevant ones have been omitted. The pooling method may, however, be used to solve this problem.

Different question types are distinguished and included in TREC QA, e.g., factoid questions and list questions, whereas opinion questions are not included. All questions do not necessarily have an answer. There is no gold standard. Instead each answer is judged independently by two human assessors. Main metrics have changed over time from mean reciprocal rank over weighted confidence to average accuracy.

Chapter 7 by Bernsen, Dybkjær, and Minker concerns the evaluation of spoken dialogue systems. It first provides a general overview of evaluation methods and criteria and then describes evaluation approaches for two specific (multimodal) spoken dialogue systems in detail. The two systems addressed are the non-task-oriented Hans Christian Andersen (HCA) system and the in-car SENECA system, both developed in European projects.

Both systems are prototypes and they are described along the same lines, i.e., a brief system overview is provided followed by a description of the evaluation of the prototype, including test set-up, evaluation method, test users, and evaluation criteria.

For the HCA system technical, as well as usability evaluation criteria are mentioned. The technical criteria encompass criteria meant to evaluate more general system issues, e.g., real time performance and robustness, as well as criteria, which relate to specific components. The components addressed include speech recognition, gesture recognition, natural language understanding, gesture interpretation, input fusion, character module, response generation,

graphical rendering, and text-to-speech each of which have specific criteria assigned, such as lexical coverage for natural language understanding and intelligibility for text-to-speech.

Most of the usability evaluation criteria address (parts of) the entire system rather than a single component. For the HCA system examples include speech understanding adequacy, frequency of interaction problems, naturalness of user speech and gesture, entertainment value, and user satisfaction.

For the SENECA system no technical evaluation criteria are mentioned. Examples of usability evaluation criteria are glance analysis, driving performance, and user satisfaction.

Data Resources and Evaluation. The two last chapters address language resources and evaluation in terms of an overview of data resources and their creation (Chapter 8) and standardisation efforts regarding annotation of language resources (Chapter 9), respectively.

Chapter 8 by Cieri gives an overview of language resources. Data is known to be a crucial driving factor for the development and evaluation of text and speech systems and the current trend is that there is an increasing demand for ever more sophisticated, diverse, and large-scale data resources – a demand, which the data providers still fail to keep pace with. The data providers include researchers and research groups, as well as data centres. The most well-known data centres are probably the Linguistic Data Consortium (LDC) in the USA and the European Language Resources Association (ELRA) in Europe. These data centres create data but also distribute data created by others, e.g., by evaluation programmes and projects. A wide variety of resources are available via these centres, including data developed in support of speech recognition, speech synthesis, language and acoustic modelling, information retrieval, information extraction, summarisation, natural language processing, machine translation, speech to speech translation, and dialogue systems.

Data is used for development, training, and evaluation. If one can find and get access to available resources, this is normally cheaper than creating them from scratch. However, the suitability of the data must often be traded off against the cost of developing new data.

Data resources are costly to develop and several steps are involved. There is already a substantial effort in just collecting them and subsequently they are often annotated in various ways, which again requires a human effort. The challenges involved in collecting data differ depending on their type. Challenges in the collection of broadcast, telephone, and meeting data are discussed. These include, e.g., variable environments, multimodality, and overlapping speech. Subject recruitment is also discussed.

Several layers of annotation may be added to the collected raw data. Annotation may or may not require expert annotators but in many cases some level of

expertise is necessary. Data quality is important and annotations must therefore be evaluated. Precision, recall, discrepancy, and structure are the parameters used to evaluate annotation quality.

Chapter 9 by Ide and Romary also addresses language resources but from the perspective of standardisation of their representation and annotation. The ISO TC37/SC4 committee on language resources management has been established with this kind of standardisation in mind. The goal is to achieve an internationally accepted standard that will enable a much more flexible use, reuse, comparison, and evaluation of language resources than is the case today.

The chapter provides an overview of earlier, as well as ongoing projects that have dealt with various aspects of language resource standardisation, such as TEI, MATE, EAGLES, ISLE, XML, and RDF/OWL. The ISO group is building on such previous and contemporary work. The idea is to allow people to continue to still use their preferred annotation representation format and structure to the extent that it can be mapped to an abstract data model using a rigid dump format. To this end the Linguistic Annotation Framework (LAF) has been established to provide a standard infrastructure for the representation of language resources and their annotation. The underlying abstract data model builds on a clear separation of structure and contents.

On the contents side ongoing work is presented on creating a Data Category Registry, which describes various data categories, such as gender and its values. Again the idea is that users may continue with their own categories, which it should then be possible to convert to the standard categories. New standard categories may be added if needed. The idea is not to impose any specific set of categories but rather to ensure well-defined semantic categories.

An illustrative example is provided to explain how the presented standardisation work may enable a person to reuse annotations from two other persons who used two different coding schemes.

Much work still remains to be done on standardisation of language resources. However, the American National Corpus project has chosen the LAF dump format for annotation representation, which may be seen as a step in the right direction.