Testing and Assessment in Translation and Interpreting Studies

edited by
Claudia V. Angelelli
Holly E. Jacobson

ata

American Translators Association
Scholarly Monograph Series XIV

John Benjamins Publishing Company

Testing and Assessment in Translation and Interpreting Studies

American Translators Association Scholarly Monograph Series (ATA)

As of 1993 John Benjamins has been the official publisher of the ATA Scholarly Monograph Series. Edited by Françoise Massardier-Kenney, under the auspices of the American Translators Association, this series has an international scope and addresses research and professional issues in the translation community worldwide. These accessible collections of scholarly articles range from issues of training, business environments, to case studies or aspects of specialized translation relevant to translators, translator trainers, and translation researchers.

Volume XIV

Testing and Assessment in Translation and Interpreting Studies. A call for dialogue between research and practice
Edited by Claudia V. Angelelli and Holly E. Jacobson

Testing and Assessment in Translation and Interpreting Studies

A call for dialogue between research and practice

Edited by

Claudia V. Angelelli
San Diego State University

Holly E. Jacobson
University of Texas at El Paso

John Benjamins Publishing Company
Amsterdam / Philadelphia

TM
The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48–1984.

Library of Congress Cataloging-in-Publication Data

Testing and assessment in translation and interpreting studies : a call for dialogue between research and practice / edited by Claudia V. Angelelli, Holly E. Jacobson.
p. cm. (American Translators Association Scholarly Monograph Series, ISSN 0890-4111 ; v. XIV)
Includes bibliographical references and index.
1. Translating and interpreting--Evaluation. I. Angelelli, Claudia (Claudia V.) II. Jacobson, Holly E.
P306.2.T495 2009
418'.02--dc22 2009025926
ISBN 978 90 272 3190 1 (HB; alk. paper)
ISBN 978 90 272 8902 5 (EB)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Part III. Professional certification: Lessons from case studies

INTRODUCTION

Testing and assessment in translation and interpreting studies

A call for dialogue between research and practice

Claudia V. Angelelli and Holly E. Jacobson

Translation and interpreting are areas of inquiry supported by substantial scholarship. The notion of quality is central to both fields, whether at initial acquisition levels as a formative assessment in educational programs, or at more advanced levels in developing instruments for professional certification, as well as in measuring the quality of translation/interpreting for instruments and processes used for research purposes. Assessment and testing in the two fields is implemented for a number of purposes. Examples include screening applicants for entry into an educational program; providing feedback to students taking a course; testing knowledge and skills at the end of a course of study; carrying out quality assessments in contexts where interpreters play an essential role in achieving interactional goals; certifying professional-level competence in translation or interpreting; determining quality of localization products in the industry, as well as measuring the impact of surveys and other instruments translated for research purposes.

Most of the discussions around theory have focused on quality in theoretical terms, particularly in translation studies. Many of the established theoretical frameworks referred to in the translation literature are based on dichotomies or continua that distinguish between translations that closely adhere to the original linguistic code and more liberal translations that achieve a structure that is less subservient to that of the source text. Nida's (1964) concepts of formal and dynamic equivalence represent one of the first approaches to defining translation quality. His framework calls for determining quality according to the response a translation produces in target readers; that is, the response to the translation by target readers should be equivalent to the response to the original by source-text readers. In turn, Newmark (1982) uses the terms *semantic* and *communicative* translation to refer to a dichotomy that is similar to Nida's formal and dynamic equivalence. Likewise, Toury (1995) established a framework to refer to two types of translations, using *adequacy* to refer to a translation that closely adheres to the “norms” of the source culture, and *acceptability* to refer to translations that respond to the norms of the target culture. Venuti (1995) coined the terms *foreignization* and *domestication* as a means of underlining the need to examine unequal power relations that influence the way translations are realized, while Bastin, at a more pragmatic level, argued for adaptation rather than translation (1998). Skopos Theory (Reiss & Vermeer 1984, in Hatim and Munday 2004) emphasizes that the skopos or purpose of the translation is the measuring stick by which translation quality should be measured. In other words, a translation must be judged by whether it meets the linguistic, social, and cultural norms of the context in which it will be used. Later researchers, including Hatim and Mason (1990, 1997), Hickey (1998) and Baker (1992) turned to disciplines such as theoretical linguistics, pragmatics, sociolinguistics, and discourse analysis to inform models of translation and description of translation quality. These researchers grounded their models of translation in theoretical frameworks that allow for a deeper analysis of translated texts, with a focus on cross-linguistic differences in text types at the level of semiotics; pragmatics; socio-cultural context in which the original and source texts are used (non-verbal aspects of texts); and discursive elements. However, House (1981, 1997, 1998) was one of the first scholars to focus specifically on translation quality assessment, basing her work on pragmatics. She posits the existence of two types of translation, which she refers to as *covert* and *overt*. An overt translation is realized as a way of providing the target world a glimpse into the source world, or of “eavesdropping” on another culture or discourse community, and retains the integrity of the original socio-cultural context. It is obviously and overtly a translation. A covert translation, on the other hand, is used “to recreate an equivalent speech event” which meets the expectations and rules of the target discourse community (House 1998: 65). Like Bastin, Nida, Newmark, Toury, Reiss and Vermeer, and Venuti, House distinguishes between texts that are more closely associated with the source text and those that distance themselves from the original linguistic code in order to achieve functional pragmatic equivalence (House 2001). According to the models proposed by all of these scholars, quality depends on the purpose and function of the translation.

The pioneering work of these translation scholars recognizes the varied contexts in which translation is carried out, and moves away from more traditional views of translation that focus on a discourse of accuracy, which is defined by Zhong (2002: 575) as a paradigm “which requires translation to be accurate [on a lexico-semantic level], faithful [to the source text], objective, and impartial”. As House (2001: 247) states, “It is obvious that equivalence cannot be linked to formal, syntactic, and lexical similarities alone because any two linguistic items in two different languages are multiply ambiguous, and because languages cut up reality in different ways.” However, none of the models of translation quality presented thus far addresses the "how to" of effectively and accurately measuring quality. The researcher is left to ponder questions related to how "reader response" can be measured and compared; how to determine the variables that demonstrate whether a translation is acceptable to the target discourse community; or how the "function" of a translation is defined in measurable terms. These are all questions that have not been clearly addressed in the literature.

Testing and assessment of interpreter performance faces a similar dilemma. Research in interpreting focused traditionally on conference interpreting after the establishment of the International Association of Conference Interpreters (AIIC) in 1953. Early empirical research emerged within psychology, focusing on the cognitive processes of simultaneous interpreting (Pöchhacker 2004). In addition, as Hsieh (2003) points out, theoretical developments in simultaneous interpreting have primarily been driven by translation practices that focus on fidelity and accuracy. Interpreting practitioners have also played a key role in establishing models of interpreting based on the concept of "conduit" according to which interpreters are to remain neutral, detached, and faithful to the original (Ibid: 12). Community interpreting eventually adopted these theories, although research indicates that such theories do not accurately reflect how mediated interaction actually takes place (cf. Angelelli 2001 and 2004a; Clifford 2005; Davidson 1999; Metzger 1999; Roy 2000; Wadensjö 1998). However, few researchers have focused on measurement of aspects of interpreting in general, quality in performance specifically, and on the problem of assessing interpreting via the implementation of valid and reliable measures based on empirical research. A few scholars have ventured into this new territory. Angelelli (2001 and 2004b), for example, developed the first valid and reliable instrument to study the role that interpreters play in the various settings where they work (i.e. the courts, the hospitals, business meetings, international conferences and schools) in Canada, Mexico and the United States using psychometrics. On the basis of empirical data collected during an ethnography she developed an assessment instrument for use in healthcare contexts that measures language proficiency and interpreters' readiness in Cantonese, Hmong and Spanish (Angelelli 2003, 2007a and b). Sawyer (2004) conducted a case study on the measurement of translation and interpreting competence in a graduate level program in the United States, and started a discussion on the political and ethical consequences of test validation in translation and interpreting. Clifford (2005) developed an interpreter certification test grounded in discourse theory. He argues for more rigorous approaches to assessment, with empirically developed constructs and competencies, and for the exploration of psychometrics in developing certification instruments. Although other assessment instruments have been developed for organizations and interpreter education programs worldwide, these instruments are not generally presented in descriptive terms, and are not based on valid and reliable approaches to testing and assessment (Clifford 2005; Angelelli 2003, 2007a and b).

There is a lack of empirical research on both translator and interpreter competence and performance, and on assessing processes and products for different purposes, i.e. those of interest to industry, pedagogy and research. In addition, little has been published on the high-stakes certification programs and standards that exist in different countries: assessments seem to be conducted in a vacuum, and the processes involved need to be accurately described in order to assure transparency.

The idea for this volume emerged after a two-year series of conferences on testing and assessment during the ATA Research Forum. For the last five years, efforts were made to bring together professional organizations granting certification, academia, government and industry (free-lancers as well as agency owners) to find a space to discuss theoretical and empirical research in Translation and Interpreting Studies during a professional meeting. This is how the ATA Research Forum was established (Angelelli, 2004) within the American Translators Association. The editors are grateful to the presenters and the participants who at the time responded to a call (Angelelli 2006 and 2007) for the ATA Forum to focus on issues of testing and assessment, including the definition and the measurement of translation and interpreting competence and quality. Some of the presenters from the ATA Forum have contributed to this volume. In addition to seeding efforts at the Research Forum, the editors posted calls for papers in national and international scholarly websites and networks such as ATISA, EST, ITIT, The Linguist List, and at universities and colleges involved in interpreting and translation education and research. Editors also approached other scholars who had previously worked in the area of testing and posted the call for papers on professional association lists such as ATA and AAAL.

As suggested by the foregoing discussion, the present volume deals with issues of measurement that are essential to translation and interpreting studies. The collection of papers explores these issues across languages and settings (including university classrooms, research settings, the private sector, and professional associations), with a focus on both processes and products. All of the contributors are researchers and educators or doctoral students of either translation or interpreting – or both – who have focused on areas of testing and assessment. The authors have approached their chapters from different perspectives, some focusing on very specific variables, and others providing a much broader overview of the issues at hand. In some cases authors go into a more micro-perspective of measuring either translation or interpreting (e.g. the measurement of text cohesion in translation; the measurement of interactional competence in interpreting; the use of a particular scale to measure interpreters' renditions; or the application of a specific approach to grading). In other cases, authors present a broader view of program assessment (such as interpreter or translator certification at the national level or program admissions processes).

This volume is divided into three sections. The articles in the first section explore the theoretical underpinnings of assessing translation and interpreting, specifically as they relate to construct definition and rubric development. The articles in the second section discuss results of empirical research implementing quasi experimental and non-experimental designs. These studies delve into evaluation methods, including holistic/intuitive-impressionistic and analytical and dichotomous items-methods, and the application of evaluation scales to grading. They also provide insight into types of assessment (e.g. meaning-oriented) and assessment constructs (e.g. cohesion). The articles in the third section present case studies that are of a broader scope, describing admissions tests and professional certification.

The boundaries between sections are clearly fluid, and were established for the practical purposes of the volume only. One of the strengths of the volume lies in the fact that there are common threads running through all the chapters, that they are linked in a number of ways. All three sections contain chapters that include different approaches to testing (e.g. theoretical, empirical or descriptive); describe a variety of text purposes (e.g. admissions or certification) and test types (e.g. rubrics, evaluation scales; aptitude); and discuss different evaluation functions (e.g. formative, as for pedagogical purposes, or summative, as in high-stakes tests for certification). The chapters are also linked by the test constructs explored (e.g. translation or interpreting competence), and the approaches taken to measurement (i.e. norm-referenced or criterion-referenced tests, holistic, analytic, or dichotomous). Throughout the volume, authors argue for dialogue between research and practice.

The volume opens with the first chapter on theoretical concerns in testing and assessment. Claudia V. Angelelli discusses the basic questions that precede the development of a test. She argues for a need to ground translation tests in testing theory, and explores a construct definition that bridges the literature in Translation Studies, Testing, and Second language Acquisition. By clearly operationalizing the construct of translation competence, and based on a self-study of a translation organization, Angelelli proposes a rubric to measure the translation ability of candidates seeking professional certification across languages.

Holly E. Jacobson addresses similar concerns about the connection between theory and construct development in her discussion of healthcare interpreting. Grounded in concepts derived from interactional sociolinguistics and conversation analysis, Jacobson points to the need to develop a more comprehensive approach to assessing interpreter performance. She argues that current approaches to measuring interpreter-mediated interaction fall short in their emphasis on lexico-semantic concerns, at the exclusion of other linguistic and interactional features. This chapter offers a step-by-step approach for developing an analytic scoring rubric for assessing interactional competence in interpreting.

The second section, which focuses on empirical approaches in translation and interpreting assessment, begins with June Eyckmans, Philippe Anckaert and Winibert Segers' discussion on norm-referenced tests for translation. According to the authors, the calibration of dichotomous items (CDI) as a method for assessing translation competence transfers the well-known "item"-concept from language testing theory and practice to translation assessment, thus representing a rupture with traditional techniques of translation testing where the evaluator judges the value of the translation based on a series of pre-established criteria. The authors compare three approaches to translation evaluation on their psychometric qualities in a controlled empirical design, and contend that the CDI method is less subjective and more reliable.

Elisabet Tiselius explores the implementation of Carroll's (1966) scales for evaluating *intelligibility* and *informativeness* in interpreter performance. The author adapts Carroll's scales – which were originally devised by Carroll for machine translation – to the context of simultaneous interpreting between English and Swedish. She uses transcripts of interpreted renditions in a pilot study to compare the grading results of non-experts (non-interpreters) and interpreters. The preliminary data suggest that interpreters and laypeople do not differ significantly in how they approach the grading of simultaneous interpreting, which, if supported by future research on a larger scale, could have an impact on selection of graders in testing and assessment.

Mira Kim's contribution addresses the lack of systematic criteria to assess translations in the university classroom and the challenges faced by translation teachers, who need to assess students' translations and provide constructive, detailed feedback on assignments. Drawing on both qualitative and quantitative data in teaching translation from English to Korean, Kim elaborates on the pedagogical effectiveness of a meaning-oriented assessment tool which is based on systemic functional linguistics (SFL). She describes how meaning-oriented translation assessment criteria have been used in the classroom, providing detailed examples of the evaluation process.

Brian James Baer and Tatyana Bystrova-McIntyre propose the use of corpora to document the differences within language pairs which can provide an empirical basis for the formulation of assessment tools. Based on data collected from English and Russian, they argue for a granular assessment tool (replicable for other language pairs) to measure three isolatable – but nevertheless frequently ignored – features of textual cohesion: punctuation, sentencing, and paragraphing. They contend that focusing assessment on such textual elements can encourage novice translators to consider the target text globally, as a professional product composed of various features above and beyond lexis.

Turning to an underrepresented area in translation assessment, Kerian Dunne explores approaches to determining the quality of localization products. He considers some practical ways in which educators and practitioners can re-think assessment and find a common framework within which to discuss, evaluate, measure, and improve localization quality. He discusses perceptions and misperceptions which currently influence localization quality assessment, and points to their limitations. Through the provision of concrete examples, he explores possible solutions, and calls for further empirical research to inform the development of evidence-based assessment approaches.

The third section opens with the exploratory work of Šárka Timarová and Harry Ungoed-Thomas, who discuss the admissions tests for a particular interpreter education program, and argue for the need to carefully study the effectiveness of similar tests in screening applicants to IEPs in Europe. By applying multiple linear and logistic regression analyses to study the relationship between the IEP's admissions test and end-of-program exam, the authors conclude that this particular admissions test, aimed at measuring aptitude, is a poor predictor of students' success rate in the program. The research of these authors is exploratory in nature, and points to the need for IEPs to not only determine the predictive validity of their admissions tests, but also to submit their programs to overall program evaluations, including psychometric studies of entry and exit exams.

In a parallel investigation, Karen Bontempo and Jemina Napier also study admissions testing, drawing on data from two previous studies involving signed language interpreter education in Australia. One study analyzed the perceptions of interpreters-in-training regarding the efficacy of IEPs in Australia, while the other identified the gaps that exist in interpreter education: that is, the skills that are not addressed in IEPs. The authors apply this data to the development and piloting of an admissions screening test designed to measure six elements considered to be predictive of performance in an IEP. The pilot study involves a group of applicants to one particular signed language IEP. The results of the admissions test are compared with program exit outcomes: the authors argue that the test is not predictive of final examination performance in the IEP. They call for urgent review of current practices, and for empirical research that will inform the overhaul of the Australian national curriculum, including instructional quality, testing approaches, and resources for IEPs.

Hildegard Vermeiren, Jan Van Gucht and Leentje De Bontridder present a critical perspective and detailed overview of the spoken-language certification process of social interpreters in Flanders, Belgium. Given current trends in migration, the authors describe national efforts to offer quality service to those who do not speak societal languages. At the same time they contend that assessment and other similar rational procedures provide ideological self-legitimization to qualified authorities. The authors argue that consequential validity issues are a driving force in societal pressure for efficiency and accountability of the assessment procedure. The certification process described involves the implementation of what the authors refer to as objectifying elements, such as criterion-based evaluation grids, guidelines for scoring, pre-determined cut-off scores, and triangulation. They call for validity, reliability, and feasibility in interpreting assessment, and discuss inter-rater reliability and grader training in administration of certification exams. The authors argue that there exists a permanent need for improvement of test materials and procedures.

Debra Russell and Karen Malcolm also address the topic of national certification in their overview of the testing processes implemented in certifying signed language interpreters in Canada. Based on an evaluation of the testing system by the Association of Visual Language Interpreters (AVLIC), comprehensive and responsive test processes were developed to support interpreters in pursuing certification. The Canadian testing model is presented in detail, including the purpose of the test, test methodology and procedures, and test construction and piloting processes. The authors contend that the AVLIC test processes, which include an online written test of knowledge and personal feedback provided to candidates preparing for the test, and are situated in authentic discourse, constitute a model for professional certification. However, they too point to the dynamic nature of certification exams, and to the constantly evolving field of interpreting studies in their argument that the new certification model must be subject to ongoing refinements.

The issues discussed in this volume – the need for clearly defined and more inclusive constructs; the value of empirical analysis of current approaches to testing; the insistence on consistency in grading; the importance of constantly reviewing and refining assessment procedures – are shaping measurement approaches in translation and interpreting. They are relevant to the myriad of contexts in which the assessment of translation and interpreting is implemented, from the interpreting classroom to national certification exam development to the industry passing judgment on quality. A systematic response to these issues is required by key players, including teachers, administrators, practitioners and researchers. This response must be grounded in testing theory; it is a response that relies on testing principles in order to reliably inform practice. It is our hope that the contributions presented in this volume will serve to instigate discussion among those engaged in testing in our field.

References

Angelelli, Claudia V. 2001. "Deconstructing the Invisible Interpreter: A Critical Study of the Interpersonal Role of the Interpreter in a Cross-Cultural/Linguistic Communicative Event." (Doctoral dissertation, Stanford University). *Dissertation Abstracts International*, 62, 9, 2953.

Angelelli, Claudia V. 2003. *Connecting World Collaborative Testing Project*. Technical Report. Commissioned by Healthy House for a MATCH Coalition and The California Health Partnership. Funded by The California Endowment.

Angelelli, Claudia V. 2004. *Medical Interpreting and Cross-cultural Communication*. London: Cambridge University Press.

Angelelli, Claudia V. 2007a. "Longitudinal Studies and the Development of Assessment for Advanced Competencies." In *The Longitudinal Study of L2 Advanced Capacities*, In Lourdes Ortega and Heidi Byrnes (eds), 264–278. New York: Routledge.

Angelelli, Claudia. 2007b. "Assessing Medical Interpreters: The Language and Interpreting Testing Project." *The Translator,* 13(1): 63–82.

Baker, Mona. 1992. *In Other Words: A Coursebook on Translation*. London: Routledge.

Bastin, George. 1998. *¿Traducir o Adaptar? Estudio de la adaptación puntual y global de obras didácticas*. Caracas, Venezuela: Universidad Central de Venezuela.

Carroll, John B. 1966. "An Experiment in Evaluating the Quality of Translations." *Mechanical Translations and Computational Linguistics,* 9(3–4): 55–66.

Clifford, Andrew. 2001. "Discourse Theory and Performance-Based Assessment: Two Tools for Professional Interpreting." *Meta,* 46(2): 365–378.

Clifford, Andrew. 2005. "A Preliminary Investigation into Discursive Models of Interpreting as a Means of Enhancing Construct Validity in Interpreter Certification." (Doctoral Dissertation, University of Ottawa). *Dissertation Abstracts International,* 66, 5, 1739.

Hatim, Basil and Mason, Ian. 1990. *Discourse and the Translator.* London: Longman.

Hatim, Basil and Mason, Ian. 1997. *The Translator as Communicator*. London: Routledge.

Hatim, Basil and Munday, Jeremy. 2004. *Translation: An Advanced Resource Book*. London: Routledge.

Hickey, Leo (ed.). 1998. *The Pragmatics of Translation*. Clevedon: Multilingual Matters.

Hsieh, Elaine. 2003. "The Communicative Perspective of Medical Interpreting." *Studies in English Language and Literature,* 11, 11–23.

House, Juliane. 1981. *A Model for Translation Quality Assessment (2nd Edition).* Tübingen: Narr.

House, Juliane. 1997. *Translation and Quality Assessment: A Model Revisited*. Tübingen: Narr.

House, Juliane L. 1998. "Politeness and translation." In *The Pragmatics of Translation*, Leo Hickey, (ed). 54–71. Clevedon: Multilingual Matters.

House, Juliane. 2001. "Translation Quality Assessment: Linguistic Description versus Social Evaluation." *Meta,* 46(2): 243–257.

Metzger, Melanie. 1999. *Sign Language Interpreting: Deconstructing the Myth of Neutrality*. Washington DC: Gallaudet University Press.

Newmark, Peter. 1982. *Approaches to Translation*. Oxford: Pergamon.

Nida, Eugene. 1964. *Toward a Science of Translation*. Leiden: Brill.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London: Routledge.

Roy, Cynthia. 2000. *Interpreting as a Discourse Process.* Oxford: Oxford University Press.

Sawyer, David. 2004. *Fundamental Aspects of Interpreter Education*. Amsterdam: John Benjamins.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond.* Amsterdam: John Benjamins.

Venuti, Lawrence. 1995. *The Translator's Invisibility: A History of Translation*. London: Routledge.

Wadensjö, Cecilia. 1998. *Interpreting as Interaction*. New York: Addison Wesley Longman.

Zhong, Yong. 2002. "Transcending the Discourse of Accuracy in the Teaching of Translation: Theoretical Deliberation and Case Study." *Meta*, 47(4): 575–585.

PART I

The development of assessment instruments

Theoretical applications

Using a rubric to assess translation ability
Defining the construct

Claudia V. Angelelli
San Diego State University

One of the first and most important steps in designing an assessment instrument is the definition of the construct. A construct consists of a clearly spelled out definition of exactly what a test designer understands to be involved in a given skill or ability. This task not only involves naming the ability, knowledge, or behavior that is being assessed but also involves breaking that knowledge, ability or behavior into the elements that formulate a construct (Fulcher 2003) and can be captured and measured by a rubric. Currently, there is no one definition of translation competence and its components that is universally accepted within the academic field of translation studies (Arango-Keith & Koby 2003). Neither is there a rubric that can capture different levels of competency in translation. Instead, there is a continuing debate about how to define translation competence and exactly how its constituent elements are to be conceptualized, broken down, interconnected and measured. This paper reviews the literature from Translation Studies, Testing and Second Language Acquisition and proposes sub-components of a rubric to assess the construct of translation competence.

Introduction

Translation has been characterized as both a process and a product (Cao 1996), more pointedly a very complex process and product. The fact that translation is a multi-dimensional and complex phenomenon may explain why there have been few attempts to validly and reliably measure translation competence/ability. This is evident when comparing the research produced in translation testing with that produced in testing in related fields.

Translation shares some of the same linguistic concerns, such as discourse and grammatical competence in two languages (to name only a few), as the field of Second Language Acquisition. Translation also involves a variety of skills, including analytical skills and strategic skills, which are also present in other fields such as of Mathematics and others. When comparing the research produced in assessment within the field of Second Language Acquisition or in Mathematics, it is evident that we have not witnessed similar progress in assessment in the field of Translation and Interpreting Studies. This should not be interpreted as if the testing of translation or interpreting were not important enough or interesting enough to be worth the effort. On the contrary, developing a valid and reliable test for translation and interpreting is of paramount importance. Both academe and the industry would benefit enormously from making accurate and sound decisions on translation ability and quality based on meaningful testing.

Valid and reliable procedures for measuring translation or interpreting (or any other construct for that matter) start by posing essential questions about the procedures (Cohen 1994: 6) such as: for whom the test is written, what exactly the test measures, who receives the results of the test, how results are used, etc. Testing for both translation and interpreting share some similarities, specifically in the application of basic principles of measurement. But, because of the differences between these two the remainder of the discussion will focus solely on translation.

The answers to the questions about test procedure guide the test development and cannot be an afterthought. Test development also starts with a clear definition of what is to be measured, i.e. the test construct. Based on a case study of a professional organization certifying translators, this chapter starts by reviewing some of the relevant questions for translation test development. It uses the lens of testing and assessment to investigate the construct of "translation ability" and the potential use of rubrics to measure this construct. In so doing, it reviews how translation competence/ability has been defined by previous research and by professional ideology. Subsequently, it offers a working definition of the construct of translation ability for the specific purpose of translation certification. It argues for the use of rubrics to assess the translation ability of individuals seeking certification. It presents a rubric as a work in progress in the hope of contributing to relevant international discussions on valid and meaningful translation assessment.

1. Initial considerations

In this section I review the key questions in the assessment of translation as they apply to high-stake tests, such as translation certification examinations. The decision-making process of test developers' as to what to assess must be grounded in theory. For the purposes of translation assessment, I am suggesting that conceptualizations of communicative translation (Colina 2003) based on Skopos theory (Nord 1991 and 1997) be broadened to include concepts from cross-cultural communication and communicative competence theories (Bachman 1990;

Hymes 1974; Johnson 2001) to allow for decisions regarding what to assess based on broader principles applicable to translation.

1.1 Questions preceding test development

When test developers begin the process of creating a test, they are guided by the following questions (Cohen 1994: 11–48):

- *What* aspects of an individual's translation ability should be assessed?
- *Why* are certain techniques, assessment methods or approaches being used instead of others?
- *How* will the assessment instruments (translation tests) be developed, and how are they going to be validated?
- *When* will the test take place, and how often is the professional organization planning to administer it?
- *Where* will the exam take place and what is the physical environment(s) of the exam?
- *Who* is the intended audience for the test? What information is available about social backgrounds cognitive skills and personal characteristics (diverse or similar) of target audience?
- *For whom* are the results on the translation test intended; for candidates themselves or for organizations which make the exam a requirement?

So far I have presented relevant questions that pertain to the *Wh*-group. Outside *wh*-questions there are other important questions that test developers need to answer as they embark on the test-development process.

1.2 Nature of the test

Among further relevant questions there are those concerning the nature of the test to be developed. Is the test a norm-referenced or a criterion referenced one? This distinction is important since it allows for different things. "A norm-referenced assessment provides a broad indication of a relative standing, while criterion-referenced assessment produces information that is more descriptive and addresses absolute decisions with respect to the goal" (Cohen 1994: 25). The norm-referenced approach allows for an overall estimate of the ability relative to the other examinees. Norm-referenced tests are normed using a group of examinees (e.g. professional translators with X amount of years of experience, or translators who have graduated from translation programs 6 months before taking the test, etc.). In the criterion-referenced approach, criterion skills or behaviors are determined and then test specifications are written. This approach is used to see if a test taker has met certain objectives or criteria rather than to see how a test taker does compared to another. In addition to the nature of the test, whether is is a criterion-referenced or norm-references, other relevant questions test developers ask are about validity and reliability of the assessment instrument.

1.3 Validity

Traditionally, validity has been present in discussions on testing and test developers have raised questions such as: Is the test measuring what it is supposed to measure? (Kelly 1927; Lado 1961; & Cronbach 1971 in Weir 2005). Additionally validity has been discussed in different types, such as construct validity, content validity and face validity, among others (Bachman 1990; Bachman & Palmer 1996; Fulcher 2003; Messick 1989). As validity is multifaceted and multi-componential (Weir 2005), different types of evidence are needed to support any kind of claims for the validity of scores on a test. In 1985 the American Psychological Association defined validity as "the appropriateness, meaningfulness and usefulness of the specific inferences made from test scores (in Bachman 1990: 243). Therefore, test validity is not to be considered in isolation, as a property that can only be attributed to the test (or test design), nor as an all-or-none, but rather it is immediately linked to the inferences that are made on the basis of test scores (Weir 2005).

As an example, let's consider construct validity. This category is used in testing to examine the extent to which test users can make statements and inferences about a test taker's abilities based on the test results. Bachman and Palmer (1996) suggest doing a logical analysis of a testing instrument's construct validity by looking at the clarity and appropriateness of the test construct, the ways that the test tasks do and do not test that construct, and by examining possible areas of bias in the test tasks themselves. Construct validity relates to scoring and test tasks. Scoring interacts with the construct validity of a testing instrument in two primary ways. Firstly, it is important that the methods of scoring reflect the range of abilities that are represented in the definition of competency in the test construct. Similarly, it is important to ask if the scores generated truly reflect the measure of the competency described in the construct. Both of these questions essentially are concerned with whether or not test scores truly reflect what the test developers intended them to reflect.

The construct validity of a testing instrument can be affected by many of the same factors of the testing situation and the test tasks which create problems in reliability (see below). Therefore, it is important that all aspects of the testing situation and the test itself be considered for possible sources of bias for or against certain candidates. Bias here refers to any factor that may affect test performance that is not a part of the test's construct or testing objectives. The test designers must assure themselves that everything is done to make sure that the setting, instructions, input and expected responses do not end up influencing the test scores. Ideally, the only thing that should influence test scores is the candidate's competence, or lack thereof, as defined by the test's construct. With the question on validity comes the question of reliability.

1.4 Reliability

Reliability is one of the terms commonly used in the assessment field and is also fairly well-known outside of the field itself. It is not uncommon to hear lay persons using the word *reliability* or *reliable* and discussing what they judge reliability to be on a given issue or how reliable something is (e.g. a car or a brand). However, in the field of assessment and testing, the word *reliability* has specific meanings and set methods for its measurement. Primarily, *reliability* is used as a technical term to describe the amount of consistency of test measurement (Bachman 1990; Cohen 1994; Bachman & Palmer 1996) in a given construct. One way of judging reliability is by examining the consistency of test scores. If a person is given the same test at different times, will he or she score more or less the same? If different graders score the same examinee's test, will their scores be similar? If a person takes different versions of the same test, will they have similar scores on both versions? These and other questions reflect aspects of the consistency in test scores which test developers and testing experts are looking at when they examine the reliability of a test. However, reliability is not just about test scores. Creating a reliable test and judging the reliability of an existing test involves looking at the ways in which the consequences of factors outside of what is actually being tested, have been minimized to the greatest extent possible (Bachman & Palmer 1996). Factors of test administration and scoring procedures can be evaluated on the basis of how they might influence the reliability of the test in light of current thinking in the field of testing.

To determine reliability, we can use the questions for making a logical evaluation of reliability as set forth in Bachman & Palmer (1996). The factors impacting reliability are: (1) variation in test administration settings, (2) variations in test rubrics (scoring tool for subjective assessment), (3) variations in test input, (4) variation in expected response, and, (5) variation in the relationship between input and response types.

Both variations in the test setting and the physical conditions under which a test is administered can become possible sources of problems in reliability (Bachman & Palmer 1996). Due to the nature of national or international organizations' site selections, there will inevitably be some variation in test settings. Some test administrations may take place in a quiet office building. Others may take place in a room just off of a busy conference room. There will be natural variations in lighting, noise levels, temperature, and workspace available to a candidate, etcetera. These variations in settings can impact the individual candidate's ability to perform the test task or exercise. Some of these variations are unavoidable. In general, organizations should set minimum guidelines for the settings in which the examination is to be administered to minimize possible variations in performance on the examination due to environmental factors. It is also advisable to have test proctors go over a checklist of environmental factors and note any variance from the "ideal setting" so that these factors may be considered in the candidate's performance. Such data will also help the organization to analyze what, if any, of these factors play a role in variations found among the administration of the test.

When test developers or researchers are focusing on variations in the test protocol, the main question is whether or not there are variations in the way that instructions are given, in the time that is allotted for completion, or in the ways in which the scoring is done, which can influence the scores generated in an unanticipated or undesired way.

Variations in the wording of instructions are sometimes unavoidable for all versions of a certification exam, particularly if they are not all written in the same language. Therefore the question of consistency of test instructions across languages should be posed. At times, this may become a translation problem in and of itself. If the instructions are in fact consistent across languages then there would probably be no threats to the reliability stemming from this aspect of the test rubric. If, however, there is variation in the language and/or phrasing of the instructions on separate versions of the test, there is a possibility of some threat to reliability. Further study of the actual instructions would be needed in order to evaluate these variations for possible threats to reliability.

On the issue of time allotted for the test, care must be taken so that there are no variations in the time allotted due to variation in the performance of test proctors. Therefore, careful proctor training and the use of checklists by proctors are important elements in preventing threats to reliability.

Another area that may affect reliability is the manner in which the candidates' responses are scored. Scoring is ideally done based on fixed and objective criteria. Each instance of scoring by a grader should be similar to other instances of scoring that the same grader performs. This quality is known as *intra-rater reliability*.

Consultation among graders threatens the fixed and objective nature of scoring by threatening the *inter-rater reliability* (i.e. the fact that the same test, using the same scoring criteria and graded by different graders should yield similar results). Graders can pull each other one way or another. This is commonly known as grading by consensus. The most common practice in testing is for such variance in scores to be sent to a third and neutral party for an additional scoring. The averages of all the scores or the two that are in closest agreement are then often used as the final score. The most important factor for the sake of reliability, however, is that each scoring be done completely independently of the others in order to maintain the integrity of grading criteria, the scoring procedure and intra-rater reliability.

A third factor that may affect reliability is the variation in test input. Variation in the input given to test candidates can create problems in reliability when such variation has an unintended consequence on the performance on different parts or versions of a test. Therefore, when testers measure a candidate's translation ability, it is important to look at the ways in which the passages for translation are delivered to the candidates and what the qualities of those passages are. Are they formatted adequately? Are the fonts clear and legible? In order to keep the input consistent, it is advisable that the texts be retyped in a clear and legible font while maintaining as many features of the source material such as headings and paragraph flow, as possible. The desire for authenticity in passages may make the use of copies of passages in the original format attractive. However, this may be unnecessary and possibly create distracting visual variations. This should be avoided to the greatest extent possible. When working with authentic passages from source language materials, controlling the variation of the linguistic features of the passages to a high degree may be difficult. However, for the sake of the reliability of the examination, it might behoove an organization who is certifying candidates in multiple languages to consider which specific linguistic features it is seeking to test and to choose passages based on the degree to which they present those challenges. Sometimes organizations have panels of linguistic experts selecting the passages. Those experts should adhere to criteria designed for the specific selection. As we discussed in the *Wh*-section questions, criteria put forward for passage selection must be specific. For example an organization can give linguists a criterion such as 'make sure that each passage has common translation challenges such as false cognates, etc.', or 'check that passages are of a certain length' or 'select a passage that is generic, and one that pertains to the legal domain.' We may not think these criteria present any problem. However, in the design of a test with maximum reliability, it might be good to have a panel of linguists to analyze these texts for the ways in which they interact with the operational construct of translator competence so that the organization can know exactly which skills within the construct are being tested in each passage. With this information, a group (or bank) of passages will create a large variety that can be based on the skills tested. Also, any variation in reliability due to the variation in the examination passage can be anticipated and controlled.

An additional type of variation that may affect reliability is the expected response. When organizations require a complete written translation as its only response mode, this aspect of variation has no foreseeable effect on reliability. Some possible areas to be aware of as organizations move into electronic-format tests are issues surrounding consistency. The test should, as much as possible, either be all on computers or all hand-written.[1] Also, the ease of use of any word processing application used for the test should be examined and piloted for possible effects on reliability before implementation. Possible problems in testing electronically are variations in the manner of character entry, variation in editing features, and variation in the visual presentation of text from those encountered in commonly used and accepted professional tools. It is important to consider, that once an electronic test format is made available, it should be either used exclusively, or that measures be taken to minimize possible effects of written versus electronic response formats on test performance and grading across languages or sittings for the sake of maximum reliability.

In terms of variation in the relationship between input and response types, two factors are highly important between the versions and the test tasks: first, the way in which the questions are presented and second, the way a candidate is expected to respond. Again, given the format of translation certification exams, there is no anticipated danger of reliability being threatened by this type of variation. One thing to be cautious about in selecting a passage is the sudden changes in the text type or genre within a piece. For example, does the text suddenly go from narrative to dialogue? If awareness of such a change is being tested, this would be fine. If it is not the skill being tested, such a change may cause unanticipated difficulty for candidates, and this could threaten reliability.

1.5 Test authenticity

Another important aspect of testing is test authenticity. Authenticity is the term that the testing community uses to talk about the degree to which tasks on a test are similar to, and reflective of a real world situation towards which the test is targeted. It is important that test tasks be as authentic as possible so that a strong re-

[1] At the time of writing this paper, only the American Translators Association has conducted pilot exams on computers. Although a paper-pencil test poses threats to authenticity, most organizations are still certifying members this way.

lationship can be claimed between performance on the test and the performance in the target situation. If test tasks are too different from the situations in which candidates will be employing the competence being tested, the possibility that good or bad performance on the test does not reflect the ability to perform in the target situation increases (Bachman & Palmer 1996). For example, if we are testing for the ability to translate but we require the candidate to discriminate between acceptable and unacceptable translations, there is a high possibility that candidates who are strong in identifying good or bad renditions would succeed. By the same token, there may be a disadvantage for candidates who are better in producing translations rather than identifying good or less than satisfactory renditions. Therefore, it is important that the target situation in which we expect the candidates to perform be clearly defined and that the test tasks mirror that situation as clearly as possible. This issue is particularly relevant for professional associations that grant certification and use in-house tests. In general professional organizations are composed primarily of working professional translators. Members probably know the real world context in which translation competence is applied better than any test developer. However, knowing a situation intimately and defining it clearly for testing purposes are two very distinct things. Definition of a target construct often takes a particular kind of expertise that is different from the expertise of a practitioner. The practitioner is in the midst of the target situation and sometimes fails to notice aspects of the situation merely because they are taken for granted. Much of what translators do, they do automatically, and therefore, unconsciously (Toury 1995).

In terms of defining the target situation, some ideas about what the primary aspects of the target context entails were set forth previously in the suggested definition of the construct of translation competence. Professional translators deal with a wide variety of source texts that were produced in diverse contexts. They are contracted by clients and agencies with diverse needs. They use a variety of tools and can work individually or as part of teams. All of these aspects form a part of the target-use situation. Although not all of these aspects can be matched with equal efficacy in an examination, it is important that the testing tasks reflect as many of these aspects as possible.

1.6 Task authenticity

When developers are creating a test, another important aspect of task authenticity is the test format, and its impact on the security of the test. When organizations require test takers to produce a translation, this task is reflective of the type of task that a professional will perform in the target situation. The main areas in which a test task may seem to be inauthentic are the response format, the availability of tools, and the lack of time for editing. Some of these are authenticity problems which are logistically difficult to solve. The handwritten nature of the response format is seen as being fairly inauthentic for most contemporary translation workplaces. However, this is not a simple problem to solve as there are implications for test security and fairness, among others. It is possible, with current technology improvements, to disable e-mail functions temporarily and prevent the exam from leaving the examination room via the internet. Additionally exam proctors can be asked to control the environment and not allow electronic devices such as flash drives into the testing room to avoid downloading exam originals. To increase authenticity, it is important that the computerized test format mirror the tools and applications currently used by professional translators as closely as possible while maintaining test security. It is important that the word processing interface be as similar to the industry standard as possible. Available tools should also be as similar to real world working situations as possible. Since complete internet access could compromise test security (e.g. candidates could e-mail translations to each other during the test), it is understandable that to a certain degree, the format offered to examination candidates would lack authenticity (e.g. current certification exams such as ATA or NAATI are done by hand). However, creative solutions should be sought to minimize this lack of authenticity to the greatest degree possible.

The concept of translation assessment may or may not include translation, editing, and proofreading as separate tasks requiring different skills, and therefore different measurements. If an organization chooses to measure them jointly, this decision needs to be addressed by weighing categories and grading procedures, as well as in the test instructions to candidates. It is only when the test developers have considered all these elements that the test development can begin. Undoubtedly, the process begins with the question asking *what*; which is asking what the test assesses. This requires a clear definition of the test construct.

2. Defining the test construct

A construct consists of a clearly spelled out definition of exactly what a test developer understands to be involved in a given ability. If we are testing an ability to translate, it is important that we first clearly and meticulously define exactly what it is that we are trying to measure. This task not only involves naming the ability, knowledge, or behavior that is being assessed but also involves breaking it down into its constituent elements (Fulcher 2003). Thus, in order to measure a translator's professional ability in translating from one specific language into another, we need to first define the exact skills and sub-skills that constitute a translator's professional ability. In order to design and develop a test that assesses the ability to translate at a professional level, we have to define what the translation ability is. We have to operationalize it. The goal is to consider what type of knowledge and skills (in the broadest sense) might contribute to an operational definition of 'translation ability' that will inform the design and the development of a test of translation competency (Fulcher 2003). That is, we must say exactly what knowledge a translator needs to have and what skills a candidate needs to have mastered in order to function as a qualified professional translator. These abilities cannot be vague or generic.

To illustrate this we look at definitions (operationalizations) of translation competence. One definition of translation competence (Faber 1998) states the following: "The concept of *Translation Competence* (TC) can be understood in terms of knowledge necessary to translate well (Hatim & Mason 1990: 32f; and Beeby 1996: 91 in Faber 1998: 9). This definition does not provide us with specific descriptions of the traits that are observable in translation ability, and therefore it does not help us when naming or operationalizing the construct to develop a test.

To find an example of a definition developed by professional organizations, we can look at the one published by the American Translators Association (ATA). The ATA defines translation competence as the sum of three elements: (1) comprehension of the source-language text; (2) translation techniques; and (3) writing in the target language. In a descriptive article, Van Vraken, Diel-Dominique & Hanlen (2008 http://www.atanet.org/certification/aboutcert_overview.php) define criterion for comprehension of the source text as "translated text reflects a sound understanding of the material presented." The criterion for translation techniques is defined as "conveys the full meaning of the original. Common translation pitfalls are avoided when dictionaries are used effectively. Sentences are recast appropriately for target language style and flow." Finally, evaluation of writing in the target language is based on the criterion of coherence and appropriate grammar such as punctuation, spelling, syntax, usage and style. In this professional organization, the elements being tested (according to their definition) are primarily those belonging to the sub-components of grammatical competence (language mechanics) and textual competence (cohesiveness and style). But while this definition is broader than that of Beeby, Faber, or Hatim and Mason (in Faber 1998), it still does not account for all the elements present in the translation task required by their test.

We could argue that translation involves various traits that are observable and/or visible which include, but are not limited to conveyance of textual meaning, socio-cultural as well as sociolinguistic appropriateness, situational adequacy, style and cohesion, grammar and mechanics, translation and topical knowledge.

These traits contribute to an operational definition of translation ability, and they are essential to the development of a test.

A test can only be useful and valid if it measures exactly what it intends to measure; that is, if it measures the construct it claims to measure. Therefore, for a translation test to be valid, it must measure the correct construct, i.e. translation ability. The first crucial task of the test developer is to define the construct clearly. Once the construct is defined clearly, then and only then can the test developer begin to create a test that measures that construct. Evidently, this process is not linear in the sense that the construct undergoes revisions and modifications, but its definition does need to occur *a priori* (Bachman 1990). As evident from testing principles, a central issue in assessment is construct validity. Establishing construct validity ensures that the right construct is being measured. In the next section we will review how the construct of translation competence has been conceptualized.

3. Review of relevant literature

A good translation is a highly complex activity that involves many diverse areas of knowledge and skill. Therefore, defining translation competence is not an easy task. It is a "dynamic process and it is a human and social behavior" (Cao 1996: 231) that results from experience, training and the feedback effects of client-translator or translator-reader interaction. (Neubert & Shreve 1992: 10 in Cao 1996: 231). Currently, there is no one definition of translation competence and its components that is universally accepted within the academic field of translation studies (Arango-Keith & Koby 2003). In fact, there is considerable debate about how to define translation competence and exactly how its constituent elements are to be conceptualized, broken down and interconnected. Despite this disagreement, the academic discussion about translation competence can be an important aid in helping to define the constructs of what makes a competent and professionally qualified translator.

As Kiraly points out "An empirical description of translation processes implies the possibility of describing what a professional translator has to know and has to do (even if much of what he or she does is subconscious) to produce a high-quality translation." (1995: 13). To begin, Wolfram Wilss (1982 in Kiraly 1995) initially described translation competence as consisting of three primary components which include (a) source language receptive competence coupled with (b) target language reproductive competence operating within (c) a super-competence which reflects the ability to transfer the message from one language to another. This description of translation competence emphasizes that it is not merely enough to transfer a message from one language to another, but rather that there is a need to be strategic about it (Valdés and Angelelli 2003). Presas (2000) helps us further define the idea of Wilss' super-competence by defining what it is that makes translation competence different from bilingual competence. She emphasizes that a competent translator uses specialized linguistic and cultural knowledge to control interference in both the reception of information from the source text and the production of the target text. According to Presas, the competent translator does this in part through making a transfer at the level of meaning rather than at the level of words and phrases between two connected but separate code systems, i.e. languages. However, in order to validly and reliably test these specialized skills and knowledge it is necessary to define them further.

Many contemporary definitions of translation competence view translation as a specialized sort of communication. They define the translator as an individual who is interpreting a text that was written to perform a function in the source language and culture while delivering it into a new form in order to perform a function in the target language and culture (Kiraly 1995; Cao 1996; Neubert 2000; Beeby 2000; Orozco 2000; Adab 2000; Colina 2003). This type of functional approach to translation views translation competence as a specialized type of communicative competence. This concept of a communicative competence comes from the field of Second Language Acquisition (SLA).

Although the fields of SLA and Translation Studies, despite focusing on similar phenomena, have not historically engaged in the sharing of knowledge and theories, greater cross-fertilization between the two has occurred in recent years. The beginnings of this can be seen in more recent works on teaching and testing for translation and interpreting (Angelelli 2000, 2003, 2004b and 2007a and b; Schäffner and Adab 2000; Colina 2002, 2003 and 2008).

SLA theory also interacts with testing theories, especially in reference to testing language abilities and analytical skills (Bachman 1990; Bachman & Palmer 1996; Cohen 1994; Fulcher 2000; Johnson 2001). Therefore, it is important to have an understanding of these theories in communicative competence in order to frame a construct of communicative translation competence that allows us to create a theoretically sound assessment.

Among the most commonly used models of communicative competence in the fields of SLA and language assessment is that proposed by Bachman (1990). His model of communicative competence is divided into organizational competence, which involves the more mechanical aspects of communication and how they are organized; and pragmatic competence, which deals with how language is used in specific situations. Bachman further subdivides organizational competence into grammatical and textual competences. Grammatical competence is composed of the individual's knowledge of the forms and structures of a language.

Textual or discourse competence refers to the way in which sentences and larger chunks of language are woven together to create a coherent and cohesive message. Pragmatic competence is further divided into illocutionary and sociolinguistic competences. Illocutionary, or strategic competence consists of the individual's knowledge of the ways in which language is used to accomplish functions and create impressions and relationships. Sociolinguistic competence is an individual's knowledge of ways of speaking and interacting through language, e.g. politeness, taboos, etc. (Bachman 1990). These different competences are used in any given act of communication. An act of translation, by virtue of being an instance of language use, is a form of communication. Therefore, these communicative competences cannot be disregarded in the construct of translation competence.

For the current discussion on the construct of translation competence, a logical starting point on communicative translational competence is to consider the definitions proposed by Cao (1996), Colina (2003), and PACTE (in Orozco 2000). In Cao's model of translation proficiency, there is a translational language competence (defined similarly to Bachman's language competence), in addition to translational knowledge structures, such as world and specialized knowledge. There also exists a translational strategic competence which is connected to both of the other two, as well as the situational context. Thus, the core of translation competence lies in matching language competence and knowledge structures to the current communicative context (i.e. the translation task). This is achieved through the application of competence in translational strategies in order to execute a communicative translation task. This model helps us to see that translation lies not only in the ability to effectively convey a message between languages but also the ability to do so in a particular context. A translation needs to be both a good rendering of the source text and the proper rendering to meet the needs of the receiver. Being able to produce the right translation in the right context is therefore seen as a part of translation competence.

Colina (2003) defines communicative translational competence as consisting not only of communicative competence in both languages, but also including an element of interlingual and intercultural communicative competence. Colina emphasizes that translation is a communicative interaction in as much as the translator is responsible for the interpretation of source text (ST) meaning and its negotiation and expression in accordance with task specifications, translational conventions, and target language conventions. Thus, the model of translation competence in her work considers the ways in which the context of a translation also operates on the other competences in communicative translation competence. The model that Colina chooses to reflect these views is one put forth by Cao (1996).

Just as Colina's (2003) model adds the element of context to our understanding of translation competence, the PACTE model as outlined in Orozco (2000) adds the element of methods of achieving communicative translation goals. The PACTE model presents two major competences: transfer competence and strategic competence. Transfer competence is defined as the "the ability to complete the transfer process from the ST to the target text (TT), taking into account the translation's function and the characteristics of the receptor" (Orozco 2000: 199). This competence is further broken down into comprehension competence, the ability to de-verbalize the message and control interference, re-expression competence, and competence in choosing the most adequate method. Transfer competence is seen as being informed by four other competences: communicative competence in two languages, extra-linguistic competence (i.e. world and specialist knowledge), psycho-physiological competence (i.e. using various cognitive, psychomotor and attitudinal resources), and instrumental-professional competence (i.e. the ability to use the tools and apply the norms of the profession). The final element in this model is strategic competence in which all these processes are used in finding and solving problems in the translation process (Orozco 2000).

The standout feature of Orozco's model is the emphasis placed on tools and processes for problem-solving. Competent translators need to be able to find and correct problems in their own translations and processes. It is also important that they are familiar with the tools and standards of their trade. The strategic use of software, on-line residing tools, and more traditional items like dictionaries, are an important part of any translator's work. Even more importantly, knowledge of common practices and standards of professional conduct are also vital to competent translation. These are the tools that translators use to overcome translation problems and, therefore, form a vital part of the competence of a professional translator.

In addition to the contributions of the models of communicative competence and translational language competence, we need to look at the mode in which language is used. There are different modes through which the overall language competence in each language is engaged in the communicative act of translating. The American Council of Teachers of Foreign Languages through the National Standards in Foreign Language Education Project (2006) has defined three primary modes of language use according to the level of contact and interaction between the participants in the act of communication: the interactional mode, the interpretive mode, and the presentational mode. The interactive mode involves situations in which all participants can participate as both the presenter and the audience in conversations. The interpretive mode refers to situations in which the language user is primarily receiving information and the original presenter/writer is not available to clarify questions, similar to what occurs while reading or listening to text in a target language. The presentational mode involves situations in which the language user is primarily presenting information (either orally or in writing) with limited possibilities to interact directly with the eventual recipient of the message such as writing a document or presenting to a large audience. Of these modes, the interpretative mode (specifically reading) plays a greater role in relation to the translator's access to source text, while the presentational mode (specifically writing for a large readership) plays a greater role in the translator's production of the target text.

Despite the difference in modes, many of the underlying sub-competences are similar. A candidate seeking translation certification must have control and awareness of the various aspects of the source language in order to competently interpret the meaning of the source text. This means viewing the text through the cultural lenses with which it was produced. This also includes: (1) the grammatical aspects of a language which encompass the ways in which words and word parts are combined to form meaning; (2) the textual aspects of the language which include the conventions for how the information in a text is linked together, structured, and presented; and (3) the pragmatic aspects which include the culturally specific limitations on what is said, how it is said, and how these create feelings and establish relationships with the readership and the subject matter. Without an understanding and awareness of the subtleties represented in these different aspects of language, a candidate is not able to fully comprehend either the source text or how these elements will affect the translation task. Similarly, a candidate must have control and awareness over the grammatical, textual and pragmatic aspects of the target language in order to competently produce a target text. These are complex skills and abilities.

In light of the current literature on what constitutes translator competence and communicative translation competence, we turn our attention to the construct as it is defined by available documents from professional organizations granting certification. This definition of the current construct is drawn from both the overt explanations of what is being tested and the implicit priorities set forth in such documents as grading protocols when available.

In a set of ATA documents written to inform the general public about the nature of translation,[2] we can see how the professional organization conceptualizes translation. The ATA through its publication "Translation: Getting It Right," emphasizes the aspect of a translator's strategic (Orozco 2000) and intercultural communication (Colina 2003) competence by mentioning how translators can bridge cultures by encouraging consumers to tell the translator what the transla-

[2] At the time of writing this chapter, the author consulted the websites of various professional associations administering translation tests, such as ATA, NAATI, NAJIT. Information on the conceptualization of the test construct was only available at ATA website.

tion is for. This shows an emphasis on the pragmatic and functional competences that professional translators possess.

However, the article mentioned earlier by van Vraken, Diel-Dominique and Hanlen about the ATA certification exam does not discuss the purpose of the translation as it mentions the three elements that the exam measures. Those are: (1) comprehension of the source-language text, (2) translation techniques and (3) writing in the target language. The article defines criterion for comprehension of the source text as "translated text reflects a sound understanding of the material presented." The criterion for translation techniques is defined as "conveys the full meaning of the original. Common translation pitfalls are avoided when dictionaries are used effectively. Sentences are recast appropriately for target language style and flow." Finally, evaluation of writing in the target language is based on the criterion of coherence and appropriate punctuation, spelling, syntax, usage and style.

The current three part construct used by the ATA seems to primarily emphasize the reading comprehension, translation ability (not operationalized) and the micro-linguistic elements of translation competence present in writing (e.g. lexicon, grammar and punctuation rather than discourse, cohesion, etc.). The first element of this construct, the comprehension of the source text, combines aspects of both the organizational and pragmatic competences as defined by Bachman (1990). In order to comprehend a source text, a translator must be able to both make sense of the letters, words, phrases, sentences and text as a whole and understand that text in terms of what it means in the original culture. Therefore, in order to make this concept more reliably measurable, we need to break it down further. The second element of translation "technique" fits in with Cao's (1996) translation knowledge structures and Orozco's (2000) transfer competence. This is another example in which, in order for the sub-components of translation "technique" to be more reliably measured, we need to break them down and outline them as specific behaviors and observable aspects of a translated text. The final aspect of this construct, writing in the target language, is focused primarily on the micro-elements of the competence when translating, such as grammar, lexicon and punctuation.

The construct as defined by professional associations (NAJIT, ATA) seems somewhat more limited than the communicative constructs that have been established in the fields of language testing, SLA, and translation studies that take into account more macro elements of cross-linguistic communication such as purpose of the translation, readership, cohesion, register, etc. The language element is quite prominent in the professional organizations' construct and communicative translation competence is not fully represented. The elements being tested under professional organizations definition are primarily those belonging to the sub-components of grammatical competence (language mechanics) and textual competence (cohesiveness and style) of Bachman's (1990) model. The pragmatic competence defined by Bachman (1990) is only partially represented in the comprehension and rendering of the passage. Candidates are only judged on their understanding of the "full meaning and purpose" (not defined), of the source text and their ability to convey that meaning (again, not defined) in the target text. It is also problematic that comprehension in one language is being tested through production in another language. One could argue that the two could be unrelated. (A better test of comprehension might be to have the candidate summarize the text in the source language, or complete a reading comprehension exercise.) Also, it appears that there is no testing as to whether the candidate can perform the communicative functions necessary to produce a text that is appropriate to the target language readership. This raises the question: is the focused emphasis on grammatical competence appropriate?

Additionally, many current tests (e.g. ATA, NAATI) which certify translators ask for the delivery of a "neutral" translation that mirrors the source text and does not take a particular audience into account. Many times candidates are discouraged from making changes in the style, register and the use of regionalisms although these may be communicatively required in certain translation situations. Brief test instructions (e.g. this text is going to be published in Readers' Digest) designed to save space or time may result in a test-taker having no clear target readership or purpose, which in turn does not allow a candidate to show the best sample of his/her ability and does not allow the grader to make a judgment about the candidate's control of register or regionalisms. Similarly, the lack of a specified audience and function for the translation does not allow for the measurement of illocutionary competence (the ability to achieve functions with language, e.g. apologize, challenge, etc.) and sociolinguistic competence (culture specific references and what is allowed or disallowed in a given context in a given culture). The inclusion of these elements in other materials (e.g. translation brochures such as ATA Translation: Getting it Right) about good-quality translation suggests that they are essential and should be included in the construct of assessment for certification.

4. An expanded framework for an expanded construct

Given what research in the areas of communicative competence, sociolinguistics, translation studies, second language acquisition and testing have shown in addition to what the professional associations granting certification state, how should the construct of communicative translational competence be defined for the specific purpose of a certification examination? What sub-components of this construct are being measured by tests currently used? What sub-components should be part of a certifying exam? What sub-components can be tested separately? What separate tests, if any, could be used as predictors of success in certification exams? It seems that, partially, associations refer to an operational construct of translation competence that is already functionalist and communicative in nature (e.g. translation brochure ATA Translation: Getting it Right). However, when it comes to defining it (Van Vraken, Diel-Dominique & Hanlen 2008) , the tendency is to focus more on the grammatical and textual competences. Is this a problem? If so, we need to ask ourselves why. While comparing professional associations' ideologies on translation competence to what research in translation studies state, we see a gap. When operationalizing the construct, professional associations tend to have a narrower definition of translation competence, and many times pragmatic and other elements are not included in the construct to be measured.

The operational construct needs to be articulated along similar lines to those used in translation studies in order to capture translation in its entirety and thus properly measure it (Bachman 1990). To this end, I will propose a working definition of the construct of communicative translation competence that includes the communicative elements of Hymes (1974) in addition to Bachman's (1990) frameworks of communication and communicative competence, and the contributions of Cao (1996), Colina (2003) and some of the instrumental elements reflected in the PACTE definition of translation competence. This new measurable construct includes the following sub-components: (1) linguistic competence, (2) textual competence, (3) pragmatic competence, and (4) strategic competence. I do not presume to present this construct as a definite operationalization of translation abilities. This construct is presented as a guide, as a lead to chart directions for research and development in translation assessment, specifically as it pertains to professional associations granting certifications. As research develops, and as we subject this framework to empirical tests, it is likely that it will undergo changes to reflect our collective growing knowledge in the area. Let us look at the subcomponents of the construct.

1. *Linguistic-level competence*. The first sub-component of our construct is the linguistic component defined here in its narrowest sense. Translation is, in many ways, the communicative act of transferring a message between two languages. This activity requires a certain degree of communicative competence in two languages. In all of the models from Translation Studies previously reviewed (Cao 1996; Colina 2003 and Orozco 2000), competence in two languages is an undisputed aspect of translation competence (the same holds true for interpreting – see Angelelli 2003 and 2007b). In language assessment, the dominant models for language competence are the ones set forth by Bachman (1990) and Johnson (2001). Bachman's (1990) model forms the basis for Cao's (1996), Colina's (2003)

Translational Language Competence and Angelelli's model of Language Competence for Interpreting (2003) which is also combined with Johnson's (2001). Cao's notion of organizational competence (1996) includes grammatical competence and textual competence.

Clear grammatical competence plays a vital role in the act of translation. Cao defines this sub component as control of vocabulary (the words of a language), morphology (the way that smaller parts combine to form words), syntax (the way that words combine to form phrases and sentences), and graphemic (writing system) knowledge.

Each of these aspects contribute both to the interpretation of the source text and the production of the target text. A breakdown in any one of these areas can affect the act of translation. Insufficient knowledge of vocabulary can lead to miscomprehension of the source text or failure to successfully communicate meaning in the target text. This competence can be aided through proper and skillful use of dictionaries and other professional resources, but only to a degree. A translator's knowledge of morphology and syntax helps both interpretation and production in the act of translation. A failure to understand the effects that syntax and morphology have on meaning can also lead to incomplete or mistaken understanding of the source text and the production of a difficult or misleading rendition in the target text. Graphemic knowledge, likewise, plays a part in both the interpretation and production aspects of translation. Failure to understand differences in meaning carried by punctuation and diacritical marking can lead to misapprehension of the source text. Lack of knowledge of writing mechanics or misapplication of graphemic knowledge can lead to interference with communication of meaning and difficulty in the comprehension of a target text. Therefore, each of the aspects of grammatical competences is vital to the act of translation and is being assessed either directly or indirectly in any authentic translation task.

To be measured, this linguistic sub-component of translation competence needs to be clearly stated. To assess it during certification exams, for example, one can start by considering a continuum of more to less successful translations and describing what they would look like, how those would reflect more or less mastery of this subcomponent. Table 1 illustrates statements in a possible 5-point-scale to assess the linguistic sub-component of translation competence.

2. *Textual competence.* Textual competence, or the ability to string ideas together as a text, is also a vital part of any act of translation. Within the purview of textual competence, Cao includes cohesive competence, the ability to use linguistic devices to connect sentences, ideas, rhetorical organization competence, and the ability to organize a text in the most appropriate way to achieve its aims in a given community. In fact, Colina (2003) points out that the literature suggests that one

Table 1. Linguistic sub-component (T = translation; TL = target language)

5	T shows a masterful control of TL grammar, spelling, and punctuation. Very few or no errors.
4	T shows a proficient control of TL grammar, spelling, and punctuation. Occasional minor errors.
3	T shows a weak control of TL grammar, spelling, and punctuation. T has frequent minor errors.
2	T shows some lack of control of TL grammar, spelling and punctuation. T is compromised by numerous errors.
1	T exhibits lack of control of TL grammar, spelling and punctuation. Serious and frequent errors exist.

of the skills that separates the professional and the novice translator is the attention to textual clues. A successful translator must understand how a source text is structured internally and the effects that such an organization has on the meaning that the author is creating and communicating. Likewise, the successful translator activates his/her knowledge of similar available devices in the target language to render a similar message and meanings in the target text. This includes such aspects as the creation (by the source text author or the translator) of tenor and tone. This competence involves understanding the rules and conventions of rhetoric and cohesion in both codes well enough to know what meanings are conveyed; through either following or breaking said conventions in the source language, in addition to being able to render similar meaning in the target language, depending on the translation task. In fact, this competence is vital to successful translation both in the interpretative and the presentational modes of using language for communicative purposes (National Standards in Foreign Language Education Project (2006). That is to say, translators will make use of this mode to both read and interpret the source text as well as to produce the target text. Therefore, any authentic translation task will to some extent assess textual competence along with grammatical competence.

Similar to the linguistic sub-component, the textual sub-component of translation competence needs to be clearly stated. Table 2 illustrates a continuum of more to less successful translations by describing how those would reflect more or less mastery of this subcomponent.

3. *Pragmatic competence.* In Bachman's model of communicative competence (1990) adopted by Cao to discuss translation competence (1996) pragmatic competence is subdivided into illocutionary competence and sociolinguistic competence. Even if we could argue that sociolinguistic competence and pragmatic competence are separate, that Cao could have treated them separately, or could

Table 2. Textual sub-component (T = translation; TL = target language)

5	T is very well organized into sections and/or paragraphs in a manner consistent with similar TL texts. The T has a masterful style. It flows together flawlessly and forms a natural whole.
4	T is well organized into sections and/or paragraphs in a manner consistent with similar TL texts. The T has style. It flows together well and forms a coherent whole.
3	T is organized into sections and/or paragraphs in a manner generally consistent with similar TL texts. The T style may be inconsistent. There are occasional awkward or oddly placed elements.
2	T is somewhat awkwardly organized in terms of sections and/or paragraphs or organized in a manner inconsistent with similar TL texts. The T style is clumsy. It does not flow together and has frequent awkward or oddly placed elements.
1	T is disorganized and lacks divisions into coherent sections and/or paragraphs in a manner consistent with similar TL texts. T lacks style. T does not flow together. It is awkward. Sentences and ideas seem unrelated.

have used sociolinguistic competence as an umbrella term, I will follow her classification for the sake of simplicity. Illocutionary competence is the knowledge of how language is used to perform functions (e.g. apologizing, complaining). This competence plays an important role in the act of translation both when the translator approaches the source text as well as when he/she produces the target one. When a translator is approaching and analyzing a source text, this competence allows to discern whether the text is a polemic, a primarily objective report of data, a proposal for action, etc. Likewise, in the production of the target text, the translator makes use of this competence to reproduce those functions in the translation. Under the label of sociolinguistic competence, Bachman includes knowledge of linguistic variations (i.e. dialects, regionalism, and national varieties) and knowledge of cultural reference and figures of speech. Knowledge of variation is important in being able to interpret a source text in a dialect other than the standard. It also may be important in helping the translator understand the particular cultural assumptions that may underlie a source text. These features may also be communicated through cultural references and the use of figures of speech. Culture reference and figures of speech can be an obstacle to successful translation. A successful translator is aware of these elements and is able to resist the temptation to translate them directly and/or find a successful way of communicating their meaning in the target text. In addition to these two aspects of socio-cultural knowledge, it is important to add register (Angelelli 2006, 2004). The degree to which a text creates a social relationship between the author and his reader must be understood and properly rendered in many translation tasks. To provide an extreme example, it would be improper to translate a formal invitation extended to a foreign dignitary to attend a closing conference ceremony that needs to be RSVP into a colloquial and informal text that informs the reader about a gathering and does not convey the sense of formality and respect intended in the original. Such failure to recognize and communicate the use of register may lead to the intended readership's misunderstanding of the true communicative intent of a document.

Additionally, individuals who take part in a communicative event (and translation is one of them), to use Hymes' terms (1974) need to have knowledge about ways of "doing" and ways of "communicating". Hymes' work on discourse communities and communication, specifically applied to speaking (see Hymes taxonomy 1974) can also be applied to writing, written communication and therefore written translation. To belong to a community of discourse (for example American engineers writing for a German journal on engineering) means that the translator needs to be a competent writer in engineering by using specialized vocabulary and exhibiting mastery of underlying structures and assumptions that are relevant to that specific field of activity. In other words, when translators engage in technical translation for engineers, they need to be able to write as if they belonged to the discourse community of engineers. Even when they do not belong to that discourse community, they have to be perceived by the reader as if they were a part of it or as a "temporary guests" able to navigate it successfully (cf. Angelelli 2000 and the notion of interpreters as temporary guests in discourse communities). In the case of scientific and technical translation, translators need to know the content area of the field. They also need to know the typical renditions of technical terms and ways of discussing different topics in a given field in each of the languages involved. Failure to properly apply such knowledge may lead to an unsuccessful translation that is neither comprehensible nor useful to the target readership.

Translators working in specialized fields must also have enough field knowledge to be able to successfully render terminology and/or concepts that may be new to the field in either language. Similarly, business and legal translation depend on knowledge of governments, legal systems, business organizations and other aspects of these fields that underlie the texts in both cultures. They must be able to communicate these differences while making the target text comprehensible to the recipient. These tasks require a working knowledge of the degrees to which the texts and cultures being mediated converge and diverge. It also requires the ability to make the technical document in one language available to the layperson in another language when necessary. This is all modified by the translator's knowledge about the recipients' knowledge of the various fields involved and cultural differences that exist. Therefore, we need to assess the knowledge of discourse and register as well as the knowledge of the socio-cultural aspects of the language.

Table 3. Pragmatic sub-component (T = translation; TL = target language)

5	T shows a masterful ability to address the intended TL audience and achieve the translations intended purpose in the TL. Word choice is skillful and apt. Cultural references, discourse, and register are completely appropriate for the TL domain, text-type, and readership.
4	T shows a proficient ability in addressing the intended TL audience and achieving the translations intended purpose in the TL. Word choice is consistently good. Cultural references, discourse, and register are consistently appropriate for the TL domain, text-type, and readership.
3	T shows a good ability to address the intended TL audience and achieve the translations intended purpose in the TL. Cultural references, discourse, and register are mostly appropriate for the TL domain but some phrasing or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.
2	T shows a weak ability to address the intended TL audience and/or achieve the translations intended purpose in the TL. Cultural references, discourse, and register are at times inappropriate for the TL domain. Numerous phrasing and/or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.
1	T shows an inability to appropriately address the intended TL audience and/or achieve the translations intended purpose in the TL. Cultural references, discourse, and register are consistently inappropriate for the TL domain. Most phrasing and/or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.

Table 3 illustrates statements of a continuum of more to less successful translations by describing how those would reflect more or less mastery of this subcomponent.

4. *Strategic competence*. The final major aspect of translation competence that is included in this proposed construct is translation strategic competence. This competence has to do with the way in which a translator approaches a translation task and the methods he/she uses to pinpoint and overcome problems within the performance of the translation assignment. According to Orozco (2000), strategies include distinguishing between primary and secondary ideas, establishing conceptual relationships, searching for information, task planning and many others. This competence is where the conscious skill of the translator enters into the translation task. Here is where interference is controlled, where culture is consciously mediated, and where the decision is made to consult resources and determine how they can be properly applied.

Within a translator's strategic competence lies what PACTE (in Orozco 2000) calls instrumental-professional competence. Use of professional tools and standards of behavior is an important part of a translator's ability to be strategic. A translator's knowledge competence can be augmented but not substituted by the use of reference materials and consultation with professionals in the given field.

Proper knowledge of how to use and access these tools, therefore, is also a part of translator strategic competence. Colina (2003) points out that the proper use of translation tools is a factor that differentiates the novice translator from her/his professional counterpart. Inexperienced translators tend to make unskillful use of these tools since they do not possess the benefit of experience to tell them what to accept and what to reject. For the contemporary translator nowadays, these instrumental-professional competences go far beyond the use of dictionaries. The internet, electronic reference materials, CAT tools (computer applications to aid in the process of translation, memory applications, machine-assisted-translation software, as well as word processing programs) are all vital part of the toolkit for today's translator. One must not only know how but when to use each. Competent professional translators must be able to perform a successful web search and be able to identify which sources to accept and reject. They must be able to maintain and properly apply computer-aided translation tools such as translation memory and databases. These skills also include having a sufficient knowledge of the language to be able to successfully accept and reject a word processor's spelling and grammar corrections.

Similarly, the ability to manage human resources is an important part of a translator's strategic competence. A working translator in today's market must know how to obtain necessary information from a manager or client (Fraser 2000). Increasingly, a professional translator must also know how to work in a team (Arango-Keeth & Koby 2003). Due to testing formats and technology limitations, it may not be possible to assess these competences in every assessment task. However, it is important to see them as part of the construct of translation competence and acknowledge whether or not a certification test chooses or does not chose to assess them. Tests may or may not target all of the subcomponents of translation competence. As long as test developers clearly define what their test intends to measure, and justify the reasons for doing so, candidates know in advance what sub-components are included in a test and what sub-components are not.

To some extent, the strategic translation competence is the translator's ability to exercise conscious control over their linguistic, cultural, field, and instrumental knowledge. This competence is involved in choosing which features of the target text to put in the foreground and which to put in the background. It involves choosing between making something explicit (e.g. explaining a point that may be unfamiliar to the target audience) and using other devices to make a point implicit. This competence may also be reflected in the use of self-editing and drafting processes. It is included both explicitly and implicitly in any translation assessment even when the grader only sees the final product that the examinee submits. The application of strategy is only evident in its effect: strategic competence is truly demonstrated in the absence of problematic translations in the final product.

Table 4. Strategic sub-component (T = translation; TL = target language)

5	T demonstrates astute and creative solutions to translation problems. Skillful use of resource materials is evident.
4	T demonstrates consistent ability in identifying and overcoming translation problems. No major errors and very few minor errors are evident. No obvious errors in the use of resource materials are evident.
3	T demonstrates a general ability to identify and overcome translation problems. However, a major translation error and/or an accumulation of minor errors are evident and compromise the overall quality of the translation. Improper or flawed use of reference materials may be reflected in the TT.
2	T demonstrates some trouble in identifying and/or overcoming translation problems. Several major translation errors and/or a large number of minor errors are evident and compromise the overall quality of the translation. Improper or flawed use of reference materials is reflected in the TT.
1	T reflects an inability to identify and overcome common translation problems. Numerous major and minor translation errors lead to a seriously flawed translation. Reference materias and resources are consistently used improperly.

Table 4 illustrates statements of a continuum of more to less successful translations by describing how those would reflect more or less mastery of this subcomponent.

Now that we have operationalized these sub-components of translation ability, we can turn to discussing ways to assess them. Because they are definable, they are also gradable. In the field of testing, it is not uncommon to see subcomponents of a construct scored with a scoring rubric. Many professional associations granting certification do not use rubrics. In the next section I explore the use of a rubric for certifying translators.

5. Using a rubric

Rubrics are commonly used in testing. They allow for a more systematic and holistic grading. A rubric generally contains all sub-components that constitute the construct. It provides descriptive statements of behaviors that candidates may exhibit in a particular sub-component.

Since a scoring rubric can be used to holistically score virtually any product or performance (Moss and Holden 1988; Walvood and Anderson 1998; Wiggins 1998), it makes sense to discuss its feasibility for scoring translation. A rubric is developed by identifying what is being assessed (i.e. translation competence). It implies identifying the characteristics of translation competence, the primary traits of the product or performance, (i.e. micro-linguistic competence, textual competence, pragmatic competence, strategic competence, etc.) and then delineating criteria used to discriminate various levels of performance, as was done earlier with each of these sub-components. For example, in order to be considered a competent professional translator, an individual must demonstrate sufficient control and understanding of the linguistic features of the source language to successfully comprehend the meaning of a source text appropriate to the translation task. In addition, sufficient control and understanding of the linguistic features and writing conventions in the target language is necessary to successfully produce a target text appropriate to the translation task. This includes: grammatical and mechanical control, control of cohesive and textual devices, control of functional and socio-cultural aspects of the languages, and sufficient relevant world and technical knowledge to successfully complete the translation task. This includes both knowledge of cultural differences in world views and ways of doing things as well as area specific knowledge of institutions, ways of working, professional conventions, concepts and terminology. In addition, a competent professional translator exhibits an ability to identify and overcome problem areas in the performance of a translation task. This includes: application of strategies, use of professional tools and resources, and the ability to work in teams and work with manager and/or clients. Additionally, each of the sub-components carries a certain weight (decided by the association based on their needs and stated on test specifications).

By constructing a scoring rubric, graders can holistically score all the elements that were considered relevant to be included in a test. This assures that in the test, the construct that was intended to be measured is not only measured by the test (as a result of careful development) but it is also scored by graders. This is an important contrast to the point-adding or point-deducting system which is many times used in schools and professional associations. A description of the best work that meets these criteria and the best performance that can be expected will describe the top score. The worst work that can be expected using the same criteria constitutes the lowest acceptable score. Intermediate level work is assigned intermediary scores, and the number of intermediary levels determines the number of levels of a scale. For example, a rubric can have a scale that runs from one to six (e.g. unacceptable translation, inadequate translation, barely adequate translation, competent translation, very competent translation, and outstanding translation), or from one to three (unacceptable, barely acceptable, clearly acceptable) or any other set that is meaningful for the organization that is developing the test.

Table 5 is an example of a five-point-scale rubric that could be used to assess translation ability by professional associations. It was drafted for the American Translators Association as a result of a self-study on their certification exam.

Table 5. Working draft for rubric to assess translation (Angelelli 2006)
T = translation; TL = target language; ST = source text

	Source Text Meaning
5	T contains elements that reflect a detailed and nuanced understanding of the major and minor themes of the ST and the manner in which they are presented in the ST. The meaning of the ST is masterfully communicated in the T.
4	T contains elements that reflect a complete understanding of the major and minor themes of the ST and the manner in which they are presented in the ST. The meaning of the ST is proficiently communicated in the T.
3	T contains elements that reflect a general understanding of the major and most minor themes of the ST and the manner in which they are presented in the ST. There may be evidence of occasional errors in interpretation but the overall meaning of the ST appropriately communicated in the T.
2	T contains elements that reflect a flawed understanding of major and/or several minor themes of the ST and/or the manner in which they are presented in the ST. There is evidence of errors in interpretation that lead to the meaning of the ST not being fully communicated in the T.
1	T shows consistent and major misunderstandings of the ST meaning.
	Style and Cohesion (addresses textual sub-component)
5	T is very well organized into sections and/or paragraphs in a manner consistent with similar TL texts. The T has a masterful style. It flows together flawlessly and forms a natural whole.
4	T is well organized into sections and/or paragraphs in a manner consistent with similar TL texts. The T has style. It flows together well and forms a coherent whole.
3	T is organized into sections and/or paragraphs in a manner generally consistent with similar TL texts. The T style may be inconsistent. There are occasional awkward or oddly placed elements.
2	T is somewhat awkwardly organized in terms of sections and/or paragraphs or organized in a manner inconsistent with similar TL texts. The T style is clumsy. It does not flow together and has frequent awkward or oddly placed elements.
1	T is disorganized and lacks divisions into coherent sections and/or paragraphs in a manner consistent with similar TL texts. T lacks style. T does not flow together. It is awkward. Sentences and ideas seem unrelated.
	Situational Appropriateness (addresses pragmatic sub-component)
5	T shows a masterful ability to address the intended TL audience and achieve the translations intended purpose in the TL. Word choice is skillful and apt. Cultural references, discourse, and register are completely appropriate for the TL domain, text-type, and readership.
4	T shows a proficient ability in addressing the intended TL audience and achieving the translations intended purpose in the TL. Word choice is consistently good. Cultural references, discourse, and register are consistently appropriate for the TL domain, text-type, and readership.

Table 5. (*continued*)

3	T shows a good ability to address the intended TL audience and achieve the translations intended purpose in the TL. Cultural references, discourse, and register are mostly appropriate for the TL domain but some phrasing or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.
2	T shows a weak ability to address the intended TL audience and/or achieve the translations intended purpose in the TL. Cultural references, discourse, and register are at times inappropriate for the TL domain. Numerous phrasing and/or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.
1	T shows an inability to appropriately address the intended TL audience and/or achieve the translations intended purpose in the TL. Cultural references, discourse, and register are consistently inappropriate for the TL domain. Most phrasing and/or word choices are either too formal or too colloquial for the TL domain, text-type, and readership.
	Grammar and Mechanics (addresses micro-linguistic sub-component)
5	T shows a masterful control of TL grammar, spelling, and punctuation. Very few or no errors.
4	T shows a proficient control of TL grammar, spelling, and punctuation. Occasional minor errors.
3	T shows a weak control of TL grammar, spelling, and punctuation. T has frequent minor errors.
2	T shows some lack of control of TL grammar, spelling and punctuation. T is compromised by numerous errors.
1	T shows lack of control of TL grammar, spelling and punctuation. Serious and frequent errors exist.
	Translation Skill (addresses strategic sub-component)
5	T demonstrates able and creative solutions to translation problems. Skillful use of resource materials is evident.
4	T demonstrates consistent ability in identifying and overcoming translation problems. No major errors and very few minor errors are evident. No obvious errors in the use of resource materials are evident.
3	T demonstrates a general ability to identify and overcome translation problems. However, a major translation error and/or an accumulation of minor errors are evident and compromise the overall quality of the translation. Improper or flawed use of reference materials may be reflected in the TT.
2	T demonstrates some trouble in identifying and/or overcoming translation problems. Several major translation errors and/or a large number of minor errors are evident and compromise the overall quality of the translation. Improper or flawed use of reference materials is reflected in the TT.
1	T reflects an inability to identify and overcome common translation problems. Numerous major and minor translation errors lead to a seriously flawed translation. Reference materials and resources are consistently used improperly.

In order to obtain feedback it was presented to ATA graders during the 48th ATA Annual Conference in November 2007. At the time of writing this article, the ATA certification committee had not made a decision to expand or change the sub-components of translation competence listed in their website, nor had they decided on the consideration of this rubric, either partially or in its entirety.

The operational categories selected in the creation of this rubric may not at first glance appear to be equal to the sub-components of the construct defined above. They are however inter-related. Some sub-categories have been collapsed and unified into a single category in order to minimize the number of categories that graders must rate and to facilitate the definition of performance levels (Bachman & Palmer 1996; Cohen 1994). Additionally the terms used in the rubric are more aligned with terminology generally used by graders. The definition of the rubric categories and their justifications are as follows:

Source text meaning is a measure of the extent to which the candidate's response (the target text) reflects or fails to reflect an adequate understanding of the themes and rhetoric of the source text. Appropriate conveyance of meaning is always present in the discourse of professional organizations when they define what the exams are targeting. This is different from language production, although many times the borders between the two areas get blurred. However, meaning is a very indirect measure of the grammatical competence of the candidate in the source language. It is difficult to call this a reliable measure of source language grammatical competence since the difficulty with target language production may also hinder successful communication of major and subtle meanings encoded in the language of the source text. If an organization wanted to measure language comprehension, which may impact the rendering of meaning, a more direct measure of source language comprehension that is not dependent on target language production would be preferable. Nevertheless, a successful communication of the meanings found in the source text may correlate highly with source text comprehension, although this would need to be demonstrated through empirical research. Obvious misinterpretations of the source text as evidenced in the target text may also be seen as possible indicators of problems with source text comprehension.

Target text style and cohesion is seen as being reflective of the candidate's knowledge of the ways in which texts are linked and organized into documents in the given target language genre, or document type within a given communicative setting. Knowledge of genre is seen as an ability to segment the document appropriately into sections and/or paragraphs, such as an introduction, statement of the problem, presentation of findings, discussion, proposals, etc. depending on the purpose and type of document being translated. Knowledge of cohesive devices and the rules for creating a coherent text is reflected in the flow of the document, the degree to which it seems to form a single unit and how it addresses textual competence.

Situational appropriateness is a measure of the candidate's ability to employ the functional and socio-cultural aspects of the target language in their translation. The functional aspects of language are judged by the degree to which the target text is successful at achieving its intended target language use. The socio-cultural aspects of language are judged in the target text's use of appropriate markers of register, i.e. degree of formality in phrasing and word choice. It addresses pragmatic competence.

Grammar and mechanics is the category which includes spelling, diacritical marks, agreement, punctuation, and other conventions of the writing and grammar of the target language. It addresses linguistic competence in the target language. Together with meaning it is the category most frequently used by professional associations and schools while scoring a translation test.

Translation skill is meant to include the application of strategies to translation problems and the use of resource materials. This category is measured by how well the target text reflects the identification of translation problems and successful solutions to them. It also includes the degree of success or failure in the appropriate use of references in overcoming gaps in language or topic knowledge. (This may appear more clearly in the misapplication of resources.) It addresses strategic competence.

6. Levels of performance of a rubric

The rubric presented above in Section 5 was designed with a high professional standard for certification as its set point. The scale goes from 1 to 5. Number 1 is seen as a true lack of ability. It is imagined that very few candidates will score "1" at this level, given that they should have self-selected out of the examination. A score of "5" is seen as being indicative of particularly outstanding performance. The desired level of performance for certification is seen as being represented by "4" on this scale. Number 3 is seen as the point at which the candidate shows evidence of skill but falls slightly short of the proficiency level desired for certification. A number "2" on the scale represents a deficient performance that is clearly below the level required to perform as a certified professional translator. It is important to point out that organizations which currently grant certification (e.g. ATA, NAATI) focus on certifying candidates in terms of language pairs and directionality. That information is provided by stating "certified in Spanish into English," for example. Other than that, certification does not give any specific information about what types of texts a candidate can translate and/or in which contexts. Even when some organizations may discriminate between certifying at a professional or paraprofessional level (e.g. NAATI), certification is generic. Therefore, the levels of the rubric simply point to distinct levels of performance that programs may need to show so that test results can be referenced to certification criteria in the event of an examination challenge. It is also believed that the use of this number of performance levels will be easily managed by graders. There is the possibility that more levels would be confusing for candidates and graders while the use of fewer levels would not allow for a clear delineation of competences upon which decisions about certification are made.

Conclusion

This paper explored what translation ability is, and how it may be measured in a meaningful way for the purpose of certifying translators. Certification examinations assess an individual's ability to function as a professional translator. This ability (test construct) can be defined as consisting of the following sub-components: linguistic competence, textual competence, pragmatic competence, and strategic competence. Because important questions are asked before conceptualizing a test, test designers are able to define the test construct based on specifications of professional organizations. As test designers engage in test development and consider what associations deem as important elements to be present in those tests, they also consider how to score them. One possible way to holistically score a test is by using a scoring rubric. Test designers develop rubrics based on the test construct sub-components. Once translation skills have been defined, it is agreed that knowledge of these skills and the ability to successfully apply them form the core of translation ability (test construct). A translation test, however, may not measure them all. It is therefore important to define *a priori* which parts (sub-components) of the construct are examined by the organization test, and then check for validity during and after the development of the test. All tests have consequences. In the case of certification tests or any other high stake test, the consequences for test takers are even more important. This is why extreme care should be taken to develop tests that measure what they are set out to measure (i.e. the construct, in this case translation competence) and that those tests measure translation competence in a valid and reliable way. Clear definitions of constructs as well as validity and reliability considerations constitute the basis from which we need to develop current and future translation assessment examinations.

Limitations and implications

The research presented here should be considered as work in progress. It is an attempt to put forward a way of measuring translation competence in a more systematic way. Until this rubric is put to the test and applied to exams of various language combinations, we cannot begin to discuss its value (or lack thereof) for professional organizations granting certification. In order to do so, we need to see more collaboration between professional organizations and researchers. As has been stated above, the little discussion on translation assessment has been done in the field of translation (and interpreting) and it is still obscured by the tension between theory and practice. Practitioners believe that expertise in testing is obtained by practical experience. Since they may not be aware of the consequences of not developing a test based on testing principles, or of not examining a test for its validity, reliability, authenticity or practicality, they continue testing candidates according to a status quo. In so doing, they are measuring other candidates' performances to the best of their abilities. This, however, may not be enough of a justification.

In sum, as this paper demonstrates, the literature on translation competence and measurement of translation suggests the current need of exploration by mainstream researchers of testing and assessment (i.e. researchers who have expertise in testing, even if they do not work in the field of Translation Studies), as they address the many complex issues surrounding the measurement of translation competence in various language pairs. For applied linguists working with testing and bilingualism, and for translation scholars working on testing, the literature on translation competence and translation measurement offers important insights about the linguistic, psycholinguistic and sociolinguistic characteristics of individual bilingualism, as well as about the testing practices used in the profession. This article concludes with a call for collaboration between practitioners and professional organizations, as well as researchers in translation and researchers in testing. This work has implications for translation teaching and testing, for translation teacher education and translation practice.

References

Adab, Beverly. 2000. "Evaluating Translation Competence". *Developing Translation Competence*, 215–228.

Angelelli, Claudia V. 2000. "Interpreting as a Communicative Event: A Look through Hymes' Lenses." *Meta* 45 (4): 580–592.

Angelelli, Claudia V. 2003. "Connecting World Collaborative Testing Project". Technical Report. Commissioned by Healthy House for a MATCH Coalition and The California Health Partnership and funded by The California Endowment.

Angelelli, Claudia V. 2004a. *Medical Interpreting and Cross-Cultural Communication.* London: Cambridge University Press.

Angelelli, Claudia V. 2004b. *Revisiting the Interpreter's Role: A Study of Conference, Court and Medical Interpreters in Canada, Mexico, and in the United States*. Amsterdam/Philadelphia: John Benjamins.

Angelelli, Claudia V. 2006. "Minding the Gaps: New Directions in Interpreting Studies". *TIS Translation and Interpreting Studies* 1 (1): 41–67.

Angelelli, Claudia V. 2007a. "Validating Professional Standards and Codes: Challenges and Opportunities." *INTERPRETING: International Journal of Research and Practice in Interpreting* 8 (2): 175–193.

Angelelli, Claudia V. 2007b. "Accommodating the Need for Medical Interpreters: The California Endowment Interpreter Testing Project." *The Translator* 13 (1): 63–82.

Arango-Keith, Fanny, and Koby, Geoffrey S. 2003. "Translator Training Evaluation and the Needs of Industry Quality Training." In *Beyond the Ivory Tower: Rethinking Translation Pedagogy*. Brian Bear & Geoffery S. Kobby (eds.) Amsterdam: John Benjamins Publishing Company, 117–135.

ATA Certification Program. 2008. (Active September 29, 2008).
(http://www.atanet.org/certification/aboutcert_overview.php)

Bachman, Lyle and Palmer, Adrian. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Beeby, Allison. 2000. "Evaluating the Development of Translation Competence." In *Developing Translation Competence*. Christina Schäffner & Beverly Adab (eds.) Amsterdam: John Benjamins Publishing Company, 185–198.

Cao, Deborah. 1996. "On Translation Language Competence." *Babel* 42 (4): 231–238.

Cohen, Andrew. 1994. Assessing Language Ability in the Classroom. 2nd edition. Boston: Heinle & Heinle Publishers.

Colina, Sonia. 2002. "Second Language Acquisition, Language Teaching and Translation Studies". *The Translator* 8 (1): 1–24.

Colina, Sonia. 2003. *Translation Teaching From Research to the Classroom a Handbook for Teachers*. Boston: McGraw-Hill Companies.

Colina, Sonia. 2008. "Translation Quality Evaluation: Empirical Evidence from a Functionalist Approach." *The Translator* 14 (1): 97–134.

Faber, Pamela. 1998. "Translation Competence and Language Awareness." *Language Awareness* 7 (1): 9–21.

Fulcher, Glenn. 2003. *Testing Second Language Speakers*. London: Pierson Longman Press.

Johnson, Marysia. 2001. *The Art of Non-Conversation.* New Haven: Yale University Press.

Kiraly, Don C. 1995. *Pathways to Translation: Pedagogy and Process*. Trans. Geoff Koby. Kent: Kent State University Press.

McNamara, Tim, and Roever, Carsten. 2006. *Language Testing: The Social Dimension*. Blackwell Language and Learning Monograph Series. Richard Young, ed. Malden, MA: Blackwell Publishing.

Messick, Samuel. 1989. "Validity". In R. Linn (ed.) *Educational Measurement.* New York: Macmillan, 13–103.

Moss, Andrew, and Holder, Carol. 1988. *Improving Student Writing.* Dubuque, IO: Kendall/Hunt.

NAATI. 2007. *Levels of Accreditation.* http://www.naati.com.au/at-accreditation-levels.html (Active September 20 2008).

National Association of Judiciary Interpreters and Translators. Webpage on National Judiciary Interpreter and Translator Certification (NJITCE). htttp://www.najit.org/ Certification/NAJIT_Cert.html (Active November 20 2008).

National Standards in Foreign Language Education Project. 2006. *Standards for Foreign Language Learning in the 21st Century*. Lawrence, KS: Allen Press, Inc.

Neubert, Albert. 2000. "Competence in Language, in Languages and in Translation." In *Developing Translation Competence*, 3–18.

Nord, Christiane. 1991. Scopos, "Loyalty and Translational Conventions." *Target* 3 (1): 91–109.

Nord, Christiane. 1997. *Translation as a Purposeful Activity: Functionalist Approaches Explained.* Manchester: St. Jerome.

Orozco, Mariana. 2000. "Building a Measuring Instrument for the Acquisition of Translation Competence in Trainee Translators." In *Developing Translation Competence*, 199–214.

Presas, Marisa. 2000. "Bilingual Competence and Translation Competence." In *Developing Translation Competence*, 19–31.

Sawyer, David B. 2004. *Fundamental Aspects of Interpreter Education.* Amsterdam: John Benjamins Publishing Company.

Schäffner, Christina, and Adab, Beverly. (Eds.). 2000. *Developing Translation Competence*. Amsterdam: John Benjamins.

Valdés, Guadalupe, and Angelelli, Claudia V. 2003. "Interpreters, Interpreting and the Study of Bilingualism". *The Annual Review of Applied Linguistics* 23: 58–78.

Walvood, Barbara E. and Anderson, Virginia J. 1998. *Effective Grading.* San Francisco: Jossey-Bass.

Weir, Cyril J. 2005. *Language Testing and Validation. An Evidence-Based Approach*. New York: Palgrave Macmilan.

White, Edward M. 1994. *Teaching and Assessing Writing.* San Francisco: Jossey-Bass.

Wiggins, Grant. 1998. *Educative Assessment: Designing Assessments to Inform and Improve Student Performance.* San Francisco, CA: Jossey-Bass, 184–185.

Moving beyond words in assessing mediated interaction

Measuring interactional competence in healthcare settings

Holly E. Jacobson
University of Texas at El Paso

This chapter focuses on assessment of community interpreter performance in U.S. healthcare settings where nearly 50 million U.S. residents speak a language different from their primary healthcare provider. It briefly discusses the way assessment of mediated interaction in patient-healthcare professional interaction falls short in its exclusive emphasis on words and phrases and "verbatim" renditions, at the exclusion of other linguistic and interactional features. Grounded in concepts derived from interactional sociolinguistics and conversation analysis, it points to the need to develop a more comprehensive approach to assessing interpreter performance, emphasizing discourse management and the use and transfer of contextualization cues. A step-by-step approach for developing an analytic scoring rubric for assessing interactional competence is presented.

Introduction

The assessment of interpreting performance is an area of research that is still in its infancy. It is an essential area of study for a number of reasons: understanding how "quality" is determined in interpreting can provide the linguist with rich information about language structure, language processes, and, most relevant to this chapter, language use. Delving into the testing and assessment of an individual's capacity to perform in interpreting is therefore a worthwhile endeavor for generating scientific knowledge for the sake of knowledge itself. In addition, interpreter performance assessment has obvious practical applications, as can be seen in the chapters of this particular volume; it is of great significance to applied linguists who specialize in interpreting pedagogy and credentialing professional interpreters. Globalization has led to the growth of language-minority populations in countries throughout the world during the past several decades, and to an ever-increasing need to respond to ethical obligations related to equal access to medical, educational, and other social services for language-minority populations. From this, concerted efforts have emerged to establish best practices in the education, preparation, and assessment of interpreter professionals.

Throughout the world, the use of community interpreters has grown exponentially in response to increases in immigrant populations (Wadensjö 1998). Community interpreting includes the mediation of face-to-face interaction in such settings as forensic interviews, immigration interviews, business meetings, attorney-client consultations, and medical interviews, and generally involves a bi-directional, consecutive mode of interpreting. Wadensjö's (1998) model of interpreting is grounded in the dialogic theory of Bakhtin, and she uses the term "dialogue interpreting" to refer to community interpreting as a means of stressing "the defining primacy of the setting" (1998: 50). Another common term found in the literature is "liaison interpreting" (Pöchhacker 2004). For the purposes of this chapter, "community interpreting" is used as a means of emphasizing the discourse communities represented by the interlocutors present in mediated interaction, given that discourse is at the center of the discussion presented here.

This chapter focuses on assessment of community interpreter performance in U.S. healthcare settings (although the discussion is relevant to other settings, as well), where nearly 50 million U.S. residents speak a language different from that of their primary healthcare provider (Moreno, Otero-Sabogal & Newman 2007). It briefly discusses how assessment of interpreting in patient-healthcare provider interaction falls short in its exclusive emphasis on words and phrases and "verbatim" renditions, at the exclusion of other linguistic and interactional features, such as turn taking signals and other paralinguistic features of language. Grounded in concepts derived from interactional sociolinguistics and conversation analysis, it points to the need to develop a more comprehensive approach to assessing interpreter performance, emphasizing discourse management and the use and transfer of contextualization cues. While there are many other features beyond those discussed here that call for additional research, the specific goal of this chapter is to demonstrate the importance of moving beyond the lexico-semantic level in interpreter performance and to propose the use of assessment tools that include interactional features.

Interpreter performance-based assessment in healthcare

The influx of language-minority patients into the U.S. health system has led to a flurry of interest in the healthcare interpreter by medical researchers, with a number of studies in recent years being published in medical and health journals to compare the performance of ad hoc interpreters to "professional" interpreters (cf. Garcia et al. 2004; Moreno, Otero-Sabogal & Newman 2007; Flores et al. 2003; Taveras & Flores 2004); to analyze misunderstanding and breakdown in communication caused by interpreters (cf. Berstein et al. 2002; Fagan et al. 2003; Elderkin-Thompson, Lee et al. 2002; Silver & Waitzkin 2001; Hampers & McNulty 2002; Jacobs et al. 2001; Woloshin et al. 1995; Woloshin, Schwartz, Katz & Welch 1997); and to develop approaches to testing the language proficiency and interpreting skills of healthcare interpreters (cf. Moreno, Otero-Sabogal & Newman 2007). A great majority of the research in interpreting that has been published in medical and health journals has been done in isolation in the sense that it has benefited little from cross-pollination between disciplines. For example, many of the theories and models of human interaction developed within sociology and linguistics could provide the basis for a better understanding of interpreted interaction in healthcare settings, and inform how interpreter performance is assessed. This will be discussed in more detail in the next section.

A look at a few of the articles conducted by medical and health professionals mentioned above shows that the focus of analysis is lexico-semantics, or "accuracy" in interpreting; that is, the unit of analysis is the word or phrase, and whether or not the interpreter produced a "verbatim" rendition. Consider, for example, the work of Flores et al. (2003: 7), who conducted an error analysis of patient-physician encounters that took place in a pediatric clinic. The researchers audio taped the visits, and identified "frequency and categories of interpreter errors." Five categories of errors were established, including (1) omission of a word or phrase; (2) addition of a word or phrase; (3) substitution of a word or phrase; (4) elaborating one's own personal views instead of interpreting a word or phrase; and (5) use of an incorrect word or phase (one that does not exist in the target language). Deviations from word-for-word renderings were only acceptable in the case of "medical jargon, idiomatic expressions, and contextual clarifications," and when the interpreter was expected to "act as a cultural broker or advocate."

Another recent study realized by Moreno, Otero-Sabogal and Newman (2007: 331–332) involved the development of a test for interpreters in a healthcare system in northern California to "assess dual-role staff interpreter linguistic competence… to determine skill qualification to work as medical interpreters." In addition to equating language competence to interpreting competence, the unit of analysis was, as in the study by Flores et al., the word or phrase. A written translation portion of the test was designed to assess "completeness" and medical terminology. This portion of the exam was graded for accuracy, which was defined as using the correct terminology, regardless of grammar, spelling, and punctuation. Although the researchers claim that the oral portion of the test was developed to assess "comprehension and effective communication," it was administered over the phone, rather than face-to-face, and included 14 questions that were designed to elicit "verbatim interpretation." Full credit was given for a word or phrase that was interpreted "verbatim." Only partial credit was given for a rendering that was provided as an "explanation" rather than verbatim.

The research approaches implemented in these studies appear to derive from a view of interpreting which Clifford (2001: 366) contends is typical of the general population; that is, interpreting is assumed to be a straightforward exercise of substituting one lexical item for another. Although there is sufficient evidence to demonstrate that interpreting is much more complex, as will be discussed in the following section, the one-dimensional, word-level approaches to assessing interpreter performance in healthcare settings represented in the two studies appear to constitute the norm in the medical literature (see references provided above). A systematic review of the literature is needed to determine whether this is so.

The recent attention paid to interpreted interaction in healthcare is laudable, and demonstrates initial steps towards a commitment to better understanding the dynamics of patient-provider communication, and to improving healthcare access. However, lexico-semantic analyses, which are conducted at the expense of all other aspects of language, and in the complete absence of variables at the interactional level, leave one to doubt whether the research has demonstrated anything of substance regarding mediated interaction. These shortcomings in assessing healthcare interpreting must be addressed, and point to the need for future research that is informed by other areas of investigation, including interactional sociolinguistics and conversation analysis. In other words, cross-pollination among disciplines is essential if equal healthcare access for language-minority populations is to be achieved. Effectively defining and assessing quality performance in interpreting will not be possible if medical and health researchers continue to work in isolation.

The following section considers why the simplified, word-for-word view of interpreting appears to persist in the medical and health research, and provides an alternative view of interpreting, as supported by the empirical research of scholars in linguistics, sociology, and interpreting studies.

The subsequent section then focuses on some of the interactional elements of interpreting performance that have previously been neglected, and demonstrates how discourse theory can contribute to a more comprehensive performance-based assessment through the development and implementation of a rubric. It is important to note that healthcare interpreters do indeed need to develop their knowledge and terminology in the many areas of healthcare in which they work. A comprehensive assessment instrument for measuring the construct of interpreter competence must clearly include "accuracy" as one of the multiple traits to be evaluated, considering such sub-competencies as inclusion of significant content; correct use of medical and other terminology; language accommodation (including accommodation to language variety and register); unambiguous use of terminology; avoiding the use of false cognates; avoiding literal renderings resulting in gibberish; among others (Jacobson 2007). Another competency that might be considered essential to an instrument designed for the formative or summative evaluation of interpreters and student-interpreters is "professionalism": this could include the sub-competencies of accountability, appropriate dress, appropriate demeanor, demonstration of diligence in developing interpreting skills, among many others (Jacobson 2007). Professionalism is also often defined as adhering to a particular code of ethics (cf. American Institutes for Research 2005; Roat, Gheisar, Putsch & SenGupta 1999). Such competencies as accuracy and professionalism, among many others not mentioned in this chapter, are certainly integral to interpreter performance assessment. However, they are beyond the scope of this chapter, which focuses strictly on some of the many features of interactional competence in interpreting. The exclusion of other areas of interpreting, such as lexico-semantic or professional variables, provided here simply as examples, is not intended to signify that they lack importance.

Interpreting as discourse

The conduit model

The work of Clifford provides an example of interpreting research informed by the theoretical frameworks of other fields. He has conducted research in interpreting (including healthcare interpreting, Clifford 2007b), applying frameworks of testing and assessment (Clifford 2001, 2005a, 2005b, 2007b), pragmatics (2001), and discourse theory (2001, 2005a). Clifford's (2004: 92) discussion of the conduit model provides insight into the persistence of the word-for-word model of interpreting in the medical and health literature. He points out that, "The conduit portrays interpreting as an exercise carried out on linguistic forms, one in which even the smallest changes in perspective…are not permitted." In his exploration of the origins of the model, he suggests that it is based more on perceived morality or ethics (e.g. the need to be faithful to the original rendition) than on empirical evidence of what constitutes effective communication. Clifford argues that the conduit model evolved from traditions in conference interpreting, in which the interpreter has little face-to-face interaction. He also discusses the contribution that sign language interpreting has made to the persistence of the model. According to Clifford, given that the model attributes machine-like characteristics to the interpreter, who is considered to be impartial, invisible, and simply a channel through which messages are conveyed, it has served to provide Deaf individuals with a greater sense of equality with interpreters, leading to a greater sense of autonomy.

In addition, the model may continue to persist as the dominant one in community interpreting in healthcare settings given its simplicity. This contention is based on the author's experience in working in healthcare contexts as researcher and consultant, and in developing curricula for healthcare interpreters (Jacobson 2007). More complex models of language and interpreting may be met with resistance in the healthcare arena due to the potential changes in infrastructure such models could entail. However, anecdotal evidence is clearly not enough, and empirical research is needed to determine why the conduit model persists in the medical and health research and in interpreter education programs despite empirical evidence demonstrating its weaknesses. Whatever the case, the conduit model provides a reductionist model of language and interpreting, which seems to facilitate the development of training curricula (cf. Roat et al. 1999) and testing and assessment approaches (cf. medical and health literature cited above) that can be implemented with ease, and with limited expertise in language and communication. The medical and health literature ignores norms of interaction (see next section) which preclude the possibility of using simple frequency counts of words and phrases to determine the impact of an interpreter's performance on interactional goals and health outcomes. Language proficiency and interpreter competence cannot be measured only at the level of lexicon. More complex models of mediated interaction are called for, even if they meet with resistance caused by the need to revamp current approaches to education and assessment, which in turn could be viewed as costly and time-consuming. The conduit model is particularly problematic in contexts such as healthcare, where language has been identified by immigrants as the number one barrier to healthcare access in the United States, above other barriers, such as poverty and lack of insurance (The Access Project 2002; Flores, Abreu, Olivar & Kastner 1998). Equal access to healthcare for language-minority populations cannot be obtained without the development of effective interpreter assessment approaches and tools informed by what is known about human interaction. An alternative model of interpreting has been called for in the literature by researchers such as Angelelli (2001, 2003), Clifford (2001, 2004, 2005a, 2007b), Davidson (1998, 2000), Metzger (1999), Roy (2000), and Wadensjö (1998). The research of these scholars indicates that the traditional conduit model is deficient in that it focuses on lexicon, which is only one of the many complex features of language in face-to-face interaction. It is essential that mediated interaction be analyzed as discourse.

Interactional sociolinguistics

Interactional sociolinguistics (IS) is one of the many areas of research in discourse analysis that is useful for demonstrating the complexity of interpreted interaction. Interactional sociolinguistics (IS) is an approach to the study of discourse that has evolved from the work of Goffman (1981) and Gumperz (1977, 1982, 1984), and further developed by Tannen (1984, 1989). The theory and methodology of IS are based in linguistics, sociology, and culture (Schiffrin 1996), and are concerned with the processes of everyday, face-to-face interaction. IS attempts to answer questions related to the devices used to signal communicative intent, and the strategies used in conversational management (Gumperz 1984). Within the framework of IS, communicative competence goes beyond knowledge of vocabulary and grammar to include interactional competence, which is defined by Gumperz (1982: 209) as "the knowledge of linguistic and related communicative conventions that speakers must have to create and sustain conversational cooperation, and thus involves both grammar and contextualization." Meaning is situated, and hearers infer speakers' meaning based on their knowledge of the context, contextualization cues (which exhibit cross-linguistic variation, and include prosodic and paralinguistic features, such as gestures, facial expressions, and pauses), expectations about the thematic progression of the interaction, and by drawing on cultural presuppositions (Schiffrin 1996). If successful communication is to take place, interlocutors must share the same interpretive frame (definition of the situation), repertoire of contextualization cues, and sociocultural knowledge and expectations. These constitute variables that drive the norms of interaction, or rules of speaking, and they vary across languages.

IS contributes to the understanding of intercultural communication in particular from the standpoint of miscommunication; it focuses on unsuccessful face-to-face encounters to determine where a breakdown in communication has occurred. Research in IS has demonstrated that miscommunication can occur when the rules of speaking of one language are transferred to another language (Saville-Troike 1989); that is, miscommunication can be triggered when sociocultural knowledge and language-bound discourse strategies and signaling devices are used to ill effect in another language.

To illustrate the importance of IS to mediated interaction, consider the use of contextualization cues. As mentioned above, contextualization cues include certain prosodic and paralinguistic features that emphasize or add particular shades of meaning to what people say. They are powerful language tools, and can even be used to signal attitude, stance, and power relationships, among other meanings. Examples of such cues include intonation (e.g. pitch, rhythm, and intonation contour, etc., cf. Gumperz 1982); body positioning (e.g. standing close or far away from a speaker; leaning forward or against an inanimate object, etc., cf. Schegloff 1998); head and eye movements (nodding, eye gaze, etc., cf. Heath 1992); silence (pauses, pause length, etc., cf. Tannen & Saville-Troike 1985); to mention only a few. Such features signal different meanings in different languages. They are acquired through prolonged contact with the language. Serious miscommunications can develop in contexts where such cues are either not accessible or not understood. As Schiffrin (1994: 101) points out: "…such misunderstandings can have devastating social consequences for members of minority groups who are denied access to valued resources, based partially (but not totally) on the inability of those in control of crucial gate keeping transactions to accurately use others' contextualization cues as a basis from which to infer their intended meanings." Davidson (1998, 2000) presents data collected in the context of a U.S. healthcare system suggesting that interpreters play the role of powerful gatekeepers in healthcare settings, often effectively blocking access to healthcare information to language-minority patients. This is not always done intentionally, and can occur due to a lack of interactional competence in the interpreter's second language (L2), including the use and transfer of contextualization cues. Miscommunication due to lack of L2 competence in prosodic features, including intonation contour, may lead to monotonous renderings, for example, which could, in some languages, signal boredom, lack of interest in one's job, and a lack of respect for the interlocutors. The inadvertent signaling of a "negative attitude" could be distracting to healthcare professionals and patients alike, regardless of use of "correct terminology." It may produce emotional reactions in the patient, who may feel slighted or even oppressed, or in the healthcare professional, who may feel her advice is not being relayed with the necessary urgency (Jacobson 2007). Lack of eye contact and inappropriate gestures are examples of other paralinguistic features that can skew meaning and intention, and ultimately impact the interactional goals of a particular communicative event, such as a medical interview.

Cross-linguistic competence in the use and transfer of contextualization cues cannot be ignored in interpreter performance assessment, nor can the impact of miscommunications resulting from lack of competence in this area be neglected in empirical research on health outcomes. However, the ability of the interpreter to manage the flow of discourse is also essential to effective mediation and positive health outcomes.

Conversation analysis

Conversation analysis (CA) represents an area of discourse analysis that sheds light on discourse management in mediated interaction. CA is derived from a research approach referred to as ethnomethodology, which is concerned with how conversation or discourse is organized (Schiffrin 1994). It considers the way conversations are structured by delving specifically into turn taking, in particular, how turns are used to indicate what came before and what will come next in a conversation (Roy 2000; Schiffrin 1994). Researchers have demonstrated that turn taking is governed by specific norms (Schiffrin 1994; Clifford 2005), and that these norms vary cross-linguistically (Tannen 1984; Valli & Lucas 1992).

Community interpreters are faced with the daunting task of managing the flow of conversation: it is from this perspective of mediated interaction that the precept mandating that interpreters be "invisible" can clearly be viewed as untenable. Interpreting, far from occurring in a neutral, noninvolved manner, requires an active, direct interlocutor who is constantly shifting roles, aligning herself with primary interlocutors, and managing the flow of conversation. The interpreter creates and takes turns, manages overlap, and initiates talk (Metzger 1999; Roy 2000). As the only interlocutor who is assumed to have a command over both languages, the interpreter is expected to take on the primary responsibilities of discourse management. Competence in the turn taking norms of both languages is essential to this role. Turn taking competence involves knowing, in both languages, whether or not interlocutors are allowed to overlap in talk; the appropriate length of time for a turn (appropriate amount of talk); the significance of a pause (for example, whether a pause indicates a relinquishing of the floor to the other interlocutor, or simply a break before continuing with the same turn); and being able to extract the appropriate meaning from other paralinguistic features, or surface signals, used to manage turn taking (e.g. eye gaze, gestures, and body positioning). In addition, interpreters must be skilled in managing their own turns: it is often necessary to avoid being cut off; to back track in order to interpret a turn that was previously misunderstood or cut off; or to take an additional turn to make or request a clarification.

To illustrate the importance of discourse management in interpreted interaction, consider the possible impact of violating turn taking rules or norms. There are at least three conditions that can lead to the mismanagement of discourse by the interpreter, and thus to miscommunication. These conditions were derived from data collected by Jacobson (2007), who conducted qualitative research with student interpreters in medical clinics in a large metropolitan area in the U.S. One condition is related to memory problems, or the inability to retain longer chunks of information at a time in memory, rather than short words and phrases. The second condition pertains to lack of familiarity with lexicon and content (thus supporting the much acclaimed need for sufficient study in terminology and topics related to healthcare). The third condition is related to lack of turn-taking competence in the interpreter's L2. When any or all of these conditions exist, the interpreter is likely to short-circuit interlocutors' turns, such as during the question-answer sequence in a medical interview. This might happen when the interpreter lets out a gasp, or produces a sharp intake of breath (due to panic or tension) when falling behind in an interpreted rendering, or when struggling with unfamiliar content. Another common pitfall might be the overuse of the upheld open hand, facing palm out toward the interlocutor, to cut off the discourse (commonly used in both signed and spoken language interpreting). Jacobson (2007) observed that frequent interruptions of this type by student interpreters in the normal turn-taking sequence is extremely problematic in a healthcare context, for example, when a physician is attempting to process information provided by the patient to arrive at a diagnosis, or when a patient is receiving treatment instructions. It is posited here that the overuse of surface signals to cut off turns allows the interpreter to dominate the discourse, and short circuits cognitive processing of interlocutors, to the detriment of interactional goals.

Further research is called for to better understand the role of the interpreter as discourse manager. In addition, from this discussion emerges the difficulty in teasing apart competencies, such as medical knowledge, memory skills, note-taking strategies, and L2 turn-taking competence. This needs to be investigated empirically. The way that mediated discourse is managed directly impacts interactional goals, and, in turn, health outcomes, and therefore constitutes an integral part of interpreter performance assessment.

Interpreting studies

The idea that interactional competence is essential to interpreter performance, and therefore to assessment, is not new. In recent decades, a number of linguistics scholars have focused on the interactive nature of interpreting through empirical studies designed to analyze how it is that interpreters actually conduct themselves in community settings. Wadensjö (1998) was one of the first researchers to describe the community interpreter as an interlocutor who actively contributes to the dialogue and takes responsibility for the progression of talk. Her data, collected in a number of different community settings, including healthcare, show that the interpreter's role cannot be expected to be that of a "non-person" who is invisible and uninvolved. Wadensjö's work is grounded in the dialogic theoretical framework of Bakhtin. Similarly, Metzger (1999) shows in her research how interpreters manage discourse, minimizing and augmenting their participation to manage information flow. Her work is grounded, in part, in the theoretical frameworks of conversation analysis described in the previous section. Angelelli (2001, 2003) has also conducted research on the visibility of the interpreter in interpreted interaction in healthcare settings in particular. Her data, also grounded in discourse theory, suggest that interpreters are active participants who are involved in the co-construction of meaning in mediated interaction. Scholars such as Clifford (2001, 2005a), Davidson (1998, 2000), Napier (2006) and Roy (2000) have also delved into sociolinguistic aspects of interpreting, demonstrating that interpreters do much more than translate words, and must make complex decisions linked to situational variables. Their research portrays the interactive nature of spoken-language and signed-language interpreting through empirical research. And yet their findings and implications have largely been ignored in testing and assessment of interpreter performance within the medical and health literature.

Testing and assessment of interactional competence of interpreters

More research on testing and assessment in interpreter performance is beginning to emerge (this volume provides an example of some of the scholars working in the area). But little has been done that is relevant to the assessment of interpreter performance at the level of discourse, moving beyond lexico-semantic concerns. That is, although scholars have conducted empirical research that points to the discursive complexity of interpreting, this has not yet borne a significant impact on the way interpreters are assessed. Clifford (2001, 2005a) is one of the few scholars who has proposed grounding interpreter assessment in discourse theory. In his 2001 article, he provides three examples of discursive competencies to be considered in interpreter assessment, including deixis (which is particularly relevant to role shifting in sign language interpreting); modality (with a focus on intonation contour); and speech acts (their appropriate rendering in the target language). He proposes the use of rubrics to assess discursive competencies. Likewise, for his dissertation research, Clifford (2005a) developed a certification test that was based on a discursive model of interpreting. He conducted analyses to determine the psychometric properties of test items, and to determine the validity and reliability of the test. His work effectively demonstrates the need for the development of assessment instruments grounded in discourse theory.

The following section presents the development of an analytic rubric to assess discursive competencies of interpreters in healthcare. This rubric is grounded in the theoretical frameworks of IS and CA reviewed previously. Rubrics can be used to provide formative feedback to interpreters and student interpreters, and in summative evaluations, such as end-of-course exams, and in professional certification. In addition, they can be used for conducting quality assessment research in healthcare settings to identify potential areas for improvement in interpreter performance, and to better understand sources of miscommunication.

Measuring interactional competence in healthcare interpreters

Developing the instrument

The use of rubrics

Rubrics are used in language testing and assessment to measure primary and multiple traits, or competencies, in language production (Cohen 1994). The term *trait* refers to a particular ability or competency that is being measured. This type of instrument provides for an analytic or holistic approach to assessment that is commonly used to measure communicative competence in writing, but that can be applied to the measurement of spoken communicative competence, as well. Rubrics are particularly useful because they can be implemented in both formative and summative assessment, and the scales (numerical, levels of achievement, etc.) can be adapted according to the purpose of the rubric. It is possible to develop analytic rubrics, in which traits or sub-traits are scored separately, or holistic rubrics, in which traits or sub-traits are scored together (Mertler 2001; Arter & McTighe 2001, as cited in Clifford 2007a).

In developing an analytic rubric for performance-based interpreter assessment, competencies inherent to effective interpreter performance are identified and defined, or operationalized (Bachman & Palmer 1996), and a rating scale is used to score each of them separately (Mertler 2001). These scales often indicate different levels of achievement obtained in performance of a particular trait (Clifford 2001; Arter &McTighe 2001 as cited in Clifford 2007), such as Superior, Advanced, Fair, and Poor, although other types of scales can be implemented, depending on the objectives of the assessment. Assessments using rubrics are desirable because they provide detailed feedback on the particular abilities of the interpreter, and help to pinpoint where miscommunications may be taking place. They can therefore be implemented in a variety of settings: for formative assessment in interpreter education; for diagnostic purposes (such as quality assurance); for high-stakes situations, such as professional certification; and for research purposes in healthcare settings.

It is essential, as pointed out by both Clifford (2001) and Cohen (1994), that performance-based measurements be based on performances that are typical. That is, performances should be elicited in situations that closely resemble what interpreters actually face on a day-to-day basis. Validity of rubrics can be best established when “the test is based on expectations in a particular setting” (Cohen 1994: 322).

Operationalizing traits in interpreter performance

In the development of an analytic rubric, it is essential that the competencies to be measured be clearly defined, and grounded in theory. The test developer identifies each competency according to established theoretical frameworks, and then breaks them down into sub-traits, or sub-competencies. In her chapter in this volume, Angelelli discusses a similar approach in developing her five-point-scale rubric for the American Translators Association certification exam. The sub-components she identified for determining translation quality are based on frameworks of communication and communicative competence. Clifford (2001) suggests a similar approach to assessment of interpreting competence, basing the particular traits to be measured on theoretical discursive frameworks relating to deixis, modality, and speech acts.

To summarize, three important factors must be considered for the development of an assessment rubric: (1) selection of competencies to be measured must be grounded in theory; (2) traits and their sub-components must be operationalized; and (3) assessment must be of authentic performances or as close to authentic as possible. The following two sub-sections delve into the development of a rubric for assessing interactional features in interpreting performance. The competencies to be measured are grounded in the theoretical frameworks of IS and CA, their sub-competencies are defined and described, and sample rubrics are designed to be used in authentic contexts. As emphasized earlier, these competencies and their sub-components are not to be considered exhaustive. A great deal of research is still needed to better understand interaction in mediated discourse and to describe traits and sub-competencies for assessment.

Contextualization cues

The previous discussion on IS pointed to the need to assess competence in the appropriate transfer of contextualization cues, or paralinguistic features that signal meaning. Examples included intonation contour, eye gaze, body position, and other paralinguistic features. Inappropriate use or transfer of contextualization cues can lead to miscommunication, as in the example provided of monotonous renderings (which lack appropriate prosodic features). It follows, then, that contextualization cues should be included as one of the multiple traits in a scoring rubric. Note that the competency referred to as "contextualization cues" was identified based on the theoretical framework of IS, as selection of competencies to be measured must be grounded in theory. The next step is to identify the sub-competencies that are constitutive of this competency. Based on the IS framework, and on Jacobson's (2007) data on student interpreters, interpreters should be expected to: (1) demonstrate the ability to understand the meaning, in both languages, of such cues as voice volume, intonation, prosody, and other paralinguistic features accompanying the utterances of primary interlocutors; (2) produce effective and natural renditions of such cues in the target language; (3) demonstrate a balanced focus on both accuracy of information and interactional features; and (4) produce consistently dynamic (not monotonous) renderings with appropriate intonation contour in the target language. These four sub-competencies can be collapsed together for a more holistic scoring of each trait, or can be scored separately to provide for a more analytic assessment that teases apart the ability of the interpreter in relation to each of the sub-competencies. In Table 1, the four sub-competencies are grouped together.

The next step, after describing what the interpreter should be able to do (at the highest level), is to describe what would be expected to be the lowest level of performance with respect to the competency. For example, at the lowest level, an interpreter would (1) be unable to understand, in one or both languages, such cues as voice volume, intonation, prosody, and other paralinguistic features accompanying the utterances of primary interlocutors; (2) be unable to produce effective and natural renditions of such cues in the target language; (3) be unable to focus on both accuracy of information and interactional features at the same time; and (4) produce renderings with inappropriate intonation contour in the target language, or that are consistently monotone, characterized by excessive backtracking and stuttering.

The next step in rubric development is to define the continuum or scale to be used. In this case (see Table 1) four levels of achievement are used, to include Superior (highest level of performance), Advanced, Fair, and Poor (lowest level of performance). Such a continuum would be appropriate in an assessment being used in a professional setting, to indicate the level of interpreting being provided by on-site staff, for example, or if the rubric is going to be used to score an IEP exit exam. (However, it is should be noted that the validity and reliability of any testing instrument must be established before recommending its widespread use in a healthcare setting. This will be discussed further in the conclusion). Other types of scales can be used, as well. For example, for pre-testing candidates applying to an IEP, a scale ranging from Beginner to Advanced (e.g. Beginner-Advanced Beginner-Intermediate-Advanced) might be used.

Table 1 provides a rubric which shows how interactive competence in the use and transfer of contextualization cues could be assessed. Note that the number of levels may differ, depending on the degree of detail and feedback required. Each level must be distinguishable from the next, as demonstrated in the narrative descriptions provided for each level. Also, the sub-competencies of Contextualization Cues identified in the narrative descriptions can be broken down and scored separately, especially for providing formative feedback.

Table 1. Sample analytic rubric for competence in use and transfer of contextualization cues

	Contextualization Cues
Superior	Demonstrates superior ability in understanding meaning of contextualization cues (voice volume; intonation; prosody; and paralinguistic features) accompanying the utterances of primary interlocutors; produces effective and natural renditions of such cues in the target language; demonstrates balanced focus on both accuracy of information and interactional features; produces consistently dynamic renditions with appropriate intonation contour in the target language
Advanced	Demonstrates advanced ability in understanding meaning of contextualization cues (voice volume; intonation; prosody; and paralinguistic features) accompanying the utterances of primary interlocutors; is usually able to interpret such cues into target language, with some difficulty at times due to inability to consistently focus on both accuracy of information and interactional features; renditions are usually dynamic and appropriate, with occasional lapses into monotone renditions
Fair	Consistently demonstrates difficulty in understanding meaning of contextualization cues (voice volume; intonation; prosody; and paralinguistic features) accompanying the utterances of primary interlocutors; is often unable to interpret such cues into target language due to inability to focus on both accuracy of information and interactional features; renditions tend to be monotone and dull, characterized by frequent backtracking
Poor	Demonstrates clear inability to understand meaning of contextualization cues (voice volume; intonation; prosody; and paralinguistic features) in the utterances of primary interlocutors; is unable to interpret such cues into target language due to inability to focus on both accuracy of information and interactional features; renditions are consistently monotone, characterized by excessive backtracking and stuttering

Discourse management

Effective management of discourse, as demonstrated in the previous discussion on CA and turn taking, is another area of interactional competence that should be included in an analytic rubric for assessing interpreters' interactive competence. Jacobson (2007) posits that discourse management requires managing overlap and interruptions; understanding paralinguistic features that signal information related to turn taking; knowing when to take turns, and the signals to use to indicate the interpreter is taking a turn. It also involves getting the mediated interaction back on track if it derails. In her observations of student interpreters, she also noted that improved rapport (based on discourse flow and data obtained through interviews with providers) is established when healthcare provider and patient have eye contact and direct statements to each other, and that the interpreter is poised to play an integral role in facilitating direct interaction between the primary interlocutors using effective paralinguistic cues, such as head nods, eye gaze, and slight movements of the hand (this is a topic for future research).

Traditionally, a competent discourse manager is also required to provide a clear and concise introductory session (when possible) to clarify to the interlocutors how discourse management will be accomplished. To provide an example, in the *Bridging the Gap* training modules developed by Roat et al. (1999: 56) for interpreting in healthcare (mentioned previously as an example of training that is based on the conduit model), exercises are provided for carrying out an effective pre-session as part of "managing the flow" of an interpreted session. Assuming that such a session should always be required is recognized as problematic, however. Organizations such as the California Healthcare Interpreters Association (2002: 34) and the International Medical Interpreters Association (1998) have established guidelines stating that standardized interpreting protocols, such as pre-sessions, depend on time limitations, context, and urgency. It is therefore important to specify in the wording of a rubric that appropriate introductory sessions are conducted when possible. Competent management may also be linked to the appropriate use of the first and third person, although Bot (2005) suggests that changes in perspective in person may not bear as greatly on the interactional exchange as traditionally thought. The CHIA guidelines (2002: 38) also state that interpreters may use the third person for "languages based on relational inferences." In the rubric presented in this chapter, consistent use of first and third person is included, although further research is needed to better understand the impact of the first versus the third person on the effective flow of communication.

Discourse Management is the trait or competency that is represented in the rubric in Table 2. The sub-competencies are derived from the theoretical underpinnings of CA linked to turn taking. However, as noted above, other variables, such as a lack of memory skills and unfamiliarity with medical terminology, can also negatively impact discourse management. The question then arises as to whether memory, note taking, and terminology should be included as sub-competencies of discourse management. More research is needed in this area to determine how to isolate turn taking and other interactional strategies per se from other types of competencies, such as effective note-taking strategies and memory. Here, memory and note-taking skills are grouped with other competencies linked to discourse management.

In the rubric presented in Table 2, at the highest level, and as based on the CA framework and Jacobson's (2007) research, interpreters should be expected to: (1) provide a clear, concise pre-session to primary interlocutors on the interpreter's role in discourse management when possible (e.g. when an emergency or other uncontrollable factor would not prevent it); (2) consistently use the first person while interpreting, switching to third person for clarifications [but see Bot (2005)];

Table 2. Sample analytic rubric for competence in discourse management

	Discourse Management
Superior	Provides a clear, concise pre-session to primary interlocutors on interpreter's role when possible; consistently uses the first person while interpreting, switching to third person for clarifications; encourages interaction, including eye contact, between interlocutors, both verbally and through other paralinguistic cues; allows interlocutors to complete turns due to strong memory and note taking skills; demonstrates effective strategies for dealing with overlap
Advanced	Provides a clear, concise pre-session to primary interlocutors on interpreter's role when possible; consistently uses the first person while interpreting, switching to third person for clarifications; usually encourages interaction between interlocutors, both verbally and through other paralinguistic cues; usually demonstrates skill in allowing interlocutors to complete turns without interrupting for clarifications, with some difficulty due to need to further develop memory and note taking skills and build vocabulary; generally deals calmly and effectively with overlaps, with demonstrated need for further practice
Fair	In most cases, provides a clear, concise pre-session to primary interlocutors on interpreter's role, although at least one or two of the principal points are usually left out; is inconsistent in using the first person while interpreting, and exhibits excessive use of the third person, leading to awkward renditions; does not often encourage interaction between interlocutors, either verbally or through other paralinguistic cues; often interrupts interlocutors mid-turn for clarifications due to need to develop memory and note taking skills, and to build vocabulary; becomes nervous when challenged by overlaps, demonstrating clear need for further practice
Poor	Does not always provide a clear, concise pre-session to primary interlocutors on interpreter's role, leaving out principal points; is inconsistent in using the first person while interpreting, and almost always uses the third person, leading to awkward renditions; does not encourage interaction between interlocutors, either verbally or through other paralinguistic cues; does not allow interlocutors to complete turns, and interrupts frequently to request clarifications, resulting in choppy discourse; note taking and memory skills are poor; does not deal effectively with overlaps, leading to interruptions in the dialogue and excessive omissions

(3) encourage interaction, including eye contact, between primary interlocutors (such as healthcare professional and patients), both verbally and through the use of other paralinguistic cues; (4) allow interlocutors to complete turns without interrupting for clarifications (5) demonstrate obvious strong memory and note taking skills; (6) demonstrate effective strategies for dealing with overlaps.

At the lowest level, the interpreter (1) does not provide a pre-session to primary interlocutors on the interpreter's role in discourse management (when it is possible to do so); (2) is inconsistent in the use of first and third person while interpreting, leading to confusion among interlocutors; (3) does not encourage interaction between primary interlocutors; (4) does not allow interlocutors to complete turns, and interrupts frequently to request clarifications; (5) exhibits poor note taking and memory skills; (6) does not effectively deal with overlaps, leading to interruptions in the dialogue and excessive omissions.

The scale used for the competency of Discourse Management is the same as for Contextualization Cues, ranging from Advanced to Poor. As in the case of Contextualization Cues, the identified sub-competencies of discourse management can be broken down and scored separately.

A comprehensive rubric for assessing interpreter performance might include both Tables 1 and 2, to allow for assessment of interactional competence in using and transferring contextualization cues and managing discourse. Again, there are other areas of interaction to be considered in assessment, and this section has simply provided some basic guidelines for approaching these variables. A more comprehensive rubric would assess other areas of competence, such as the area traditionally referred to as "accuracy" (lexico-semantics) among many others. What is most essential to consider is that interpreter performance assessment must be grounded in theory and empirical research, and that the discursive elements of mediated interaction cannot be neglected.

Conclusion

This chapter has focused on assessing interpreter performance, with a focus on interactional competence in healthcare interpreting. Its purpose has been to look beyond words in assessment, and to explore discursive competence from within the frameworks of interactional sociolinguistics and conversation analysis. Examples were provided of ways in which miscommunication can occur due not to problems of accuracy or fidelity, but to a lack of competence in the use and transfer of paralinguistic features, and to the inability to effectively manage discourse.

The use of analytic rubrics was proposed for assessing interactional competence. Rubrics are developed for performance-based assessment through a step-by-step process that includes: (1) identifying the multiple competencies to be assessed, based on theoretical frameworks; (2) identifying the sub-traits of each competency and clearly defining them; (3) developing a scale or continuum that fits the purpose of the assessment (formative, summative, etc.). In addition, it has been noted that rubrics should be used to score performances that are elicited in situations as similar as possible to real-life interaction.

A number of limitations should be mentioned. First of all, the competencies and their sub-components described above and used in the sample rubrics are based on findings from research conducted by Jacobson (2007) with student interpreters in healthcare settings. This research was exploratory in nature, and a great deal more needs to be done to determine how effective communication can be facilitated by interpreters. In turn, it is impossible to know how effective mediated interaction can take place unless the interactional goals of any particular communicative event are fully understood and stipulated. A great deal remains to be done before the role of the healthcare interpreter, and corresponding competencies, can be established. In addition, it is important to keep in mind that testing instruments, such as rubrics, are living and dynamic, and subject to revision as findings from empirical research come to light.

Secondly, the reliability and validity of assessment instruments must be established before implementation on a wide scale, such as in quality assessment in a hospital setting, or for credentialing. This relates back to the need to define and establish the interpreter's role and relevant competencies for particular contexts. If a scoring rubric does not truly reflect the construct being measured (interpreter competence) any inferences or decisions made on the basis of the tests will be faulty (McNamara 1996). In addition, instruments must be validated *a posteriori* via data obtained through assessment of authentic performances (empirical and statistical validation) (Bachman & Palmer 1996). Reliability factors must also be considered to determine whether a rubric is consistent in its results. It is clear that scoring rubrics that will be used as diagnostic tools, or in other summative evaluations, cannot be developed in a haphazard manner, but must be grounded in theory, as discussed above.

This chapter has focused on the relevance of discourse theory to interpreter performance, with an emphasis on contextualization cues and turn taking. There are many other interactional features to consider in interpreting, and a great deal of research is still needed not only to define them, but to better understand how they impact mediated discourse. Such research will prove to be essential to both interpreter education and performance-based testing and assessment.

References

American Institutes for Research. 2005. *A Patient-Centered Guide to Implementing Language Access Services in Healthcare Organizations.* National Standards for Health Care Language Services Contract, Contract No. 282-98-0029, Task Order No. 48. Washington, DC.

Angelelli, Claudia. 2001. "Deconstructing the Invisible Interpreter: A Critical Study of the Interpersonal Role of the Interpreter in a Cross-Cultural/Linguistic Communicative Event." (Doctoral dissertation, Stanford University). *Dissertation Abstracts International,* 62, 9, 2953.

Angelelli, Claudia. 2003. "The Visible Collaborator: Interpreter Intervention in Doctor/Patient Encounters." In *From Topic Boundaries to Omission: New Research on Interpretation,* Melanie Metzger, Steven Collins, Valerie Dively and Risa Shaw (eds), 3–25. Washington, DC: Gallaudet University Press.

Bachman, Lyle F., and Palmer, Adrian S. 1996. *Language Testing in Practice.* New York: Oxford University Press.

Bernstein, Judith, Benrstein, Edward, Dave, Ami, Hardt, Eric, James, Thea, Linden, Judith, Mitchell, Patricia, Oishi, Tokiko, and Safi, Clara. 2002. "Trained Medical Interpreters in the Emergency Department: Effects on Services, Subsequent Charges, and Follow-Up." *Journal of Immigrant Health,* 4(4): 171–176.

Bot, Hanneke. 2005. "Dialogue Interpreting as a Specific Case of Reported Speech." *Interpreting,* 7(2*)*: 237–261.

California Healthcare Interpreters Association. 2002. *California Standards for Healthcare Interpreters: Ethical Principles, Protocols, and Guidance on Roles & Intervention.* Retrieved on January 20, 2009 at http://www.chia.ws/standards.htm.

Clifford, Andrew. 2001. "Discourse Theory and Performance-Based Assessment: Two Tools for Professional Interpreting." *Meta,* 46(2): 365–378.

Clifford, Andrew. 2004. "Is Fidelity Ethical? The Social Role of the Healthcare Interpreter." *TTR: traduction, terminologie, rédaction,* 17(2): 89–114.

Clifford, Andrew. 2005a. "A Preliminary Investigation into Discursive Models of Interpreting as a Means of Enhancing Construct Validity in Interpreter Certification." (Doctoral Dissertation, University of Ottawa). *Dissertation Abstracts International,* 66, 5, 1739.

Clifford, Andrew. 2005b. "Putting the Exam to the Test: Psychometric Validation and Interpreter Certification." *Interpreting* 7, 1: 97–131.

Clifford, Andrew. 2007a. "Grading Scientific Translation: What's a New Teacher to Do?" *Meta,* 52(2): 376–389.

Clifford, Andrew. 2007b. "Healthcare Interpreting and Informed Consent: What Is the Interpreter's Role in Treatment Decision-Making?" *TTR: traduction terminologie, rédaction.* 18(2), 225–247.

Cohen, Andrew D. 1994. *Assessing Language Ability in the Classroom. (2nd Edition).* Boston: Heinle & Heinle Publishers.

Davidson, Brad. 1998. *Interpreting Medical Discourse: A Study of Cross-Linguistic Communication in the Hospital Clinic.* Unpublished doctoral dissertation, Stanford University.

Davidson, Brad. 2000. "The Interpreter as Institutional Gatekeeper: The Social-Linguistic Role of Interpreters in Spanish-English Medical Discourse." *Journal of Sociolinguistics,* 4(3): 379–405.

Elderkin-Thompson, Virginia, Cohen Silver, Roxane and Waitzkin, Howard. 2001. "When Nurses Double as Interpreters: A Study of Spanish-Speaking Patients in a US Primary Care Setting." *Social Science & Medicine,* 52: 1343–1358.

Fagan, Mark J., Diaz, Joseph A., Reinert, Steven E., Sciamanna, Christopher N. and Fagan, Dylan M. "Impact of Interpretation Method on Clinic Visit Length." *Journal of General Internal Medicine,* 18: 634–638.

Flores, Glenn, Abreu, Milagros, Oliver, M. A. and Kastner B. 1998. "Access Barriers to Healthcare for Latino Children." *Archives of Pediatrics and Adolescent Medicine,* 152, 11: 1119–25.

Flores, Glenn, Laws, M. Barton, Mayo, Sandra J., Zuckerman, Barry, Abreu, Milagros, Medina, Leonardo and Hardt, Eric J. 2003. "Errors in Medical Interpretation and Their Potential Clinical Consequences in Pediatric Encounters." *Pediatrics* 111: 6–14.

Garcia, Estevan A., Roy, Lonnie C., Okada, Pamela J., Perkins, Sebrina D. and Wiebe, Robert A. 2004. "A Comparison of the Influence of Hospital-Trained, Ad Hoc, and Telephone Interpreters on Perceived Satisfaction of Limited English-Proficient Parents Presenting to a Pediatric Emergency Department. " *Pediatric Emergency Care,* 20(6): 373–378.

Goffman, Erving. 1981. *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Gumperz, John J. 1977. "Sociocultural Knowledge in Conversational Inference." In *The 28th Annual Roundtable on Language and Linguistics,* Muriel Saville-Troike (ed), 225–258. Washington DC: Georgetown University Press.

Gumperz, John J. 1982. *Discourse Strategies*. Cambridge: Cambridge University Press.

Gumperz, John J. 1984. "The Retrieval of Sociocultural Knowledge in Conversation." In *Language in Use: Readings in Sociolinguistics,* John Baugh and Joel Sherzer (eds), 127–137. Austin, TX: University of Texas Press.

Hampers, Louis C. and McNulty, Jennifer E. 2002. "Professional Interpreters and Bilingual Physicians in a Pediatric Emergency Department." *Archives of Pediatrics & Adolescent Medicine,* 156: 1108–1113.

Heath, Christian. 1992. "Gesture's Discreet Tasks: Multiple Relevancies in Visual Conduct and in the Contextualisation of Language." In *The Contextualization of Language,* Peter Auer and Aldo Di Luzio (eds), 101–128. Amsterdam: John Benjamins.

International Medical Interpreters Association. 1998. *Medical Interpreting Standards of Practice*. Retrieved on January 20, 2009 at http://www.imiaweb.org/standards/standards.asp.

Jacobson, Holly E. 2007, April. *Mediated Discourse and the Role of the Interpreter: What Can Be Learned from Foregrounding the Profile of the Mediator?* Paper presented at the meeting of the American Association of Applied Linguistics, Costa Mesa, CA.

Jacobs, Elizabeth A., Lauderdale, Diane S., Meltzer, David, Shorey, Jeanette M., Levinson, Wendy and Thisted, Ronald A. 2001. "Impact of Interpreter Services on Delivery of Health Care to Limited-English-Proficient Patients." *Journal of General Internal Medicine,* 16(7): 468–474.

Lee, Linda J., Batal, Holly A., Maselli, Judith H. and Kutner, Jean S. 2002. "Effect of Spanish Interpretation Method on Patient Satisfaction in an Urban Walk-In Clinic." *Journal of General Internal Medicine*, 17: 641–646.

McNamara, Tim F. 1996. *Measuring Second Language Performance*. London: Longman.

Mertler, Craig A. 2001. "Designing Scoring Rubrics for Your Classroom." *Practical Assessment, Research & Evaluation,* 7(25). Retrieved November 15, 2008 at http://PAREonline.net/getvn.asp?v=7&n=25.

Metzger, Melanie. 1999. *Sign Language Interpreting: Deconstructing the Myth of Neutrality*. Washington DC: Gallaudet University Press.

Moreno, María R., Otero-Sabogal, Regina, Newman, Jeffrey. 2007. "Assessing Dual-Role Staff-Interpreter Linguistic Competency in an Integrated Healthcare System." *Journal of General Internal Medicine,* 22 (Suppl 2): 331–335.

Napier, Jemina. 2006. "Effectively Teaching Discourse to Sign Language Interpreting Students." *Language, Culture and Curriculum,* 19(3): 251–265.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London: Routledge.

Roat, Cynthia E., Gheisar, Bookda, Putsch, Robert and SenGupta, Ira. 1999. *Bridging the Gap: A Basic Training for Medical Interpreters (3rd Edition)*. [Training Guide] Cross Cultural Health Care Program: Seattle.

Roy, Cynthia B. 2000. *Interpreting as a Discourse Process*. Oxford: Oxford University Press.

Saville-Troike, Muriel. 1989. *The Ethnography of Communication*. Oxford: Blackwell Publishers.

Schiffrin, Deborah. 1994. *Approaches to Discourse*. Oxford: Blackwell Publishers Inc.

Schiffrin, Deborah. 1996. "Interactional Sociolinguistics." In *Sociolinguistics and Language Teaching,* Sandra McKay and Nancy H. Hornberger (eds), 307–328. Cambridge: Cambridge University Press.

Schegloff, Emanuel A. 1998. "Body Torque." *Social Research* 65(3): 535–596.

Taveras, Elsie M. and Flores, Glenn. 2004. "Appropriate Services to Children in the Emergency Department." *Clinical Pediatric Emergency Medicine,* 5: 76–84.

Tannen, Deborah. 1984. Conversational Style: Analyzing Talk among Friends. Norwood, NJ: Ablex.

Tannen, Deborah. 1989. *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press.

Tannen, Deborah and Saville-Troike, Muriel. (1985). *Perspectives on Silence*. New Jersey: Ablex.

The Access Project. 2002. *What a Difference an Interpreter Can Make: Healthcare Experiences of Uninsured with Limited English Proficiency*. Boston: National Health Law Program.

Valli, Clayton and Lucas, Ceil. 1992. *Linguistics of American Sign Language: A Resource Text for ASL Users.* Washington DC: Gallaudet University Press.

Wadensjö, Cecilia. 1998. *Interpreting as Interaction*. New York: Addison Wesley Longman.

Woloshin, Steven, Bickell, Nina A., Schwartz, Lisa M., Gany, Francesca, and Welch, H. Gilbert. 1995. "Language Barriers in Medicine in the United States." *Journal of the American Medical Association,* 273(9): 724–728.

Woloshin, Steven, Schwartz, Lisa M., Katz, Steven J., and Welch, H. Gilbert. 1997. "Is Language a Barrier to the Use of Preventive Services?" *Journal of General Internal Medicine*, 12(8): 472–427.

PART II

The development of assessment instruments

Empirical approaches

The perks of norm-referenced translation evaluation

June Eyckmans, Philippe Anckaert and Winibert Segers

In this chapter we propose a method for translation assessment that implies a rupture with traditional methods where the evaluator judges translation quality according to a series of pre-established criteria. The Calibration of Dichotomous Items (CDI) is a method for evaluating translation competence that essentially transfers the well-known "item"-concept from language testing theory and practice to translation assessment. We will present the results of a controlled empirical design in which three evaluation methods will be compared with each other: a holistic (intuitive-impressionistic) method, an analytical method that makes use of assessment grids, and the calibration of CDI-method. The central research question focuses on the reliability of these methods with reference to each other.

1. Introduction

In recent years the field of translation studies has professed the need for more empirical evidence for the quality of translation tests (Waddington 2004; Anckaert & Eyckmans 2006; Anckaert et al. 2008). Although educational as well as professional organizations have implemented certification of translation skills on the basis of test administration, the reliability and validity of those tests remain underexplored. It seems that translation assessment around the world is more dependent on codes of practice than on empirical research.

Translation tests have a long history of serving as an indicator of language proficiency in schools, universities, and colleges around the world, although some language testers have raised serious objections to this practice (e.g. Klein-Braley 1987). With the current revival of the contrastive approach in Second Language Acquisition (Kuiken 2001; Laufer 2005) one might even say that translation is back in vogue as a measure of foreign language abilities. Ironically, it was remarked by Robert Lado in the sixties that "translation tests that are so common in testing language proficiency skills are not available as tests of the ability to translate" (Lado 1961: 261). To our knowledge, no method has been developed or disseminated to relate performance indicators to the underlying translation competence in a psychometrically controlled way (Anckaert et al. 2008). At least two explanations might account for this lack of psychometrically sound test development when assessing translation ability. First of all, the lack of validity of the translation test as a measure of language proficiency caused a certain loss of popularity of the format during the years of the Communicative Approach to language teaching (Widdowson 1978). As translation tests were not considered to be valid test formats for measuring language proficiency, they were not subjected to the same psychometric scrutiny as other language testing formats (such as the cloze or c-test). A second explanation concerns the epistemological gap that is experienced between protagonists of hard sciences (i.e. biology, chemistry, etc.) versus the humanities (literature, linguistics, translation and interpreting studies, etc.). The preconception that it is impossible to objectify language knowledge or translation quality without surrendering its essence is very tenacious among translator trainers and language teachers whose corporate culture exhibits a marked reticence towards the use of statistics (Anckaert et al. 2008). The fact that the teaching and testing of both translation and interpreting skills has generally been more in the hands of practitioners than of researchers has not helped in this regard either. Thanks to the introduction of psychometrics, numerous studies on the reliability and validity of language tests have been carried out in the domain of language assessment, but the domain of translation and interpreting studies has lagged behind in this respect. Today's practice of assessing translations is still largely characterized by a criterion-referenced approach. Both in the educational and the professional world, assessment grids are used in an attempt to make the evaluation more consistent and reliable (Waddington 2001; House 1981; Horton 1998; Al-Qinai 2000; Schmitt 2005). These grids reflect the criteria the evaluator (or the organization) sees as essential for determining the quality of the translation. They traditionally consist of a near-exhaustive taxonomy of different kinds of mistakes and/or bonuses (i.e. grammaticality, text cohesion, word choice, etc.) combined with a relative weight that is attributed to these categories (major or minor mistake). Although the use of assessment grids is motivated by the evaluator's wish to take the different dimensions of translation competence into account, one could argue that they fall short in adequately reducing the subjectivity of the evaluation, since the identification of dimensions of translation competence in itself is pre-eminently subjective (Anckaert et al. 2008). Aside from the inherent subjective nature of establishing sub-competences of translation ability, there are other factors that threaten the reliability of the test. Let's start with the evaluator's challenge of being consistent in her/his assessment when confronted with the task of scoring several tens of translations within a certain time span. Not only will the scoring be susceptible to order effects (contrast effects as well as halo effects, i.e. unintentional or unconscious preconceptions versus students with a weak/strong reputation), it is also difficult to maintain a sound testing practice in which one does not only distinguish the really good from the really bad translations, but where one can also discriminate within the large group of 'average quality' translations. Moreover, all scores have to be justifiable, and nowadays students exercise their rights (as they should) on equitable assessment of their performances.

Fortunately, researchers from the fields of translation studies (Pacte 2000; Waddington 2001; Conde Ruano 2005) as well as applied linguistics (Anckaert et al. 2008) are now taking up issues such as inter- and intra-rater reliability of translation assessment, and the construct validity of translation tests (see also the symposium "Testing in a multilingual context: The shift to empirically-driven interpreter and translator testing" organized by Helen Slatyer, Claudia Angelelli, Catherine Elder and Marian Hargreaves at the 2008 AILA World Congress of Applied Linguistics in Essen). Gradually the methodology of educational measurement together with the insights of language testing theory are being transferred to the field of translation and interpreting studies in order to arrive at valid and reliable ways to measure translation competence.

It is within this context that we will put forward a norm-referenced method for assessing translations, which we have developed with the aim of freeing translation assessment from construct-irrelevant variables that arise in both analytic (i.e. evaluating by means of pre-conceived criteria) and holistic (i.e. evaluating the performance as a whole in a so-called "impressionistic" or "global intuitive" way) scoring. The approach largely consists of transferring the well-known "item"-concept of traditional language testing theory and practice to the domain of translation studies and translation assessment. This implies a rupture with traditional methods of translation assessment, where the evaluator judges translation quality according to a series of pre-established criteria. A norm-referenced method is independent of subjective a priori judgments about the source text and the translation challenges it may encompass. Instead, the performance of a representative sample of the student population is used in order to identify those text segments that have discriminating power. Every element of the text that contributes to the measurement of differences in translation ability acquires the "item"-status. These items are selected in a pre-test procedure. Afterwards these items function as the sole instances on which to evaluate the translation performance in the actual summative/diagnostic test. As in language testing, the norm-referenced method presupposes a dichotomous approach of the text segments: a translated segment is either acceptable or not (i.e. correct or not). There is no weighing of mistakes and/or bonuses against other alternatives. This does not imply that there is only one appropriate translation for the text segment in question; it merely means that for each translated segment it is agreed between graders which alternatives are acceptable and which are not. Since the method is based on the practice of calibrating segments of a translation, we call it the Calibration of Dichotomous Items-method (CDI-method). The different steps that lead to the calibration of items and tests (pre-testing, establishing discrimination power and estimating the test reliability, which will be discussed in Section 4) allow the construction of standardized tests of translation (across languages). It is clear that this does not constitute a time-saving procedure; therefore the method is only to be promoted for use in summative contexts (high stake situations where decisions have to be made). This norm-referenced approach holds the promise that it is a stable and evaluator-independent measurement that bridges the gap between language testing theory and the specific epistemological characteristics of translation studies. It is exactly this promise that we will put to the test in the empirical study that we will report in this chapter. Three approaches to translation evaluation will be compared on their psychometric qualities in a controlled empirical design: the holistic method, the analytic method (both criterion-referenced approaches), and the CDI-method (norm-referenced approach).

In this article we will gather data within an educational context. However, the resulting knowledge and insights concerning the assessment of the construct of translation competence can also be applied in professional contexts.

2. Research hypothesis

The central research question of the study focuses on the reliability of the three different methods with reference to each other. Reliability coefficients express the extent to which a test or a method renders consistent (and therefore reliable) results. In the case of translation assessment, one could argue that a scoring method is reliable when a translation product or performance is assessed similarly by different assessors who use the same method. Put in laymen's terms, a reliable method or test means that a student or translator can be confident that his or her translation performance will obtain a similar score whether it is assessed through one method or the other, or whether it is corrected by one assessor or another.

When a test consists of separate items (as in a Multiple Choice Language Test or in a Cloze Test), the traditional way of estimating its reliability is through the calculation of Cronbach's Alpha. The reliability of the translation scores obtained through the CDI-method can therefore be calculated directly by means of the Cronbach's Alpha reliability index (Bachman 2004). When there are no discernable items in the test – as is the case with the holistic and analytic method – the reliability is estimated indirectly by means of a correlation coefficient. This correlation coefficient expresses the extent to which the scores of different graders coincide with each other. More particularly, the correlation coefficient calculates the overlap between the graders' rank orders of the translators' performances. The reliability of the holistic and analytic method will be estimated indirectly by means of Spearman rank correlation coefficients (Bachman 1990; Bachman & Palmer 1996) of the different assessors' scores. This is called the inter-rater reliability.

The research question is split up into different research hypotheses, which read as follows:

1. The inter-rater reliability between the graders will be the weakest in the case of the holistic evaluation method.
2. The inter-rater reliability between the graders who used the analytic method will be in between that of the holistic method and that of the CDI-method (which will be maximal, given the fact that the test scores are based on a calibrated selection of items).
3. Because of a possible lack of reliability of both the holistic and the analytic methods, they will not be able to discriminate the students' performances sufficiently. In the case of the CDI-method, the discriminating power is maximized because the "items" (translation segments) are selected according to their r_{it}-value (the amount in which the item contributes to the test's global reliability).

3. Method

3.1 Participants

A total of 113 students participated in this study. They were enrolled in the first (n = 22), second (n = 39) or third year (n = 38) bachelor-level Dutch-French translation courses, or the Dutch-French master (n = 14) translation course from the four francophone Translation Colleges in Belgium (University College Brussels, University College Francisco Ferrer, University College Leonardo da Vinci and University Mons-Hainaut) (for details on the number of participants per college: see Table 1). In targeting all four Translation Colleges and gathering translation performances from bachelor as well as master students, we have

Table 1. Number of participants per grade and per college

	BA, 1st	BA, 2nd	BA, 3rd	Master	Total
University College Brussels	7	13	21	8	49
University College Francisco Ferrer	4	1	2	1	8
University College Leonardo da Vinci	0	13	15	5	33
University Mons-Hainaut	11	12	0	0	23
Total	22	39	38	14	113

Notes: Due to practical constraints during the data collection (illness or absence of students), some of the cells contain few or no participants

reached the entire population of Dutch-French translation students in Belgium. Their ages range from 18 to 25. The translation assignment was carried out in class during college hours under supervision of one of the researchers and their translation trainer.

3.2 Materials and procedure

The translator students of the four different translation colleges were asked to translate a relatively short text (346 words) from Dutch (language B) into French (language A) (see Appendix 1). The text in question deals with the advertising world and can be characterized as a non-specialized text written in standard journalistic style. It does not contain terminological challenges, the linguistic complexity is average and the content matter is easily accessible (not requiring prior knowledge about advertising or marketing). As such, the length, type and difficulty of the text can be considered indicative of the materials used in the students' translation courses. The choice for this text in terms of register and subject was motivated by the fact that the assignment should not be too difficult for the BA1 students, yet challenging enough to discriminate between the performances of the master-level students.

The translations were handed to two graders who were asked to correct them using a holistic approach (see Appendix 2), and two graders who were asked to correct them along the lines of a clearly specified analytic method (see Appendix 3). Care was taken not to train the graders with reference to this experiment since they were specifically selected on the basis of their longstanding experience in translation assessment and/or concordance expertise. The research team meanwhile graded the translation performances according to the norm-referenced CDI-method. First, they identified all eligible text segments (in other words, all translation mistakes) that could serve as items of a calibrated translation test. The identification of eligible text segments worked as follows: two members of the research team (both bilinguals Dutch/French and experienced translator trainers) separately went through all student translations and registered all translated segments that could possibly be construed as mistakes (be it misinterpretations, grammatical errors, unsuited lexical choices, omissions, additions, spelling errors, literal translations, doubtful register use, lack of idiomaticity or stylistic mistakes). In a second phase, they compared the results of their prospective grading work with the goal of arriving at one long list of translation errors. Establishing this list is a cumulative process: all text segments that have been identified as a mistake by one of the graders are taken into account and added to the list. The result is a list that contains the total number of translation errors that have been made by the 113 students when performing this particular translation. In a second phase the discriminating power of (r_{it}-values) all these translation errors is calculated in order to determine the set of calibrated items (in other words: to distinguish items with good discriminating power from items with weak or negative discriminating power), and the internal consistency of the test is calculated by means of Cronbach's Alpha.

The students were given 90 minutes to perform the translation. The translation was carried out with pencil and paper. The students were also allowed to access their computers in order to use whatever electronic or other reference they saw fit because we wanted to make sure that heuristic competence was included in the translation performance, since we consider it an integral part of all translation work.

The students were not allowed to confer with each other during the assignment. They were given instructions in their mother tongue (French) (see Appendix 1), and the data gathering was supervised by their regular translation trainer and one of the authors.

3.3 Data processing

The graders had been informed about the quasi-experimental design before they agreed to participate. They received a reasonable payment for the amount of time it took them to score all the translations. We approached several graders and we selected them according to their profile. We also granted them the liberty of choosing the correction method they preferred to work with (either the holistic method or the analytic method). This resulted in a selection of four graders for the experiment. Two of them are experienced translation trainers (one male and one female) with more than 20 years of experience in assessing students' translation performances within a university college context. Two other graders (one male and one female) were selected because of their longstanding experience in the

Concordance Service of the Belgian Council of State. The Concordance Service is the highest authority in the verification of the equivalence of legal and official source documents and their translations. Translators who work in this service are called revisers. The graders' age ranged between 44 and 63.

We made the deliberate choice not to train the graders (for information on the advantages and disadvantages of rater training see Weir 1994 and 2005). Instead all four graders were selected because of their longstanding experience in translation quality assessment. We made sure that they could follow the assessment approach to which they have been accustomed throughout their professional careers. The profiles of the graders who participated in the experiment are as follows:

Grader 1 (henceforth "Train-An" for Trainer-Analytic) is a translator trainer who has been using an analytic method throughout his career. His scoring sheet was examined and approved by Grader 3, who was also going to mark the students' translations by means of an analytic method.

Grader 2 (henceforth "Train-Ho" for Trainer-Holistic) is a translator trainer who has always used a holistic method when assessing students' performances. She is convinced of the merits of her approach.

Grader 3 (henceforth "Re-An" for Reviser-Analytic) is a professional reviser who favors working with an analytic assessment sheet. He approved the assessment sheet of grader 1.

Grader 4 (henceforth "Re-Ho" for Reviser-Holistic) is a professional reviser who approaches translation assessment holistically and who has agreed to continue doing so in this study.

4. Results and discussion

During the process of test calibration, two members of the research team worked in close collaboration in order to identify all translation errors in the translations (n = 113). Both members have more than 15 years of experience in scoring translations in an educational as well as a professional context. Whenever these graders were in disagreement about whether or not a particular translation 'item' was to be accepted or not, the item was included. This means that differences in the graders' appreciation of translation alternatives were never discarded. Instead, possible variance of what could be called graders' subjectivity was embraced and included for further analysis. Every instance of translation performance can be considered as an item (lexical choice, grammatical mistakes, spelling, punctuation, register, style, etc). Their work resulted in a list of 170 items. A matrix was developed in which all these items were listed and every student received a 0 or 1 score on the basis of his or her performance on the item in question. On the basis of this matrix, it was possible to apply some basic psychometrics from classical language testing theory.

For all 170 items, the corrected item-total correlations were calculated in order to distinguish good discriminating translation items from other ones. The corrected item-total correlation (henceforth called r_{it}-value) calculates the correlation between the item and the rest of the scale, without that item considered as part of the scale. If the correlation is low, it means the item is not really measuring what the rest of the test is trying to measure. With r_{it}-values higher than .3 considered a threshold value for good discriminating items (Ebel 1979), 77 translation items were withheld as items to be included in a calibrated translation test (see Appendix 4). When we calculated the test reliability on the basis of these items only, we arrived at a Cronbach's Alpha of .958. This is a very high reliability coefficient, as was expected given the fact that the items had been selected according to their discriminating power.

Let us now look at how the graders did in terms of reliability when they proceeded holistically or analytically in the marking process. Table 2 shows the Spearman rank correlation coefficients between the four graders. The correlation coefficients range between .635 and .780. With reference to our first research hypothesis, we can confirm that the graders of the holistic method obtain a lower agreement on the rank orders of the participants than the graders of the analytic method (r = .669 versus r = .740). This seems to indicate that the use of an analytic assessment grid results in a higher inter-rater reliability than holistic marking. However, the Spearman rank correlation coefficient of .740 is still far from the reliability estimate that was obtained with the CDI-method (Cronbach's Alpha: .958). This confirms the second research hypothesis of this study: the inter-rater reliability obtained between the graders who used the analytic method is in between the inter-rater reliability of the graders who used the holistic method and the reliability obtained with the CDI-method.

Table 2. Spearman rank correlation coefficients between graders (N = 113)

	Train-An	Train-Ho	Re-An	Re-Ho
Train-An		.735**	.740**	.720**
Train-Ho			.780**	.669**
Re-An				.635**
Re-Ho				

** Correlation is significant at the 0.01 level (2-tailed)

It should also be remarked that although the Spearman rank correlation coefficients are shown to be statistically significant, they do not reflect a satisfying inter-rater reliability within this particular assessment context. Even the highest correlation coefficient between the graders (r = .780) explains merely 61% (r^2 = 60.84) of the variance between both graders. The predictive value of one grader's scores for the outcome of another grader's scores is, in other words, limited. This means that the rank orders of the scores that the graders attributed to the same translation performances vary substantially. The different scores that have been attributed to the same translation performances by the four graders are illustrative of this. The translation of participant 44, for example, receives a zero from Trainer Analytic, four out of twenty from Trainer Holistic, three out of twenty from Reviser Analytic, but no less that fourteen out of twenty from Reviser Holistic. The translation of participant 94 receives much higher but equally disperse marks from the respective graders (9.5/20, 18/20, 11.5/20, 15/20). It is clear that, notwithstanding the conscientious grading process of both the translation trainers and the revisers, the inter-rater reliability between the graders of both the holistic and the analytic evaluation method is unconvincing. We can only conclude that if we were to transfer these data to a realistic setting – a translation exam within an educational context or the selection of translators within a professional context – the participants' chances of succeeding at the test would be highly dependent on the grader in question.

When we look at the descriptive statistics of the scores that have been attributed by the four graders, we notice (see Table 3) small differences in the obtained means, with the exception of the translation trainer who used the analytic method (mean of 2.58 out of twenty versus 9.77, 8.34 and 9.62). Further analysis of these data reveals that the trainer in question attributed many zero scores to the students' translation performances: the mode and median for this teacher are zero. Also, the range of scores that he has attributed is lower than that of the others, possibly indicating the smaller discriminating power of the scores this grader attributed. Although it is apparent that this grader's marking behavior is much stricter than that of the other graders, this did not significantly influence the inter-rater reliability (see Table 2). There is, however, a clear bottom effect at the low end of the scale and consequently no opportunity to discriminate between the many translation performances that obtained a zero according to this grader's use of the assessment grid. It should be noted that the reviser-grader who scored analytically used the same assessment grid and marked the same translation performances, yet distributed much higher marks than his colleague.

Table 3. Descriptive statistics of the graders' scores (N = 113)

	Train-An	Train-Ho	Re-An	Re-Ho
Mean	2.58	9.77	8.34	9.62
Median	.00	11.00	8.00	10.00
Mode	.00	12.00	.00	10.00
Standard deviation	3.82	4.92	4.99	4.57
Range	13.00	18.00	17.00	16.00
Minimum	.00	.00	.00	2.00
Maximum	13.00	18.00	17.00	18.00

The third research hypothesis of this study concerned the discriminating power of the assessment methods. We hypothesized that a possible lack of reliability of both the holistic and the analytic approach might reveal a flawed discrimination between the participants' performances, in contrast with the CDI-method that is designed to maximize the discriminating power through the selection of "items" according to their r_{it}-value (the amount in which the item contributes to the test's global reliability). In order to investigate this question, scores obtained with the CDI-method were calculated and compared with the scores that resulted from the analytic and holistic method. In order to determine a score on the basis of the calibrated items, the classical procedure for calculating test scores on the basis of raw scores was used: (((raw score – mean)/SD)*desired SD) + desired mean). In other words: a participant's test score equals his or her raw score (on a total of 77 calibrated items) minus the mean (59.38 in this case), divided by the standard deviation (SD 14.95). The resulting number is then multiplied by a chosen standard deviation, and to this a chosen mean different from zero is added (in our case: 14.63) in order to maximize the range of scores. This classical procedure was followed in order to obtain scores that would be comparable to the scores awarded by the graders of the analytic and holistic method (who were asked to grade along the traditional 0 to 20 scale).

To check whether the use of the different methods resulted in scores that discriminate sufficiently between the participants, the different educational levels were taken into account (BA1, BA2, BA3 and Master level). Table 4 and Figure 1 show the score distribution for the three methods.

Table 4. Score distribution for the three methods per level

	Analytic method (mean)	Holistic method (mean)	CDI (mean)
BA1 (n = 22)	2.40	6.80	11.67
BA2 (n = 39)	5.19	9.55	14.92
BA3 (n = 38)	6.20	10.20	15.23
MA (n = 14)	9.02	12.93	16.88

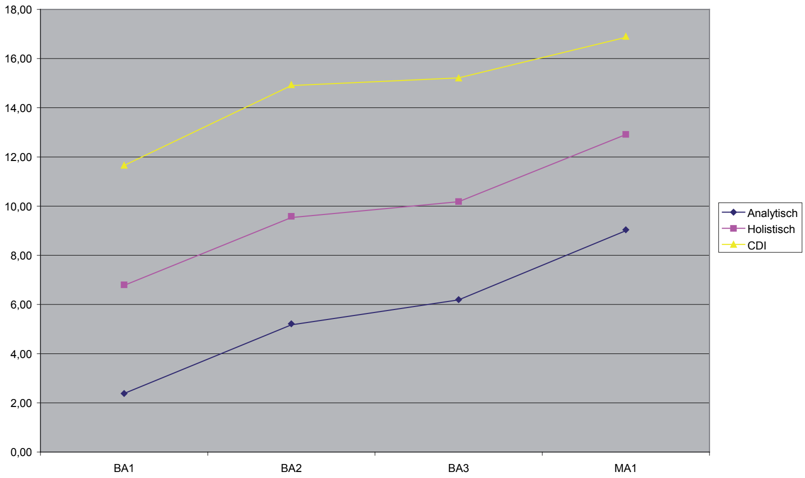


Figure 1. Graphic representation of the score distribution for the three methods.

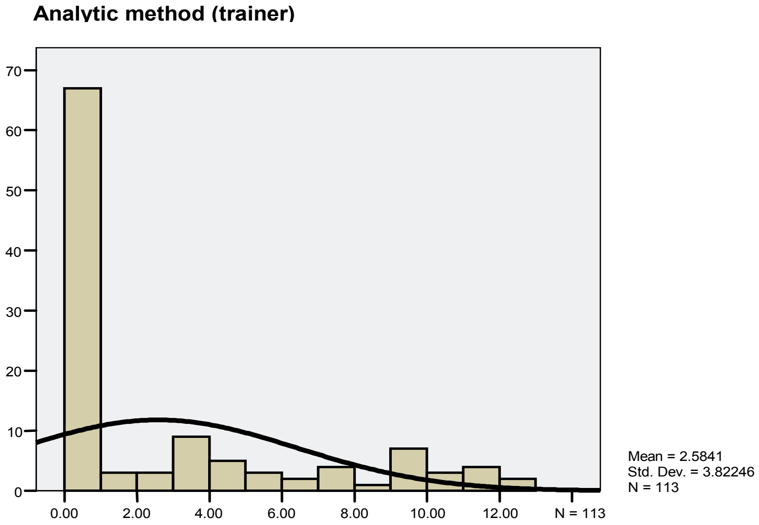


Figure 2. Distributions of scores awarded by the translation trainer who graded according to the analytic method

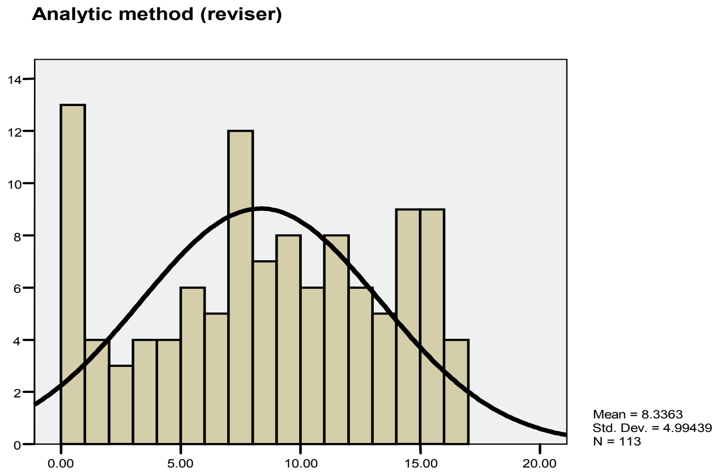


Figure 3. Distributions of scores awarded by the reviser who graded according to the analytic method

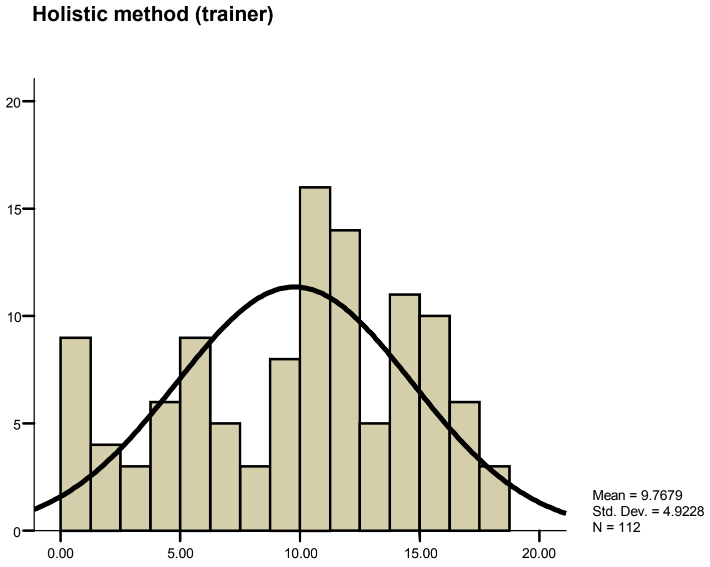


Figure 4. Distributions of scores awarded by the translation trainer who graded according to the holistic method

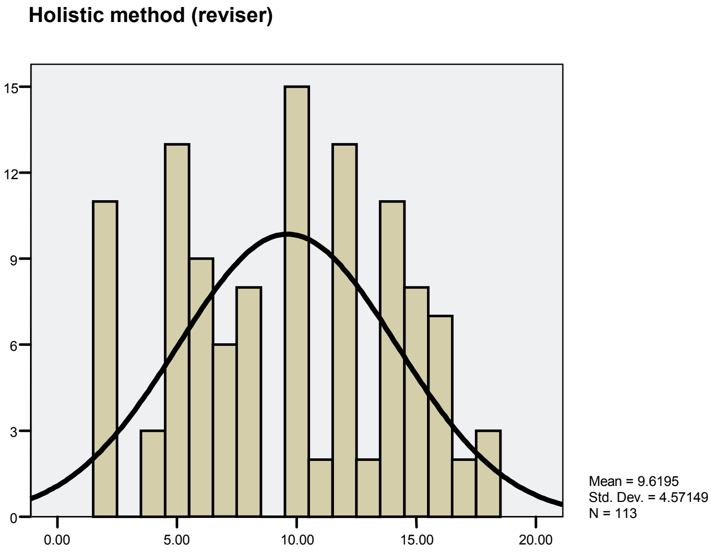


Figure 5. Distributions of scores awarded by the reviser who graded according to the holistic method

Although there are relatively large differences in the obtained means, the score distribution looks quite similar for the three methods. This seems to indicate that all three methods perform well in distinguishing weak students from intermediate students and advanced students. However, it has to be pointed out that these means represent global effects for groups and are the result of scores that have been attributed by two graders. This conceals the large differences in individual grader variation that we have pointed out earlier in this results section. When we look at the individual score distributions of the different graders (Figures 2 to 5), the differences in score allocation depending on the grader and the particular method that was applied are apparent.

Figures 2 to 5 illustrate how the means in Table 4 conceal the marked differences in the ways the different graders discriminate between student performances, even when they are applying the same correction method.

5. Conclusion: Implications and limitations of the study

5.1 Implications

On the basis of the results obtained with 113 participants, we have to conclude that the holistic and analytic methods that were used by the graders in this study fall short in terms of test reliability and discriminating power. The analytic method seems to lead to a higher level of inter-subjective agreement between graders than the holistic method, but it still contains too much variability when the individual score distribution is looked at in detail. Both methods undoubtedly have their merits in formative classroom evaluation of translation performances, but they are difficult to apply consistently because of the doubtful stability of criteria that are used to evaluate. However, holistic as well as analytic assessment approaches are still regularly used in summative assessment (translation exams within an educational context, certification of translation skills in a professional context, etc.) and caution is warranted. Since these methods are essentially based on subjective approaches towards translation assessment, their reliability is questionable, and the justification of scores can turn out to be very problematic (lack of justification of scores might even lead to litigation, as has been the case a couple of times in Belgium already).

The CDI-method relates the performance indicators to the underlying translation competence in a more reliable and consistent way, since it is based on the selection of discriminating items or translation segments. However, it is important to note that the representative nature of the sample is of overriding importance when it comes to identifying the discriminating items in a translation test. This means that the implementation of the method requires constant monitoring. Therefore the method is only to be promoted for use in summative contexts (high stake situations where decisions have to be made). An important advantage of the method is that it ties in nicely with the European call for test standardization in the sense that it can contribute to the reliable and valid certification of translation competence in a European context. It allows an exploration of test robustness so that tests can be validated for different language combinations and language-specific domains. However, the construction of a reliable battery of translation tests will be a time-consuming process – as it has been in the domain of language testing – and sufficiently large populations will be needed in order to safeguard item stability.

5.2 Limitations

Among the limitations of this study are the relatively small number of participants at the BA1 and Master level, and the fact that the translations were carried out with paper and pencil. In a replication of the study on a larger scale, care will be taken to create an environment that mimics a real professional translation assignment by letting the participants perform the translation on computer.

Although the CDI-method is not difficult to apply, it does require a thorough analysis of all possible translation errors and a minimal understanding of the basics of psychometrics. If the assessment of translations is to become more scientific, an introduction to the fundamental principles of language testing is warranted for translation trainers. With this chapter, we hope to have contributed to what we would like to call *translation assessment literacy* by bridging the gap between language testing theory and the specific epistemological characteristics of translation studies.

Further research on the already assembled data will be directed at an in-depth investigation of the professional profiles of the graders with regard to the score distributions. The data assembled in this study also seem to coincide with the theoretical assumption that the CDI-method is inclusive of every possible dimension of translation ability in so far as the interaction of a particular text with a particular population gives rise to the different dimensions of translation ability. This will be taken up in a forthcoming article. Finally, a future research project will rise to the methodological challenge of achieving equivalent standards for translation competence across languages and applying the CDI-method with more distant language pairs.

References

Al-Qinai, Jamal. 2000. "Translation Quality Assessment. Strategies, Parameters and Procedures." *Meta* 45(3)*:* 497–519.

Anckaert, Philippe and Eyckmans, June. 2006. "IJken en ijken is twee. Naar een normgerelateerde ijkpuntenmethode om vertaalvaardigheid te evalueren." In *Vertalingen objectief evalueren. Matrices en ijkpunten*, Van de Poel, Chris and Segers, Winibert (eds.), 53–67. Louvain/Voorburg: Acco.

Anckaert, Philippe, Eyckmans, June and Segers, Winibert. 2008. "Pour une évaluation normative de la compétence de traduction." *ITL International Journal of Applied Linguistics* 155: 53–76.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, Lyle F. & Palmer, Adrian S. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, Lyle F. 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Conde Ruano, Tomás. 2005. *No me parece mal. Comportamiento y resultados de estudiantes al evaluar traducciones*. University of Granade: Unpublished doctoral dissertation.

Ebel, Robert L. 1979. *Essentials of Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall.

Horton, David. 1998. "Translation Assessment: Notes on the Interlingual Transfer of an Advertising Text." *IRAL* 36(2*):* 95–119.

House, Juliane. 1981. *A Model for Translation Quality Assessment*. Tübingen: Gunter Narr.

Klein-Braley, Christine. 1987. "Fossil at Large: Translation as a Language Testing Procedure." In *Taking Their Measure: The Validity and Validation of Language Tests*, Grotjahn, Rüdiger, Klein-Braley, Christine and Stevenson Douglas K. (eds.). 111–132. Bochum: Brockmeyer.

Kuiken, Folkert. 2001. "Contrastief en taakgericht: een contrast." In *Perspectieven voor de internationale neerlandistiek in de 21ste eeuw: handelingen Veertiende Colloquium Neerlandicum, Katholieke Universiteit Leuven, 27 August-2 September 2000 / IVN, Internationale Vereniging voor Neerlandistiek*, Elshout, Gerard et al. (eds.). 353–362. Münster: Nodus-Publ.

Lado, Robert. 1961. *Language Testing. The Construction and Use of Foreign Language Tests. A Teacher's Book*. London: Longmans.

Laufer, Batia. 2005. "Focus on Form in Second Language Vocabulary Learning." In *EUROSLA Yearbook Volume 5*, Foster-Cohen Susan H., Garcia-Mayo María and Cenoz, Jasone (eds.). 223–250. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pacte. 2000. "Acquiring Translation Competence: Hypotheses and Methodological Problems in a Research Project." In *Investigating Translation*, Beeby, Allison, Ensinger, Doris, Presas, Marisa (eds.). 99–106. Amsterdam: John Benjamins.

Schmitt, Peter A. 2005. "Qualitätsbeurteilung von Fachübersetzungen in der Übersetzerausbildung. Probleme und Methoden." Paper presented at Vertaaldagen Hoger Instituut voor Vertalers en Tolken, 16–17 March 2005,
http://www.hivt.be/onderzoek/vertaaldagen2005_verslag.htm.

Waddington, Christopher. 2001. "Different Methods of Evaluating Student Translations: The Question of Validity." *Meta* 46(2): 331–325.

Waddington, Christopher. 2004. "Should Student Translations be Assessed Holistically or through Error Analysis?" *Lebende Sprachen* 49(1): 28–35.

Widdowson, Henry G. 1978. *Teaching Language as Communication*. Oxford: Oxford University Press.

Weir, Cyril. 1994. "Effects of Training on Raters of ESL Compositions." *Language Testing* 11(2): 197–223.

Weir, Cyril. 2005. *Language Testing and Validation*. Basingstoke/New York: Palgrave Macmillan.

Appendix 1. Translation assignment and instructions

Consignes

Un éditeur français vous a pressenti pour assurer la traduction d'un ouvrage de vulgarisation, paru il y a peu en néerlandais, sur l'inefficacité de la publicité commerciale. Afin de décrocher ce contrat de traduction, vous devez lui soumettre, en guise d'échantillon, la traduction française du texte ci-dessous. Vous ne devez vous soucier d'aucun aspect lié à la mise en page, mais vous apporterez un soin particulier à tous les aspects linguistiques (correction grammaticale, précision lexicale, ponctuation, ton, etc.).

Vous pouvez recourir à tous les outils d'aide à la traduction que vous jugerez nécessaires (ouvrages de référence, ressources en ligne) et disposez de 90 minutes pour vous acquitter de cette tâche. Bon travail!

Instructions

A French editor is considering you for the job of translating a non-specialist work on the ineffectuality of commercial advertising which recently appeared in Dutch. In order to secure this contract, you need to present him with the French translation of the text below as a sample. You do not need to concern yourself with any aspect related to the formatting of the text, but you do need to pay close attention to all linguistic aspects (grammatical correctness, lexical precision, punctuation, tone, etc.).

You are free to utilize any translation tool deemed necessary (reference works, on-line resources). You have 90 minutes at your disposal to bring this assignment to a successful completion. Good luck !

Texte / *Text*

Media en amusement trekken alle aandacht naar zich toe

Een van de grootste misvattingen die ik bij de meeste mensen heb vastgesteld over reclame, is dat zij ervan uitgaan dat al het geïnvesteerde reclamegeld van bedrijven terechtkomt in de 'reclamewereld', waarmee zij de wereld van reclamebureaus en reclamemakers bedoelen. Dat is niet zo. Het overgrote deel van de reclame-investeringen van het bedrijfsleven gaat naar de aankoop van ruimte in de media, en komt dus terecht op de bankrekeningen van de mediagroepen met hun tijdschriften, kranten, radiozenders, tv-stations, bioscopen, billboards... Bedrijven zijn immers op zoek naar een publiek om hun producten bekend en geliefd te maken, in de hoop dat publiek ervan te kunnen overtuigen hun producten ten minste eens te proberen. Dat publiek wordt geleverd door de media. De bedrijven kopen dus pagina's of zendtijd, en kunnen zo in contact treden met het publiek van die media. Op die manier ontstaat er een miljardenstroom van reclamegeld (in België meer dan 1,75 miljard euro per jaar), die van de bedrijven naar de media stroomt.

De reclamebureaus staan slechts aan de oevers daarvan. Zij zijn de kleine vissers, hengelend naar een opdracht van de bedrijven die er vooral in bestaat de aangekochte ruimte te vullen met inhoud. De reclamebureaus zorgen dus voor de ontwikkeling van de boodschap van het merk en voor de verpakking van die boodschap. In ruil daarvoor krijgen ze een percentage van de reclame-investeringen: vroeger ging dat om 15%, nu is het meestal minder dan 10%. Steeds meer worden deze percentages afgeschaft en vindt de betaling plaats via een maandelijks honorarium, dat door de bedrijven zwaar onder druk wordt gezet. Zij moeten immers steeds meer betalen voor hun mediaruimte en willen die meerkosten zo veel mogelijk terugverdienen, onder meer via de reclamebureaus. Deze worden verplicht steeds sneller te werken voor steeds minder geld. Dat wil niet zeggen dat de reclamebureaus armoedezaaiers zijn. Veel reclamemakers verdienen goed. Door mee te surfen op de golven van de reclame-investeringen zijn er multinationale en beursgenoteerde reclamenetwerken ontstaan. Maar de échte reclamewereld, waarin het grote reclamegeld omgaat, is eigenlijk de mediawereld.

Appendix 2. Instructions for the holistic graders

Avant de corriger, veuillez prendre connaissance des consignes qui ont été fournies aux étudiants et qui sont reprises au-dessus du texte à traduire.

Lors de la correction des traductions selon la méthode holistique, vous soulignerez dans la traduction tout ce qui ne va pas, mais sans préciser la nature de la faute ou appliquer un quelconque barème, et vous attribuerez une note entre 0 et 20 correspondant à votre impression globale en tant que correcteur.

Before marking the translations, please make sure you have carefully read the instructions given to the students, to be found above the text to be translated.

As you grade, you will underline anything in the translation that does 'not sound right', in line with the holistic method, without giving specific information about the nature of the error or applying any kind of scoring parameter. At the end, you will supply a grade between 0–20 which you feel corresponds to the impression you obtained from the translation as a whole.

Appendix 3. Instructions for the analytic correctors

Avant de corriger, veuillez prendre connaissance des consignes qui ont été fournies aux étudiants et qui sont reprises au-dessus du texte à traduire.

La méthode analytique consiste à corriger selon la grille d'évaluation reprise ci-dessous. Cette manière de procéder implique que le correcteur souligne chaque faute (de langue ou/et de traduction) et indique dans la marge la nature de la faute sous forme d'abréviation (p.ex. 'CS' pour contresens, 'GR' pour grammaire, etc.). Ensuite, il retranche du total de 20 un nombre de points par faute (ex. –2 pour un CS, –0,5 pour une faute GR, etc.).

Before grading the translation, please make sure you have carefully read the instructions given to the students, to be found above the text to be translated.

The analytical method entails that the translations be marked according to the evaluation grid provided below. This method implies that the corrector underlines every error (both regarding the language and/or the translation) and provides information in the margin as to the nature of the error (e.g. 'CT' for content errors or misinterpretations; 'GR' for grammatical errors, etc.). Finally, a number of points will be deducted from a total of 20 points for each error found (e.g. –2 per CT mistake ; –0.5 per GR mistake, etc.).

SENS (*meaning or sense*)	Toute altération du sens dénotatif : informations erronées, non-sens… J'inclus dans cette rubrique les oublis importants, c'est-à-dire faisant l'impasse sur une information d'ordre sémantique. *Any deterioration of the denotative sense: erroneous information, nonsense, important omissions …*	−1
CONTRESENS (*misinterpretation*)	L'étudiant affirme le contraire de ce que dit le texte : information présentée de manière positive alors qu'elle est négative dans le texte, confusion entre l'auteur d'une action et celui qui la subit … *The student misinterprets what the source text says: information is presented in a positive light whereas it is negative in the source text, confusion between the person who acts and the one who undergoes the action…*	−2
VOCABULAIRE (*vocabulary*)	Choix lexical inadapté, collocation inusitée… *Unsuited lexical choice, use of non-idiomatic collocations*	−1
CALQUE (*calque*)	Utilisation d'une structure littéralement copiée et inusitée en français. *Cases of literal translation of structures, rendering the text un-French*	−1
REGISTRE (*register*)	Selon la nature du texte ou la nature d'un extrait (par exemple, un dialogue) : traduction trop (in)formelle, trop recherchée, trop simpliste… *Translation that is too (in)formal or simplistic and not corresponding to the nature of the text or extract*	−0,5
STYLE (*style*)	Lourdeurs, répétitions maladroites, assonances malheureuses… *Awkward tone, repetitions, unsuited assonances*	−0,5
GRAMMAIRE (*grammar*)	Erreurs grammaticales en français (par exemple, mauvais accord du participe passé, confusion masculin/féminin, accords fautifs…) + mauvaise compréhension de la grammaire du texte original (par exemple, un passé rendu par un présent…) et pour autant que ces erreurs ne modifient pas en profondeur le sens. *Grammatical errors in French (for example, wrong agreement of the past participle, gender confusion, wrong agreement of adjective and noun, …) + faulty comprehension of the grammar of the original text (for example, a past event rendered by a present tense…), provided that these errors do not modify the in-depth meaning of the text.*	−0,5
OUBLIS (*omissions*)	Voir SENS. *See Sense/meaning*	−1
AJOUTS (*additions*)	Ajout d'informations non contenues dans le texte (sont exclus de ce point les étoffements stylistiques). *Addition of information that is absent from the source text (stylistic additions are excluded from this category).*	−1

ORTHOGRAPHE (*spelling*)	Erreurs orthographiques, pour autant qu'elles ne modifient pas le sens. *Spelling errors, provided they do not modify the meaning of the text*	–0,5
PONCTUATION (*punctuation*)	Oubli ou utilisation fautive de la ponctuation. Attention : l'oubli, par exemple, d'une virgule induisant une compréhension différente du texte, est considéré comme une erreur de sens. *Omission or faulty use of punctuation. Caution: the omission of a comma leading to an interpretation that is different from the source text, is regarded as an error of meaning or sense.*	–0,5

Appendix 4: Translation assignment with calibrated items marked in bold.

Media **en amusement** trekken alle aandacht naar zich toe

Een van de grootste misvattingen die ik bij de meeste mensen heb vastgesteld **over reclame, is dat zij ervan uitgaan dat al het geïnvesteerde reclamegeld** van bedrijven terechtkomt in de 'reclamewereld', **waarmee zij** de wereld van reclamebureaus en reclamemakers **bedoelen**. Dat is niet zo. **Het overgrote deel van de reclame-investeringen** van het bedrijfsleven gaat naar de aankoop van ruimte in de media, en komt dus terecht **op de bankrekeningen** van de mediagroepen met hun tijdschriften, kranten, radiozenders, **tv-stations**, bioscopen, **billboards**... Bedrijven zijn immers op zoek naar een publiek om hun producten **bekend en geliefd te maken**, in de hoop dat publiek **ervan te kunnen overtuigen** hun producten ten minste eens te proberen. **Dat publiek** wordt **geleverd** door de media. De bedrijven kopen dus pagina's of zendtijd, en kunnen zo in contact treden met het publiek **van die media**. **Op die manier ontstaat** er een miljardenstroom van reclamegeld (in België meer dan 1,75 miljard euro per jaar), **die van de bedrijven** naar de **media** stroomt.

De reclamebureaus staan slechts aan de oevers daarvan. **Zij zijn de kleine vissers**, **hengelend naar een opdracht** van de bedrijven die er vooral **in bestaat de aangekochte ruimte** te vullen met inhoud. **De reclamebureaus zorgen dus voor** de ontwikkeling van de boodschap **van het merk** en voor de verpakking van die boodschap. In ruil daarvoor krijgen ze een percentage van de reclame-investeringen: vroeger **ging dat** om 15%, nu is het meestal minder dan 10%. Steeds meer **worden deze percentages afgeschaft en vindt de betaling plaats** via **een maandelijks honorarium**, **dat door de bedrijven zwaar onder druk wordt gezet**. **Zij moeten immers** steeds meer betalen voor hun mediaruimte en willen die meerkosten **zo veel mogelijk terugverdienen**, **onder meer** via de reclamebureaus. Deze worden verplicht **steeds sneller te werken voor steeds minder geld**. **Dat wil niet zeggen dat** de reclamebureaus armoedezaaiers zijn. Veel reclamemakers **verdienen goed**. **Door** mee te surfen op de golven van de reclame-investeringen zijn er **multinationale en beursgenoteerde reclamenetwerken** ontstaan. **Maar de échte reclamewereld,** waarin het grote reclamegeld **omgaat**, **is eigenlijk** de mediawereld.

Revisiting Carroll's scales

Elisabet Tiselius
Stockholm University

This pilot study describes the assessment of interpreting through the application of scales originally devised by Carroll (1966) for machine translation. Study participants (interpreters, n = 6; non-interpreters, n = 6) used Carroll's scales to grade interpreted renditions (n = 9) in simultaneous mode by conference interpreters with three different levels of experience. Grading was conducted using transcripts of the interpreted renditions. Although the numbers of graders and graded renditions were small, the data indicates that interpreters and laypeople agree on the grading of *intelligibility* and *informativeness* in interpreted renditions.

1. Introduction

Tiselius (2008) conducted a longitudinal study of expertise in simultaneous interpreting from English into Swedish considering both product and process. As the assessment of interpreter performance, or the "end product" of interpreting, was one of the principal areas of focus of the longitudinal study, a literature review was conducted to identify available valid and reliable assessment instruments (cf. Angelelli 2004, 2007; Moser 1995; Gile 2003). The aim was to identify an instrument that would allow for grading of interpreter performance by non-experts in interpreting, given that interpreters are often assessed by non-experts in the field (Angelelli 2007).

Carroll's scales (1966) were selected for their ease of implementation, and because they could be adapted in a context where lay people, or people who were not professional interpreters, acted as graders. However, further exploration was necessary to determine their appropriateness for grading interpreter performance, and using non-professionals as graders. The scales were developed to measure quality in machine translation. They measure the *intelligibility* and *informativeness* of the target text in relation to the source text. They have never been tested on a large scale for interpreting. Despite this, they were used by two interpreting researchers (Gerver 1971 and Anderson 1979) and they served as a basis for developing a certification test for court interpreters in the U.S. (FCICE)[1] (a certification test that has been challenged, cf. Clifford 2005). Rating scales constitute one of many instruments used to assess interpreting both in research and in schools (cf. Lee 2008), and Carroll's scales were the first instrument of this type to be used in interpreting research.

An advantage of applying Carroll's scales to interpreting performance is their non-componential potential. Most tools implemented as user-expectation surveys in simultaneous interpreting are structured as Gile proposed in 1983 (also mentioned in Gile 2003): that is, asking separate questions on different components, such as fluency, adequacy, and so forth. This has the obvious advantage of ease of use to measure the weight of different components in an overall assessment. In the context of the 2008 longitudinal study, however, where non-interpreters were to act as graders, it was deemed more appropriate to use a tool that measured performance from a holistic perspective. A study was therefore conducted to explore the applicability of Carroll's scales for holistic grading of interpreter performance, which this chapter describes. The study of the applicability of Carroll's scales for grading interpreter performance described below dealt strictly with simultaneous conference interpreting, and with the language combination English (C) into Swedish (A).[2]

1.1 Purpose and research questions

The purpose of this study was to investigate whether Carroll's scales are appropriate for assessing simultaneous conference interpreting performance at a holistic level, and whether they represent a potential, easy-to-use tool by non-professionals. The term "non-interpreter" in this context refers to laymen to interpreting who are otherwise educated individuals, i.e. individuals with a university degree or university students or individuals with an equivalent level of instruction. The study investigates the ratings of two groups of non-experts: a group of experienced interpreters and a group of laymen to interpreting.

There were two sub-sets of research questions that contributed to the overall purpose of the study:

[1] http://www.ncsconline.org/D_Research/Consort-interp/fcice_exam/index.htm

[2] C-language – the AIIC (International Association of Conference Interpreters) language classification of a language of which one has full understanding, Generally interpreters are not expected to interpret into a C language. A-language – mother tongue level according to the AIIC language classification. http://www.aiic.net/ViewPage.cfm?page_id=199#langclassif

(1) *Does the application of the scales to interpreter performance produce valid results?*
The first question concerned the validity of the scales for the assessment of interpreting: The scales were conceived for grading machine-translated texts. The field of interpreting research and the field of testing research have evolved since the scales were developed. Furthermore, the scales have been challenged by Anderson (1976), and Clifford (2005). Therefore, although the scales appeared to be potentially useful in an interpreting context, it was essential to determine whether an application of the scales would be appropriate for assessing the quality of interpreting. The first part of this study set out to determine whether an adapted version of the scales would be valid for measuring *intelligibility* and *informativeness* in interpreting. In this part of the study it was assumed that the renditions by very experienced interpreters who had acquired a high level of professional credentials, such as accreditation from European Institutions or membership in AIIC (the International Association for Conference Interpreters) would be graded higher than renditions by novice interpreters. If this proved to be the case then the scales would at least be considered to have face validity.

(2) *Can the scales be used by non-experts to assess interpreting?*
The second question concerned who should do the grading: Professional interpreters usually have some experience in assessing interpreting, and therefore can be assumed to be able to perform this task. Are non-interpreters also able to assess interpreting if they are given the same task, including the same training and education (outside of interpreter education)? Most people who use the services of interpreters are laypeople, and the assessment of the end-user would be expected to be relevant. Laypeople are regularly asked for their opinion of interpreting quality (e.g. Moser 1995 or SCIC customer survey 2008), but the way they grade and assess interpreting has not been studied. The aim of the second part of this study was to determine whether there were differences in grading between trained professional interpreters and laypeople using the scales.

2. Background

2.1 Carroll's scales

John B. Carroll was an American psychologist who developed language aptitude tests. Carroll conducted seminal research on developing useful assessment tools for language testing (Stansfield & Reed 2003). Machine translation was another of his research areas, and in 1966, he developed two scales for evaluating machine-translated texts (Carroll 1966). In his work, Carroll challenged the Discrete Point Theory in language testing (Stansfield & Reed 2003). The discrete point approach is an analytical approach to language testing, in which each test question is meant to measure one distinct content point. Carroll was in favor of using an integrative testing design, in which each question requires the test-taker to use more than one skill or piece of knowledge at a time, which he claimed may result in a more natural representation of the test-taker's knowledge of the language. This preference for an integrative testing design can also be seen in his argumentation of how to design a method for testing machine-translated text.

Although Carroll assumed that, "the evaluation of the adequacy of a translation must rest ultimately upon subjective judgments, that is, judgments resulting from human cognitions and intuitions" (1966: 55), he believed that if sufficient care was taken, it would be possible to obtain "acceptable levels of reliability and validity that yield satisfactory properties of the scale or scales on which measurements are reported" (Ibid.). One of the ways to achieve this was to "[provide] a collection of translation units that would be sufficiently heterogeneous in quality to minimize the degree to which the judgments on the evaluative scales would be affected by varying subjective standards" (Ibid.). Carroll drew up several more requirements to obtain an evaluation, and these led him to design his scales. The original scales are reproduced here under Section 3.1 (1966: 55–56). He established the need for two scales (based on two constructs: *intelligibility* and *informativeness*), as he claimed that a translation could be perfectly intelligible but lack fidelity to the original, while another text could be completely unintelligible and yet be completely faithful to the original. Neither of the two alternatives is, according to Carroll, considered a good translation (1966: 57).

When designing the scales, Carroll picked random sentences from one machine translation and one human translation, from Russian into English. He then sorted them into nine different groups for *intelligibility* and nine different groups for *informativeness*, depending on how intelligible or informative they were, compared to the original. He then elaborated definitions for nine different grades for each scale: these definitions are included in Tables 1 and 2, under heading 3.1. Then, using the scales, 18 students of English with high scores on the SAT (a standardized test for college admission in the United States) and 18 professional translators from Russian to English graded the translated sentences, as compared with the originals.

The scales have holistic qualities, since they were designed to grade output from a perspective of general understanding. The rendition is graded holistically and focus is placed on understanding the rendition, as well as on obtaining all of the information from the original.

2.2 Applying grading scales to interpreting

As mentioned above, Anderson (1979) and Gerver (1971) used Carroll's scales to assess interpreting. Both Anderson and Gerver had two graders grade interpreters' renditions using transcripts. Anderson used full transcripts (i.e. with "false starts, hesitations, repetitions and gropings [*sic*] [for words] left in," 1979: 27), while Gerver used transcripts without these details. Gerver did not provide any critical analysis of the application of the scales, but Anderson raised certain doubts about them. She did not obtain any significant treatment effects in her first two experiments, which made her question whether the scales were fine-tuned enough for measuring the output of interpreting. However, neither Anderson nor Gerver made any specific adaptations of the scales to interpreting, nor did they use them in a larger study.

Lee (2008) also conducted a study on grading scales (not Carroll's, but her own) for assessing interpreting, in which she draws the conclusion that they had good inter-rater reliability and that graders found them easy to use, but that further research was needed before any conclusions could be drawn from the results of her study. Lee used three analytical grading scales that she designed, and concluded that, "since interpreting performance assessment does not allow time for thorough analysis, graders have to judge performance quality based on the immediate interpreting of selected criteria. For these practical reasons, grading scales appear to be an appealing method for interpreting performance assessment" (2008: 170).

As stated before, Carroll's scales were developed for written translation. Admittedly, it may seem awkward to use an instrument developed for assessing written translation to assess interpreting. In order to apply them to interpreting the difference between interpreting and translation has to be clarified. Pöchhacker defined interpreting as "a form of translation in which a **first and final rendition in another language** is produced on the basis of a **one-time presentation** of an utterance in a source language" (2004: 11, bold in the original). Without going into any detailed definition of translation [for such a definition see for instance Toury (1980), Gutt (1991) or Pym (2004)], it can be pointed out that the key differences between translation and interpreting were in fact highlighted by Pöchhacker, above. The first rendition of a translation is, in most cases, not the final one. The translator may have several opportunities to revise the target text. The translator has, in most cases, access to an original text which can be consulted continuously. These differences have to be taken into account when applying the scales to interpreting.

In order to determine whether Carroll's constructs of *intelligibility* and *informativeness* are applicable to interpreting constructs, they were compared to two of the constructs mentioned by Shlesinger (1997). Carroll's term *intelligibility* is similar to Shlesinger's term *intratextual* (i.e. a product in its own right, that can be examined on its own), and Carroll's term *informativeness* corresponds to Shlesinger's *intertextual* (i.e. a comparison of the source text and the target text, where the examination is based on similarities and differences) (Shlesinger 1997: 128). These terms were chosen in this context since they focus more on interpreting product-as-text and not as activity. Shlesinger also took the communicative act of interpreting into account when suggesting the third term *instrumentally* which is based on the usefulness and comprehensibility of the target text, thereby encompassing some of the communicative aspects of interpreting. The two constructs compared here do not take all components of the interpreted communicative event into account (cf. Wadensjö 1999 and Angelelli 2004). For the present study, given the interest in identifying an effective, holistic approach to grading transcribed versions of simultaneous interpreting performance, it was not judged to be of crucial importance.

A possible problem when using the scales to evaluate interpreting, especially if graders do not evaluate a whole text but only smaller units, is that there is a risk of graders' attention being diverted from the fact that they are grading a communicative event. In addition to this, Carroll's scales do not deal with the speaker's possible responsibility for achieving communication with the addressee via the interpreter. A successfully interpreted event is not solely the responsibility of the interpreter, as Vuorikoski pointed out (2002). In the present study, it was assumed that Carroll's statement above (that a translation could be perfectly intelligible but lack fidelity to the original, while another text could be completely unintelligible and yet be completely faithful to the original and that neither of the two alternatives is generally considered a good translation) is valid for interpreting, as well. It should be pointed out that meaning in oral discourse is subject to co-construction (see for instance Wadensjö 2000), but because of the design of this study it was not addressed here. This is a weakness of the scales.

In addition, in a study of the validity of the FCICE test, Clifford (2005) found that the two constructs of *intelligibility* and *informativeness* correlated to such a high degree that there was reason to suspect that they were not separate constructs (2005: 122). Clifford did not expect this, and he concluded that "we may wish to revisit the theory and question its assumptions, but for the moment at least, the test does not do what it has set out to do" (ibid). For the purposes of this study it should be pointed out that the FCICE scales are not similar to Carroll's original scales. Furthermore, they are not applied in the same way as in Clifford's test. Therefore, it will continue be assumed for the purposes of this study that the two constructs are different. However, the correlation of the two scales will necessarily need to be investigated in the future.

In conclusion, when applied to simultaneous conference interpreting, Carroll's scales can be assumed to account for central aspects of the interpreted event but not for its entirety as a communicative event. Despite this and other objections raised in this section, the scales still seemed to serve the purpose of being an easily accessible, easy-to-use tool that can be implemented by laypeople in order to assess a transcribed version of a simultaneous interpreting performance. For these reasons, it was decided to investigate the applicability of the scales. The following section describes the study and how the scales were applied.

3. Data and method

In the present study nine interpreters with three different levels of experience (no experience, short experience and long experience) produced nine 10-minute renditions. Carroll's scales were adapted to interpreting. The nine renditions (eliciting material) were turned into transcripts, divided into smaller units, mixed randomly and graded following Carroll's scales by two groups of trained graders (interpreters and non-interpreters, n = 12). The results from the different groups of graders were compared to each other.

3.1 Adaptation of the scales

As already mentioned, Carroll's scales do not take features of spoken language into account. To remedy this, the scales were adapted to interpreting (i.e. to spoken language). Adaptation is used as the overall term of the process of changing the scales. The adaptation consisted of: (1) deleting scale steps and references to written text and translation; (2) adding references to spoken language and interpreting; (3) changing some formulations (see Tables 1 and 2).

First, references to spoken language (Swedish, in this case) and interpreting were added to the definitions, such as "like ordinary spoken Swedish." It was also considered whether terms such as fluent, coherent, and clear needed to be added to the scales, but it was decided that "ordinary spoken Swedish" would encompass fluency, coherence, and clarity. Therefore, no additional components were added.

Furthermore, as Cohen et al. (1996: 224) mentioned, in grading scales there may be several dimensions underlying the grading being made, meaning in this case that *intelligibility* can have the underlying dimensions of fluency, clarity, adequacy and so forth. If scales are multidimensional, more than one dimension is likely to influence the grader's response. Secondly, the number of grades was reduced to six, since a pilot study indicated that six grades were easier to handle

Table 1. Scale of *intelligibility*, adapted version and original (Carroll 1966: 58)

Original scale of intelligibility	Scale of intelligibility (as adapted in the present study)
9. Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.	6. The rendition is perfectly clear and intelligible. Like ordinary spoken Swedish with few if any stylistic infelicities.
8. Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected."	5. Generally clear and intelligible but with minor grammatical or stylistic peculiarities or unusual word choices, nothing that hampers the understanding.
7. Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.	–
6. The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in nearly acceptable form.	4. The general idea is intelligible, but full comprehension is interfered with by poor word choice, poor style, unusual words and incorrect grammar. The Addressee will have to make an effort to understand the utterance.
5. The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.	–
4. Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.	3. Masquerades as an intelligible utterance, but is actually more unintelligible than intelligible. Nevertheless, the idea can still be comprehended. Word choices, syntactic arrangements, and expressions are generally unusual and words crucial to understanding have been left out.
3. Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.	–
2. Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.	2. Almost completely unintelligible. Although it does not seem completely nonsensical and the Addressee may, with great effort, discern some meaning.
1. Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.	1. Totally unintelligible and completely without meaning.

Table 2. Scale of *informativeness*, adapted version and original (Carroll 1966: 58)

Original scale of informativeness	Scale of informativeness (as adapted in the present study)
9. Extremely informative. Makes "all the difference in the world" in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation.)	6. Reading the original changes the whole understood meaning. (6 should be given when reading the original completely changes the meaning that the rendition gave).
8. Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely.	5. Reading the original clarifies the understood meaning. The original's differences in syntax, words and phrases alter the listener's impression of the meaning to some extent.
7. (Between 6 and 8.)	–
6. Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader "on the right track" as to the meaning intended.	–
5. (Between 4 and 6.)	4. Reading the original gives some additional information about syntax and words. It can also clarify minor misunderstandings in the rendition.
4. In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.	–
3. By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different "twist" to the meaning conveyed by the translation. It adds no new information about sentence structure, however.	3. By correcting one or two meanings, mainly on word level, the reading of the original gives only a minor difference in meaning compared to the rendition.
2. No really new meaning is added by the original, either at the word level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended.	2. No new meaning is added through reading the original, neither at the word level nor at the grammatical level, but the Addressee is somewhat more confident that s/he really comprehends the meaning intended.
1. Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced.	1. No new meaning is added by the original, nor is the Addressee's understanding of the rendition increased.
0. The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.	0. The original contains less information than the rendition.

Table 3. Scale of *intelligibility* on grading sheet

1. Totally unintelligible	2. Generally unintelligible	3. Seems intelligible	4. General idea intelligible	5. Generally intelligible	6. Completely intelligible

Table 4. Scale of informativeness on grading sheet

0. Original contains less information than rendition.	1. Without any new information.	2. No new information, strenthens the intended meaning.	3. Minor changes in meaning.	4. Gives some new information.	5. Original explains and improves.	6. Only new information.

than nine, in a fairly quick grading of spoken language. This will, of course, limit the variability. However, for attitude verbal grading scales or verbal description scales (i.e. scales measuring a person's experience of something (in this case an interpreted rendition) by attributing to them a verbal description (here, for instance, "totally intelligible" or "totally unintelligible"), each grade has to have a meaningful description which becomes difficult above six or seven scale steps. It is also preferable that the scales do not have a middle value (Gunnarson 2002).

However, having adapted the scales as described above, it was estimated that they had a high componential element in them, and each step covered not only implicitly but also explicitly several aspects of interpreting performance, such as adequacy at syntax level or word level. Therefore, the graders were provided with shorter verbal descriptive scales, as in Tables 3 and 4, on each sheet of grading paper. The adapted scales in Tables 1 and 2 were used as background information when training the graders (see below), but the actual grading was performed with verbal descriptive scales, as in Tables 3 and 4.

It should also be stressed that the scale of *intelligibility* has *six* as the best score and *one* as the lowest, whereas the opposite is true for the scale of *informativeness*. For the scale of *informativeness*, *one* denotes the highest correspondence with the original and is thereby the highest score, while *six* denotes low correspondence with the original and is thereby the lowest score. Appendix 1 provides a Swedish version of the scales as presented to graders.

3.2 Eliciting material

3.2.1 *The speech*

The material used to elicit the samples for grading was based on a source text from the European Parliament. It was a ten-minute speech given in English at the European Parliament by Commissioner Byrne (Byrne 2002). The criteria for choice of speech were authenticity, general topic with little specialized terminology, and length. The speech was re-recorded with a different speaker, to reduce difficulties due to speed or accent. The speed in the original speech was an average of 141 words per minute (wpm), compared to 119 wpm in the re-recorded speech. Speeches in the European Parliament are published in a verbatim report immediately after the session. They are also recorded and can be obtained from the audio-visual services at the European Parliament. Official translations of the verbatim report are published at a later stage by the European Parliament on their website.[3]

3.2.2 *The interpreters*

Nine interpreters with three different levels of experience rendered the speech from English into Swedish. The interpreters were recruited at Stockholm University and at the European Parliament. The three different levels of experience were:

i. No experience; language students familiar with the principles of simultaneous interpreting but without any professional experience of interpreting.
ii. Short experience; interpreters with formal interpreter training at university level, but with only short professional experience (<2 years).
iii. Long experience; interpreters with formal interpreter training at university level, and long professional experience (more than 20 years).

Table 5 shows the age and experience of the interpreters. All of the trained interpreters had Swedish as their mother tongue. The trained interpreters had English as a C-language (the AIIC definition of a language of which one has full understanding, but into which does not generally interpret),[4] and the untrained interpreters studied English at the university level.

Table 5. Age and experience of the interpreters

	Age span	Years at university	Interpreting school	Years of experience
Group (i) n = 3 No experience	20–30	4	No	0
Group (ii) n = 3 Short experience	30–40	4	Yes	<2
Group (iii) n = 3 Long experience	50–60	4	Yes	>25

[3] http://www.europarl.europa.eu/activities/plenary/cre.do?language=SV#

[4] http://www.aiic.net/ViewPage.cfm?page_id=199#langclassif

3.2.3 *Preparing the transcripts*

Each of the nine ten-minute renditions was first carefully transcribed using the Childes software in .ca format (MacWhinney 1991), to mark pauses, pronunciation, and intonation, and then made into a written text by adding punctuation according to intonation. This means that, in the transcripts used for grading, all meta-textual markers of pauses, pronunciation, and intonation were omitted, leaving only traditional markers such as a full stop or a comma.

The text version of each rendition was then divided into 18 interpreting units. The division into units was based on the following: The graders in Carroll's original study (1966) worked with sentences, since he argued that the translations should be divided, "to be measured into small enough parts (translation units) so that a substantial number of relatively independent judgments could be obtained on any given translation, and so that the variance of measurement due to this kind of sampling could be ascertained" (1966: 55). In this context, however, it was considered that although sentences could be identified by following intonation patterns, interpreting is too complex an exercise to be evaluated at the sentence level (this can of course be argued for translation as well). Units of meaning (Lederer 1978: 330) or translation units (Gile 1995: 101) have been used to describe the pieces of utterance with which interpreters work. Gile (1995: 102) stated that a unit can be a single word, or a long sequence. He also emphasized that it is the interpreter who decides the contents and limits of the unit. The term interpreting unit will be used here, as described by Vik-Tuovinen (2002: 22). In deciding what was to be considered an interpreting unit, two criteria were taken into consideration: intonation and idea. The interpreter's intonation indicated the end of a unit, and ideas were kept together, as in this example of an interpreting unit (English original speech): *We have developed and proposed this directive, which we consider a qualitative step forward in protecting public health. This work has been done within the legal framework for completion of the internal market. The directive before you today will represent a significant improvement on our current legislative position and fill many of the gaps, which have made the current rules ineffective.* Each unit comprised 20 to 45 seconds of listening time. This process yielded a total of 162 interpreting units to be graded.

Each interpreting unit was then printed on a separate page, with the interpreted rendition at the top and the original at the bottom. The *intelligibility* scale (as in Table 3) was at the very top of each page; the *informativeness* scale (as in Table 4), at the very bottom. For an example of a grading sheet, see Appendix 2. In order to have each grader grade units from all nine renditions, the units were coded and then mixed randomly. Naturally, in all discourse the interpreting of one unit is dependent on the preceding unit. Yet, since ideas were kept together when dividing the speech into units, each unit was deemed sufficiently self-contained to be evaluated independently of the preceding and subsequent units, at least from the perspective of both *intelligibility* and *informativeness*. The units were not sorted in chronological order. Each interpreting unit was graded by two graders from the students' group, and two graders from the interpreters' group, which was also consistent with Carroll's assumption that "for each translation unit, obtain judgments from more than one grader so that the variance of measurement attributable to graders [can] be ascertained (Carroll 1966: 56)". Each set of units to be graded was made up of 54 units.

The interpreter graders were provided with the original, verbatim speech at the bottom of the page. The non-interpreter graders were provided with a Swedish translation of the source speech by the translation service at the European Parliament. The translation was provided given that non-interpreters were chosen for having Swedish as mother tongue, and not for their command of English. It could be argued that this interjects a further complication to the grading. The original speech is then already processed once, by a translator into a translation. However, the mere act of translating does not necessarily divert or change the information and meaning in an utterance *per se*. Furthermore, since the focus of this study was to test the grading scales and the graders' ability to use them, it was decided to use a translation, thereby avoiding yet another screening of graders.

The main reason for having the graders work with a transcribed speech was to prevent graders from recognizing the voices of the interpreters, some of whom are the graders' colleagues. The transcribed texts were also deemed as being sufficiently transparent for the purposes of this study.

3.3 The grading procedure

3.3.1 *The graders*

The graders in the study were native speakers of Swedish, divided into two groups. The first group consisted of university students (n = 6, 2 male and 4 female), who were not trained in interpreting and were thus similar to potential addressees/users of interpreting. They were recruited at Stockholm University. The second group consisted of simultaneous conference interpreters (n = 6, all women), each of whom had at least eight years of professional experience, including training and evaluating interpreters. Therefore, it was possible to assume that they were professional graders of interpreting. The second group of graders was recruited at the European Parliament.

3.3.2 *Grader training*

At the beginning of the grading session, the graders were trained for their tasks. For the students (non-interpreter) grader training and grading were carried out during class hours in their regular class rooms. Two grading sessions were held with three students at each session. For the interpreters, grader training and grading were conducted at their workplace, either during lunch break or after working-hours. Three interpreters participated in one session, and the other three interpreters had individual sessions.

Training consisted of introducing the scales as presented in Tables 2 and 3. Each scale step was run through and examples were given. After this introduction, three mock units were graded together with the test leader (the author of this chapter). At this point, graders had the possibility to ask for clarification of scale steps or grading. The introduction and training part took approximately ten minutes.

3.3.3 *Grading*

Immediately following the grader training session the graders were asked to perform their grading task. They graded individually, they were requested not to consult with anybody else while grading. Each grading session took approximately one hour.

The graders received a set of 54 interpreting units, with each page folded in such a way that they first read only the unit rendered into Swedish and graded it for *intelligibility*. Then the graders unfolded the sheet and compared the rendition in Swedish with the original English (interpreter graders) or the translation into Swedish (non-interpreter graders) and graded the rendition for *informativeness*, i.e. its correspondence to the original.

3.4 Measuring significant difference and inter-rater reliability

When the grading exercise was done, all the units were returned back to the original rendition and two average scores for each rendition were calculated, one score for the non-interpreter graders and one score for the interpreter graders. The *p*-values were calculated and the result was used to determine whether the average scores showed significant difference or not between the renditions by highly experienced versus the renditions by less experienced interpreters and the renditions by interpreters with no experience. Furthermore, *p*-values were calculated and used to determine possible significant difference in grading between non-interpreter graders and interpreter graders.

A small *p*-value is strong evidence against the null hypothesis, the null hypothesis being in this case no difference between scores obtained by the different groups of interpreters. A small p-value is then strong evidence for the fact that the differences observed in grading would be at least reproduced under the same conditions. The *p*-values in this study were obtained by using a two-tailed *t*-test with unequal variance: two-tailed to investigate whether there was a difference or not, without assessing that difference, and unequal variance because different groups were measured. The reason for using *p*-values in the comparison was to determine whether or not the observed differences in the raw data were statistically significant. The differences in grading between interpreter graders and non-interpreter graders were also compared using *p*-values (obtained with a *t*-test, as above), to determine whether there were significant differences between the groups of graders. A *p*-value below 0.05 ($p < 0.05$) indicates significant difference and a *p*-value above 0.05 ($p > 0.05$) indicates no significant difference. Some comparisons in the study yielded a *p*-value lower than 0.01 ($p < 0.01$), which provided an even stronger support for the claim of significant difference.

Inter-rater reliability was tested using the Pearson product-moment correlation coefficient *r* which measures pair-wise correlation among raters using a scale that is ordered. Perfect correlation gives a value of –1 or 1 and no correlation a value of 0.

4. Results

This section provides an overview of the results of the 12 graders scoring the nine renditions, using Carroll's scales to grade the *intelligibility* of an interpreted rendition and its *informativeness* in comparison with the original speech.

4.1 Inter-rater reliability

The inter-rater reliability test gave *r* 0.6 for interpreter graders grading *intelligibility* and *r* 0.65 for interpreter graders grading *informativeness*. Non-interpreter graders grading *intelligibility* gave *r* 0.3, and non-interpreter graders grading *informativeness* gave *r* 0.5.

4.2 *Intelligibility*

Table 6 gives the *p*-values for the significance of the scores for *intelligibility* between the different renditions: long experience, short experience, and no experience, as graded by non-interpreter graders. The average score for each rendition is given together with the rendition.

Table 7 gives the *p*-values for the significance of the scores for *intelligibility* between the different renditions: long experience, short experience, and no experience, as graded by interpreter graders. The average score for each rendition is given together with the rendition.

As expected, graders gave higher scores to renditions by more experienced interpreters. In the non-interpreter graders' scores the difference is statistically significant for the grading of the renditions by long-experience interpreters versus the grading of the renditions by no-experience interpreters. The same is true for the non-interpreter graders scoring renditions by short-experience interpreters versus those of no-experience interpreters. Non-interpreter graders' scores show no significant difference for the renditions by long- and short-experience interpreters. The interpreter graders' scores also show significant difference in the grading of the renditions by the long-experience interpreters versus the renditions by the non-experienced interpreters. The interpreter graders' scores also show significant difference for the renditions of short-experience interpreters versus the renditions produced by non-experienced interpreters. There is no significant difference in grading of the renditions by long- and short-experience interpreters graded by interpreter graders.

Table 6. Significance in gradings of *intelligibility* by non-interpreters (n = 6)

Renditions	**Intelligibility**		
	No-experience 3.79	**Short-experience 5.25**	**Long-experience 5.42**
No-experience 3.79	–	0.001**	0.001**
Short-experience 5.25	0.001**	–	0.1
Long-experience 5.42	0.001**	0.1	–

**p < 0.01

Table 7. Significance in gradings of *intelligibility* by interpreters (n = 6)

Renditions	**Intelligibility**		
	No-experience 3.16	**Short-experience 4.88**	**Long-experience 5.11**
No-experience 3.16	–	0.001**	0.001**
Short-experience 4.88	0.001**	–	0.1
Long-experience 5.11	0.001**	0.1	–

**p < 0.01

4.2.1 Intelligibility *graded by non-interpreter graders vs. interpreter graders*

Table 8 shows the average scores of *intelligibility* for all nine renditions, as graded by interpreter graders and non-interpreter graders. It also shows the *p*-values for the significance in grading between interpreters and non-interpreters.

The *p*-values for the significance of the difference in grading by non-interpreters and interpreters are given for each experience level. As can be seen in Table 8, there is a significant difference in grading between non-interpreter graders and interpreter graders for the renditions produced by long-experience and no-experience interpreters. The raw data in Table 8 might indicate that interpreter graders were somewhat more severe in their grading, and this conclusion is supported by the significance. The difference in the grading of the renditions by the short-experience interpreters is not significant.

Figure 1 shows that the two groups of graders vary in the same way, although they differ slightly.

Table 8. Average scores of *intelligibility* for all nine renditions graded by non-interpreters (n = 6) and interpreters (n = 6)

Renditions	Intelligibility		Significance
	Non-interpreter graders	Interpreter graders	
No-experience	3.79	3.16	0.018*
Short-experience	5.25	4.88	0.078
Long-experience	5.42	5.11	0.015*

*p<0.05

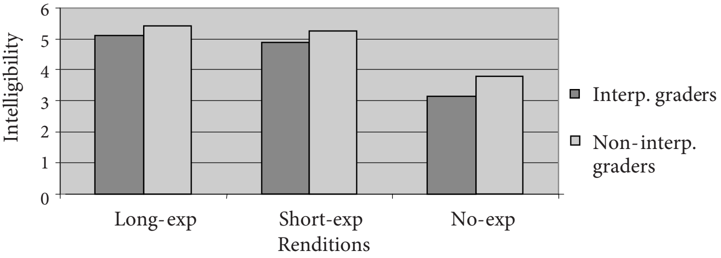


Figure 1. Average scores for *intelligibility* graded by interpreters (n = 6) and non-interpreters (n = 6)

Table 9. Significance for grading of *informativeness* by non-interpreters (n = 6)

Renditions		Informativeness		
		No-experience 4.42	**Short-experience 3.15**	**Long-experience 2.31**
No-experience	4.42	–	0.001**	0.001**
Short-experience	3.15	0.001**	–	0.001**
Long-experience	2.31	0.001**	0.001**	–

**p < 0.01. Note. The Lower the Score the Better the Performance

Table 10. Significance for grading of *informativeness* by interpreters (n = 6)

Renditions		Informativeness		
		No-experience 5.13	**Short-experience 3.42**	**Long-experience 2.60**
No-experience	5.13	–	0.001**	0.001**
Short-experience	3.42	0.001**	–	0.001**
Long-experience	2.60	0.001**	0.001**	–

**p < 0.01. Note. The Lower the Score the Better the Performance

4.3 *Informativeness*

Table 9 shows the *p*-values for the significance in grading of *informativeness* between the different renditions: long experience, short experience, and no experience, as graded by non-interpreter graders. The average score of *informativeness* for each rendition is given in the corresponding heading.

Table 10 shows the *p*-values for the significance of the grading of the different renditions: long experience, short experience, and no experience, as graded by interpreter graders. The average score of *informativeness* for each rendition is given in the corresponding heading.

The graders' scores, both for non-interpreter graders and interpreter graders, show a significant difference in the scores attributed to the renditions by long-experience interpreters vs. short-experience interpreters and to renditions by short-experience interpreters vs. no-experience interpreters. The raw data, supported by the significance, once again indicate that years of experience were consistent with better (lower) scores for *informativeness*, a sign of a perception of better rendition among these graders.

4.3.1 Informativeness *graded by non-interpreter graders vs. interpreter graders*

Table 11 shows the scores for *informativeness*, the rendition's correspondence to the original, as graded by interpreters and non-interpreters. The values are average

Table 11. Significance of grading of *informativeness* graded by non-interpreters (n = 6) and interpreters (n = 6)

Renditions	Informativeness Non-interpreter graders	 Interpreter graders	Significance in gradings between non-interpreters and interpreters
No-experience	4.42	5.13	0.001**
Short-experience	3.15	3.42	0.20
Long-experience	2.31	2.60	0.17

**p < 0.01. Note. The Lower the Score the Higher the Correspondence

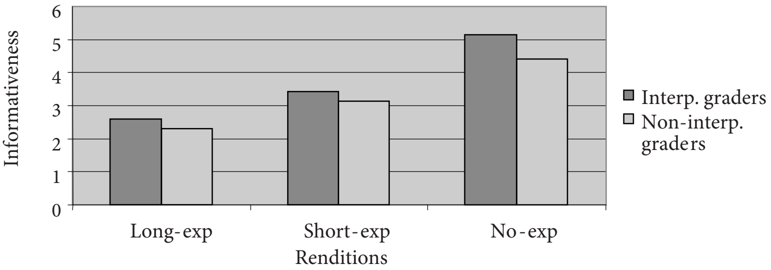


Figure 2. Average scores for *informativeness* graded by interpreters (n = 6) and non-interpreters (n = 6)

scores for all nine renditions. It also shows the p-values for the significant differences in grading between interpreters and non-interpreters.

There is no significant difference in the grading of renditions by short- and long-experience interpreters. The p-values for both groups are well over 0.05. Data again support the observations stated above, i.e. that non-interpreter graders may be more generous than interpreter graders. There is a significant difference between the two groups in the grading of renditions by the no-experience interpreters, again supporting the assumption that non-interpreter graders were more generous in their grading.

Figure 2 shows how the two groups of graders share the same tendencies. Although not in total agreement, they vary in the same way.

4.4 Spontaneous comments from graders

After each grading session, some of the graders were interviewed (informally) on their impressions of the grading. In general, graders found the scales easy to use and had no problem grading. Some graders (3) expressed a certain "grading-fatigue" towards the end of the grading.

5. Discussion

The study presented in this chapter investigated whether Carroll's scales could be applied to assess the performance of simultaneous conference interpreters. Furthermore, it investigated whether it was possible for graders who are not interpreting professionals to use the scales. This section discusses the limitations to this empirical research, as well as the results. Areas for future research are suggested.

5.1 Limitations

There are a number of limitations to this study that should be mentioned. First, the size of this study limits the possibility of drawing conclusions that can be generalized. This investigation was exploratory in nature, so caution must be taken in interpreting the results. It is clear that all graders gave higher scores to the renditions by experienced interpreters (the scores of interpreter graders and non-interpreter graders corresponded), and that interpreter graders were more severe in their assessment than non-interpreter graders. However, given that the number of graders is so low, it is uncertain whether this tendency would hold in a larger sample, and it is not possible to speculate as to the reason for it.

Secondly, the way the study was conducted takes the whole interpreted communicative event out of its context, in the following two ways:

A. Interpreters did not interpret for a live audience and did not have a live speaker from which to interpret. This takes the interpreter out of his or her context and may influence the rendition.
B. The graders were not allowed to listen to one interpreter for the whole speech, thereby creating an altogether new interpreted communicative event. The renditions were divided up into units; in addition, the graders graded from transcripts.

Furthermore, as mentioned above under Section 2.2, not all aspects of the interpreted communicative event were taken into account. However, the justification for this artificial design was that it would allow for a focus on the ability of graders to grade and on the validity of the grading scales, which was deemed appropriate for this context.

Thirdly, in order to test the grading scales, an alternative design would have been to manipulate the renditions on the grading sheets so that the grading samples contained interpreting units potentially representing all scale steps, and thereby test whether one specific interpreting unit was graded according to its assumed scale steps. Since the study used authentic renditions, the assumption was that the fact of using interpreters varying from very experienced to completely inexperienced would produce interpreting units representative of all the scale steps. In the study it could also be observed that graders made use of all the scale steps. See also the quotation of Carroll about providing sufficiently heterogeneous material: "[provide] a collection of translation units that would be sufficiently heterogeneous in quality to minimize the degree to which the judgments on the evaluative scales would be affected by varying subjective standards" (1966: 55).

5.2 Discussion of the results

Grading with the scales gave unambiguous results regardless of the graders' experience. All graders performed in line with the initial assumption that renditions by very experienced interpreters who had also acquired a high level of professional competence such as accreditation at the European Institutions or membership in AIIC would be graded higher than renditions by novice interpreters or laypeople to interpreting. This result provides some support for the validity of the grading scales since they were designed with renditions and interpreting units that were assumed to differ (experienced interpreters score better than inexperienced interpreters) and the scales reflected that difference. However, the correspondence of the scores from different groups of graders may also be due to possible flaws in the scales or the constructs. Thus further studies will have to be done, for instance, studying the correlations of the constructs, as Clifford (2005) did in his research. Furthermore, years of experience are not the only factor in predicting interpreting quality. Both the long- and short-experience renditions are based on a convenience sample (i.e. not necessarily a sample that is an accurate representation of a larger group or population). Therefore, it is quite possible that scores could vary within the sample, i.e. that one participant might perform much better or worse than the others. The results indicate that, in this study, years of experience are consistent with better scores within all grader groups and in all grading. To draw any major conclusions on years of experience and the possibility of predicting higher scores on that basis, a larger sample of renditions would have to be studied.

The inter-rater reliability is stable for both groups. The correlation is higher for interpreter graders, which may be due to the fact that they have a similar background. However, there is a sufficient correlation for non-interpreter graders when grading *informativeness*.

While these scales could be valid as an instrument for grading different aspects of interpreting quality, a larger sample needs to be studied. It is, however, important to note that the scales in this study proved easy to use, partly due to the fact that both training of the graders and sorting of the results are straightforward.

The only type and mode of interpreting tested here was technically aided simultaneous conference interpreting. It is possible that these grading scales could also be applicable to other types of interpreting, including consecutive. However, the way the scales are used in this study does not allow for a real-life evaluation, which can, of course, constitute a drawback. Furthermore, this study only used transcripts as the basis for grading: it would also be interesting to compare the results of this study to grading made from sound files.

Although drawing conclusions from this limited study is premature, some tentative ideas emerge from the research. An explanation for interpreters being slightly more severe in their grading may be their education and experience. Even interpreters who are not trained, such as teachers or examination jurors, are taught to evaluate themselves and their colleagues as part of their education. It is naturally a responsibility of the interpreter to make sure that as much information as possible is conveyed from the speaker to the addressee. The addressee has little or no possibility to check the *informativeness* or correspondence between the original and the rendition. But, when given the possibility, as in these tests, we can conclude that the same features of the interpreting performance seem to be important to non-interpreters and interpreters alike. An interesting twist is that this result contradicts Gile (1999) who found that interpreters are more lenient in their assessment of fidelity in interpreting than other graders, especially when grading transcripts.

Since the tendencies are similar between interpreter graders and non-interpreter graders, it would be feasible to use non-interpreter graders to grade renditions, at least in certain contexts. This study suggests that grading interpreter performance as part of studying their development over time, or the difference between different groups of interpreters in a research context, can be achieved with non-interpreter graders.

Finally, the fact that each rendition in the design of the study was divided into small units and randomly mixed enabled each rendition to be graded by many different graders in a fairly easy and straightforward manner. Having each grader grade nine renditions would be much more time-consuming, and definitely create "grader-fatigue." If given a whole rendition to grade, there is the risk of an inexperienced grader being misled by single features in one rendition, e.g. grading a whole performance highly because towards the end of the performance it gave a good impression. It would be interesting, in future studies, to compare the results of grading of a whole speech, using the same tool, to the results here. Furthermore, the fact that the renditions were divided into smaller units and the fact that each grader graded units from different renditions also diminished that risk. Another benefit of this type of non-componential, verbal descriptive scale was that graders found the scales, at least in this case, fairly easy to understand. Graders also found it easy to relate to the task.

5.3 Conclusion

For a project on expertise in interpreting, an instrument was needed for the assessment of interpreter performance where the assessment could be conducted by non-experts in interpreting. The reason for this was to avoid bias if the researcher was either to grade the performance of her colleagues herself, or ask other interpreter colleagues to perform such a task. Some support is found in the results of the present study to continue using this instrument.

It is beyond the scope of this study to speculate whether these scales can be used in other contexts, but the hope is that the study described here will enable other researchers to replicate this study with a greater number of subjects.

References

Anderson, Linda. 1979. *Simultaneous Interpretation: Contextual and Translation Aspects*. (psychology). Unpublished Masters Thesis. Concordia University.

Anderson, Linda. 1994. "Simultaneous Interpretation: Contextual and Translation Aspects." In *Bridging the Gap: Empirical Research in Simultaneous Interpretation*. Sylvie Lambert and Barbara Moser-Mercer (eds), 101–120. Amsterdam: John Benjamins.

Angelelli, Claudia V. 2004. *Revisiting the Interpreter's Role: A Study of Conference, Court and Medical Interpreters in Canada, Mexico, and the United States.* Amsterdam/Philadelphia: John Benjamins.

Angelelli, Claudia V. 2007. "Assessing Medical Interpreters. The Language and Interpreting Testing Project." *The Translator* 13 (1): 63–82.

Byrne, Padraig. 2002. *Debate in the European Parliament on the Tobacco Products Directive*. European Union.

Carroll, John, B. 1966. "An Experiment in Evaluating the Quality of Translations." *Mechanical Translations and Computational Linguistics* 9 (3–4): 55–66.

Clifford, Andrew. 2005. "Putting the Exam to the Test: Psychometric Validation and Interpreter Certification." *Interpreting* 7 (1): 97–131.

Cohen, Ronald Jay, Swerdlik, Mark, E., Phillips, Suzanne, M. 1996. *Psychological Testing and Assessment: An Introduction to Tests and Measurement.* Mountain View, CA: Mayfield.

Gerver, David. 1971. *Simultaneous Interpretation and Human Information Processing.* (psychology). Unpublished Doctoral Dissertation, Oxford University.

Gile, Daniel. 1995. *Basic Concepts and Models for Interpreter and Translator Training.* Amsterdam: John Benjamins.

Gile, Daniel. 1999. "Variability in the Perception of Fidelity in Simultaneous Interpretation." *Hermes—Journal of Linguistics* 22: 51–79.

Gile, Daniel. 2003. "Quality Assessment in Conference Interpreting: Methodological Issues." In *La Evaluación de la Calidad en Interpretación: Investigación,* Angela Collados Aìs, Manuela Fernández, Angela Sánchez and Daniel Gile (eds), 109–123. Granada: Editorial Comares.

Gunnarson, Ronny. *Skattningsskalornas Statistik*. Department of Primary Health Care, Göteborg University, Research methodology web site, http://infovoice.se/fou. (3/25/2009).

Gutt, Ernst-August. 1991. *Translation and Relevance. Cognition and Context*. Oxford: Basil Blackwell.

Lederer, Marianne. 1978. "Simultaneous Interpretation: Units of Meaning and Other Features." In *Language Interpretation and Communication*, David Gerver and H. Wallace Sinaiko (eds), 323–333. New York/London: Plenum Press.

Lee, Jieun. 2008. "Rating Scales for Interpreting Performance." *The Interpreter and Translator Trainer* 2:2: 165–184.

MacWhinney, Brian. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, NJ: Lawrence Erlbaum.

Moser, Peter. 1995. *Survey on Expectations of Users of Conference Interpretation*. Genève: AIIC.

Pym, Anthony. 2004. "Propositions on Cross-Cultural Communication and Translation." *Target* 16(1): 1–28.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London & New York: Routledge.

SCIC. *Customer Satisfaction Survey.* European Commission, Directorate General for Interpretation, http://scic.ec.europa.eu/europa/upload/docs/application/pdf/2008–02/scicnet-customer-satisfaction-survey-results-nov2007.pdf.

Shlesinger, Miriam. 1997. "Quality in Simultaneous Interpreting." In *Conference Interpreting: Current Trends in Research*, Yves Gambier, Daniel Gile, and Christopher Taylor (eds), 123–131. Amsterdam/Philadelphia: John Benjamins.

Stansfield, Charles W. and Reed, Daniel J. 2003. "The Story Behind the Modern Language Aptitude Test: An Interview with John B. Carroll (1916–2003)." *Language Assessment Quarterly* 1 (1): 43–56.

Tiselius, Elisabet. 2008. "Exploring Different Methods for Studying Expertise." In *Proceedings of the 49th Annual Conference of the American Translators Association,* Nicholas Hartmann (ed), 119–148. Alexandria: ATA.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam/Philadelphia: John Benjamins.

Wadensjö, Cecilia. 1999. *Interpreting as Interaction.* London: Longman.

Wadensjö, Cecilia. 2000. "Co-constructing Yeltsin – Explorations of an Interpreter-Mediated Political Interview." In *Intercultural Faultlines,* Maeve Olohan. (ed.), 233–252. Manchester: St Jerome.

Vik-Tuovinen, Gun-Viol. 2006. *Tolkning på Olika Nivåer av Professionalitet*. Vasa: Acta Wasaenisia.

Vuorikoski, Anna-Riitta. 2004. *A Voice of its Citizens or a Modern Tower of Babel?: The Quality of Interpreting as a Function of Political Rhetoric in the European Parliament*. Tampere: Tampere University Press.

Appendix 1. Carroll's scales in Swedish

Skala för förståelse (*Intelligibility*)	Skala för informativitet (*Informativeness*)
6. Tolkningen är helt tydlig och förståelig. Som vanlig talad svenska, inga eller mycket små stilistiska svagheter. *The rendition is perfectly clear and intelligible. Like ordinary spoken Swedish with few if any stylistic infelicities.*	6. Att läsa originalet förändrar hela den avsedda betydelsen. (6 ska ges när läsning av originalet totalt förändrar den förståelse som tolkningen gav). *Reading the original changes the whole understood meaning. (6 should be given when reading the original completely changes the meaning that the rendition gave).*
5. I stort tydlig och förståelig men med smärre grammatiska eller stilistiska egenheter eller annorlunda ordval, dock ingenting som hindrar förståelsen. *Generally clear and intelligible but with minor grammatical or stylistic peculiarities or unusual word choices, nothing that hampers the understanding.*	5. Att läsa originalet förtydligar den förstådda meningen. Genom förändringar i meningsbyggnad, ord och fraser ändrar originalet i viss mån lyssnarens intryck. *Reading the original clarifies the understood meaning. The original's differences in syntax, words and phrases alter the listener's impression of the meaning to some extent.*
4. Huvudtanken är förståelig, men den totala förståelsen hindras av dåligt ordval, stilistiska svagheter, underliga ord eller uttryck och grammatiska felaktigheter. Lyssnaren får anstränga sig för att förstå meningen. *The general idea is intelligible, but full comprehension is interfered with by poor word choice, poor style, unusual words and incorrect grammar. The Addressee will have to make an effort to understand the utterance.*	4. Att läsa originalet ger ytterligare information om meningsbyggnad och ord. Det kan också förtydliga mindre missförstånd i tolkningen. *Reading the original gives some additional information about syntax and words. It can also clarify minor misunderstandings in the rendition.*
3. Verkar vara en förståelig mening men är i själva verket mer oförståelig än förståelig. Huvudtanken kan kanske ändå urskiljas. Ordval, syntax och uttryck är ovanliga och ord som är avgörande för förståelsen kan ha utelämnats. *Masquerades as an intelligible utterance, but is actually more unintelligible than intelligible. Nevertheless, the idea can still be comprehended. Word choices, syntactic arrangements, and expressions are generally unusual and words crucial to understanding have been left out.*	3. Genom att rätta en eller två meningar framför allt på ordnivå ger läsningen av originalet en liten skillnad av betydelsen i tolkningen. *By correcting one or two meanings, mainly on word level, the reading of the original gives only a minor difference in meaning compared to the rendition.*

2. I princip helt oförståeligt. Verkar dock inte helt osammanhängande och lyssnaren kan möjligen urskilja någon betydelse med stor ansträngning. *Almost completely unintelligible. Although it does not seem completely nonsensical and the Addressee may, with great effort, discern some meaning.*	2. Ingen ny betydelse läggs till genom att läsa originalet vaken på ord nivå eller grammatiskt, men lyssnaren känner sig säkrare på att han eller hon verkligen förstått den avsedda betydelsen. *No new meaning is added through reading the original, neither at the word level nor at the grammatical level, but the Addressee is somewhat more confident that s/he really comprehends the meaning intended.*
1. Helt oförståeligt och helt utan mening. *Totally unintelligible and completely without meaning.*	1. Ingen ny betydelse har lagts till och lyssnarens förståelse av tolkningen har inte ökat. *No new meaning is added by the original, nor is the Addressee's understanding of the rendition increased.*
	0. Originalet innehåller om möjligt mindre information än tolkningen. *The original contains less information than the rendition.*

Appendix 2. Example of grading sheet

Skala för förståelse (*Intelligibility*)

1	**2**	**3**	**4**	**5**	**6**
Helt oförståeligt *Totally unintelligible*	I princip oförståeligt *Generally unintelligible*	Verkar förståeligt *Seems intelligible*	Huvudtanken förståelig *General idea intelligible*	I stort förståeligt *Generally intelligible*	Fullt förståeligt *Completely intelligible*

Vi har tagit fram och föreslagit detta direktiv som vi anser verkligen är ett kvalitativt steg framåt för att skydda folkhälsan och det här arbetet har gjort inom den juridiska ramen för att då färdigställa den inre marknaden och det direktiv som ni har framför er idag kommer att utgöra en klar förbättring när det gäller lagstiftningen och fylla i många luckor som har gjort att de nuvarande reglerna visat sig ineffektiva.

(Gloss rendition: We have developed and proposed this directive, which we consider really is a qualitative step forward in order to protecting public health and this work was done within the legal framework to then complete the internal market and the directive that you have before you today will make a clear improvement when it comes to the legislation and fill many gaps, which have made that the current rules have proven ineffective.)

Vi har utvecklat och föreslagit detta direktiv, som vi anser vara ett kvalitativt steg framåt för att skydda folkhälsan. Detta arbete har gjorts inom gränserna för den rättsliga grunden för den inre marknadens fullbordande. Det direktiv som ni har framför er i dag kommer att innebära en betydande förbättring av vår nuvarande lagstiftning och fylla många av de luckor som har gjort de nuvarande bestämmelserna ineffektiva.

(Verbatim original speech: We have developed and proposed this directive, which we consider a qualitative step forward in protecting public health. This work has been done within the legal framework for completion of the internal market. The directive before you today will represent a significant improvement on our current legislative position and fill many of the gaps, which have made the current rules ineffective.)

Skala för informativitet (*Informativeness*)

0	**1**	**2**	**3**	**4**	**5**	**6**
originalet innehåller mindre information än tolkningen *Original contains less information than rendition.*	Utan någon ny information *Without any new information.*	Ingen ny information stärker avsedd betydelse *No new information, strenthens the intended meaning.*	Lite förändring i betydelsen *Minor changes in meaning.*	Ger viss ny information *Gives some new information.*	Originalet förklarar och förbättrar *Original explains and improves.*	Enbart ny information *Only new information.*

Meaning-oriented assessment of translations

SFL and its application to formative assessment

Mira Kim
Macquarie University

One of the critical issues in the field of translation assessment is a lack of systematic criteria that can be used universally to assess translations. This presents an enormous challenge to translation teachers, who need to assess students' translations and provide constructive, detailed feedback. This chapter discusses how meaning-oriented translation assessment criteria have been used to address the challenges in teaching English to Korean translation over several years at Macquarie University. The meaning-oriented criteria have been devised using a text analysis approach based on systemic functional linguistics (SFL). The pedagogical effectiveness of such an assessment tool will also be discussed drawing on both qualitative and quantitative data.

Introduction

The area of translation assessment has been under-researched (Cao 1996: 525; Hatim and Mason 1997: 197) and regarded as a problematic area (Bassnett-McGuire 1991; Malmkjaer 1998; Snell-Hornby 1992) primarily due to "its subjective nature" (Bowker 2000: 183). As a consequence, there appears to be a lack of systematic criteria that can be used universally to assess translations (Bassnett-McGuire 1997; Hönig 1998; Sager 1989). This presents an enormous challenge to translation teachers, who need to assess students' translations for both formative and summative purposes, and provide constructive, detailed feedback on their translations. This chapter discusses how text analysis based on systemic functional linguistics (SFL) has been used over several years to address the challenge in teaching and assessing English-Korean translation at Macquarie University.

In earlier research, I explored the possibility of analyzing students' translation errors by categorizing them into different modes of meaning using SFL-based text analysis (Kim 2003, 2007a). The study not only showed that it is feasible to distribute translation errors into different categories of meaning; it also indicated that there are a number of potential pedagogical benefits of using text analysis as a tool in teaching translation. Since then, I have used text analysis as the main tool to discuss translation issues in class, give feedback on students' translations, and assess translation examinations. This chapter, therefore, can be regarded as a follow-up report on the use of text analysis as a formative assessment tool and on its pedagogical efficacy.

Although text analysis has been used in all levels of translation classes from introductory to advanced, the discussion in this chapter will specifically focus on how it has been used as a formative tool in assessing the components of a second semester translation course entitled *Translation Practice*, through which students can be accredited as professional translators in Australia. The reason for choosing the course as the focus for the discussion is that the pedagogical efficacy of using meaning-oriented translation assessment criteria developed on the basis of using text analysis for a formative assessment tool has become evident through quantitative data (that is students' performance in the end-of-year translation exams) and qualitative data (students' learning journals). The following section provides background information on the circumstances in which this research was conducted. The meaning-oriented assessment criteria are then presented, following a brief account of underlying theories of the criteria. The pedagogical efficacy of using the criteria is then discussed on the basis of survey results and students' learning journals, where applicable, as well as on the basis of data on students' performance in the end-of-semester translation examinations over the last five years.

Background

Macquarie University offers a suite of graduate programs in Translation and Interpreting (T&I). They include what is referred to as "the Postgraduate Diploma" (1 year full-time) and a Master program in Translation and Interpreting (1.5 year full-time) programs, which are accredited by the National Authority of Accreditation for Translators and Interpreters (NAATI). NAATI accreditation is the minimum industry standard required to practice as a translator or interpreter in Australia.

Students enrolled in the T&I programs at Macquarie University become NAATI-accredited translators[1] if they meet certain requirements. One of these requirements is that students must pass the end-of-semester paper-based

[1] They can be also be accredited as interpreters but, as this chapter is concerned with translation assessment, the discussion is limited to the requirements for translator accreditation.

translation examination of *Translation Practice*, the second-semester translation practice course. The exam is required to be carried out in the same way the NAATI accreditation exam is administered, including the same number of texts (two to be completed out of a choice of three), subject areas (social, economic, health etc), length (each about 250 words), exam time (2.5 hours plus 20 minutes for reading), and assessment method and criteria.

One of the challenges faced by teachers of translation courses is translation assessment. As the end-of-semester translation examinations of *Translation Practice* are required to be graded according to the NAATI assessment criteria, most teachers of the course use the NAATI criteria as a formative assessment tool to make sure that they use consistent criteria for both assignments and examinations. However, I found it difficult to use the NAATI translation assessment criteria for formative as well as summative purposes. Reasons for the difficulty will be explained in detail in the next section.

NAATI assessment criteria for translation tests

As mentioned earlier, most students enrolled in the second-semester translation practice course wish to achieve NAATI accreditation. Therefore the translation course is designed to develop knowledge and skills in translating short texts (250 words) in non-specialized areas. One of the routine activities in the course is that students translate a short passage as homework and hand it in, and the teacher gives feedback on the translations following NAATI examination grading guidelines.

For the last 30 years NAATI has adopted an error deduction method of translation assessment. Deductions are made from a maximum of 45 points for each text. Deductions of between 0.5 and 5 points per error are made depending on the level of "seriousness". The decision as to the seriousness of an error is left to the grader, as stated on the NAATI website (http://www.naati.com.au/at-deduction.html). Also the grader can deduct up to 5 points based on his or her overall impression. The NAATI assessment criteria are based on errors in the categories presented below:

a. Too free a translation in some segments
b. Too literal a translation in some segments
c. Spelling
d. Grammar
e. Syntax
f. Punctuation

g. Failure to finish a passage
h. Unjustifiable omissions
i. Mistranslations
j. Non-idiomatic usage
k. Insufficient understanding of the ethics of the profession[2]

(NAATI Translation Handbook: A Guide for Test Preparation 1997: 14)

The usefulness of the criteria as a formative assessment tool is significantly limited. First of all, some of the descriptors used, such as "too free", "too literal", "unjustifiable omission" and "mistranslation", are too general and there are no detailed guidelines to define the criteria. When the teacher uses such general criteria and is not able to explain when and why a translation is "too free" or "too literal," or when and why an omission is justifiable or unjustifiable, students tend to be reluctant to make their own translation choices. Instead they try to adhere as closely as possible to the source text (ST) structures for fear of losing a point by adding or omitting anything, which often leads to a translation that is too literal and that does not function as a natural text in the target language.

In addition, the NAATI criteria seem to focus too much on one aspect of meaning, which is *experiential* (e.g. who does what to whom, why, when and how) at the word or sentence level, but do not include other necessary categories related to whether a translation is accurate and natural in delivering other aspects of meaning, such as *interpersonal* meaning (e.g. formality or personal attitude) and *textual* meaning (e.g. coherent flow of information). Due to the lack of such categories, the teacher also tends to focus on lexical and syntactic errors. For instance, the criteria do not allow the teacher to effectively evaluate a translation that is correct in terms of "syntax", "grammar", and "idiomatic expression" at the clause or sentence level, but does not read well at the text level. This limited view of translation and meaning contradicts the current research on translation studies that supports the importance of creativity of the translator to produce a translation according to the translation brief, such as skopos theory and functionalism (cf. Reiss & Vermeer 1984; Nord 1997; Hönig 1998). The contradiction epitomizes the gap between theory and practice, and students are often puzzled as to how to apply theory in the actual process of producing a translation.

The lack of guidelines for deducting points (from 0.5 to 5 points) is another source of difficulty in using the criteria as a formative assessment tool, although there are general instructions stating that the criteria related to the quality of

[2] The issues related to ethics are not an immediate concern of this chapter.

naturalness are less serious than those related to the quality of accuracy.[3] When the teacher is applying the criteria, there are many occasions when errors at the word level should be treated differently due to different levels of "seriousness". There are also instances when the teacher wants to deduct or add some points based on overall impression, which is allowed according to the NAATI guidelines. However, unless evidence can be provided that supports an overall impression systematically, there is a risk of being completely subjective. Given this subjectivity, it is hard to exercise professional judgement and convince others of the veracity of adding or deducting a particular point for overall impression. As a consequence, the teacher tends to use a very narrow range of deduction points and to repeatedly take off points for minor errors. This type of assessment discourages students from making their own translation choices creatively, and instead encourages them to copy the teacher's translation style. As a result, when they get a bad grade, rather than trying to analyze reasons why some of their translation choices were identified as errors, they tend to think they were given the grade because they did not follow the teacher's style.

In order to address the drawbacks of the NAATI criteria, I have developed and used meaning-oriented assessment criteria, primarily drawing on SFL-based text analysis, which is taught in *Introduction to Text Analysis*, a core course offered at Macquarie University. The following section will briefly introduce theories that are fundamental to the meaning-oriented assessment criteria.

Underlying theories of meaning-oriented assessment criteria

The meaning-oriented assessment criteria proposed in this chapter are grounded in meaning analysis as proposed by systemic functional grammarians (Halliday 1978; Halliday & Matthiessen 2004), in addition to Skopos theory and functionalism in translation studies (Reiss & Vermeer 1984; Nord 1997). In fact, there appears to be fundamental compatibility between linguistic theory and translation theory, although few attempts have been made to explore the compatibility as yet. This section introduces the underlying theories in turn.

[3] The instructions do not seem to be based on empirical evidence. This is certainly an area that requires rigorous empirical study.

Systemic functional linguistics (SFL) theory

SFL theory has a strong social orientation stemming from the early period of its development. The theory was influenced by Firth's model of language in the initial conceptual period (Firth 1957), and was developed into a full-fledged theory of language by Halliday and other SFL scholars. One of the most distinguishable features of SFL is the incorporation of situational context and cultural context, based on the research of Malinowski (1935), into the linguistic model. Malinowski was an anthropologist who studied the culture of people living on the Trobriand Islands in Papua New Guinea. When he was translating some of the texts that he collected from his fieldwork, he realized that his translations did not make much sense to his target English-speaking readers due to their lack of understanding of the situational and cultural contexts.

SFL has provided a theoretical framework for a number of language-related disciplines. In translation studies, Halliday's systemic functional model has provided a solid theoretical basis for Catford (1965), House (1977/1997), Hatim and Mason (1990, 1997), Bell (1991), Baker (1992), Munday (1997), Trosborg (2002) and Steiner (2002, 2004), to name a few.[4] House (1977/1997), in particular, has made a substantial contribution to the field of translation quality evaluation, and Trosborg (2002) discusses the role of discourse analysis in training translators. Both of the scholars use Halliday's SFL theory as the primary framework for their work. Centrality of meaning and the shared view of meaning seem to be core links between SFL and translation studies. Newmark (1987: 293) explains:

> Since the translator is concerned exclusively and continuously with meaning, it is not surprising that Hallidayan linguistics, which sees language primarily as a meaning potential, should offer itself as a serviceable tool for determining the constituent parts of a source language text and its network of relations with its translation.

Systemic functional linguists regard language as a meaning-making resource through which people interact with each other in given situational and cultural contexts. They are mainly interested in how language is used to construe meaning. Therefore, language is understood in relation to its global as well as local contexts. This fundamental view of language is expressed through several strata or levels in SFL theory, as the diagram below, adopted from Matthiessen (1992), demonstrates.

The levels depicted in Figure 1 are context, which includes both context of situation and context of culture; discourse semantics; lexicogrammar; and

[4] For a detailed discussion, see Steiner (2005).

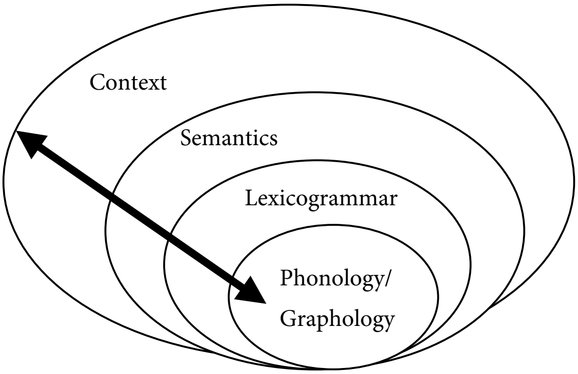


Figure 1. Levels of language

phonology/graphology. It can be said that a higher level provides a context for its lower level, and that a higher level cannot exist without its lower level. For instance, unless a word is expressed in a spoken or written form, we cannot talk about grammar. Unless an utterance is made at the lexicogrammatical level, we cannot create a text or discourse at the semantic level. Therefore, in SFL, it is common practice to study lexicogrammar, which is mainly concerned with meaning at the clause level, in relation to semantics, which is primarily concerned with meaning at the text or discourse level, and vice versa. This is one of the reasons for the strong relevance of SFL theory to translation studies. Translators cannot create a text without working on meaning at the clause level, and cannot produce a coherent text without working on meaning at the text level.

In SFL, grammar is a way of describing lexical and grammatical *choices* rather than a way of prescribing a set of grammatical *rules*. The choices are interpreted as linguistic resources which the speakers of the language use to realize meaning. Halliday (1994) states:

> One way of thinking of a functional grammar … is that it is a theory of grammar that is orientated towards the discourse semantics. In other words, if we say we are interpreting the grammar functionally, it means that we are foregrounding its role as a resource for construing meaning. (Halliday 1994: 15)

Halliday (1994: 35) asserts that a distinctive meaning is construed through three different kinds of meanings: ideational, which includes both experiential and logical resources; interpersonal; and textual. Experiential meaning represents our experience of the world, namely who (participant 1) does what (process) to whom (participant 2), how, when, and why (circumstances). Logical meaning refers to logical relations between the experiences. Interpersonal meaning expresses interaction and the relationship between the speaker and the listener or a personal attitude. *Textual* meaning organizes ideational and interpersonal meanings into a coherent linear whole as a flow of information. Each abstract mode of meaning is realized through a particular linguistic system, namely TRANSITIVITY, MOOD and THEME.[5] At the same time, these modes of meaning are associated with the situational aspects of register (Halliday 1978, 1994). Halliday's register theory basically suggests that there are three variables in any situation that have linguistic consequences and they are field, tenor, and mode. Field refers to the focus of our activity (i.e. what is going on); tenor refers to the way the speaker relates to other people (e.g. status in relation to power); and mode refers to the communication channel (e.g. spoken or written). (For a detailed explanation, see Martin 1992 and Eggins 2004.)

Each aspect of meaning is interpreted based on the evidence of linguistic resources at the clause level. Therefore systemic functional grammar (SFG) is the same as other grammars in the sense that it looks at linguistic features at the clause level, but is significantly different from the others in that it does not interpret them as a set of rules but rather describes them as resources for interpreting different aspects of meaning. Furthermore, it is viewed in relation to the context. This correlation can be presented diagrammatically, as in Figure 2.[6]

Ideational meaning is realized through the TRANSITIVITY system in association with the field of the text; interpersonal meaning is realized through the MOOD system in association with the tenor of the text; and textual meaning is realized through the THEME system in association with the mode of the text. Martin (2001: 54) explains the importance of the correlation as follows:

> This correlation between register categories and functional components in the grammar is very important. It is this that enables systemicists to predict on the basis of context not just what choices a speaker is likely to make, but which areas of the grammar are at stake. Conversely it allows us to look at particular grammatical choices and to understand the contribution they are making to the contextual meaning of a sentence. *This makes it possible for systemic linguists to argue on the basis of grammatical evidence about the nature of field, mode and tenor at the same time as it gives them a way of explaining why language has the shape it does in terms of the way in which people use it to live.* (italics mine)

[5] Following SFG conventions, the names of linguistic systems are written in capital letters (e.g. system of THEME), whereas the names of structural functions are written with an initial capital (e.g. Theme and Rheme).

[6] This is a simplified diagram to illustrate the correlation between grammar, semantics and context. There are of course other systems that are used as resources to construe different meanings.

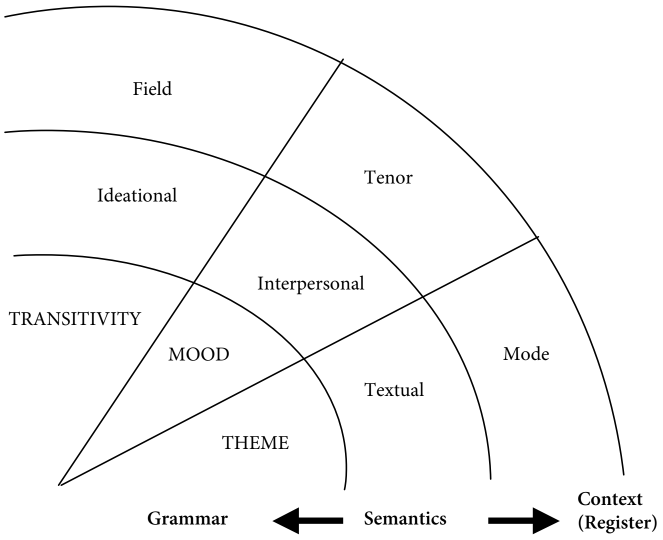


Figure 2. The correlation between grammar, semantics and context

The correlation between contextual variables (register) and grammatical choices described by Martin is also highly important in translation in general and in translation assessment in particular. In order to produce a translation that functions within a specific register (field, tenor, and mode), translators may have to "legitimately manipulate" (House 2001: 141) the source text at all these levels using a "cultural filter" (ibid.: 141) and linguistic knowledge of both languages. Therefore translator teachers as well as translators in training should consider the target text's register and assess whether or not linguistic resources (lexicogrammar) have been used adequately to create different kinds of meaning (semantics) within the register. This assessment approach is significantly different from one that focuses on whether or not a translation contains any grammatical errors given that a translation without any grammatical errors may still be regarded as inappropriate if it does not recreate the required register.

A similar argument has been made for the assessment of the discourse of learners of English as a second language in tertiary education. The evaluation of causal explanation, an essential part of academic literacy, was examined in Mohan and Slater (2004). The study revealed that current models designed to assess second language competence are only efficient in checking whether the writer has violated the basic rules of the language. The graders who participated in the study intuitively judged one text as 'more advanced' than the other but they admitted that the assessment instrument would not account for the discrepancy (ibid.: 265). Most translation assessors would also have had this experience at least once or twice, if not frequently. Mohan and Slater argue:

> The obvious implication for the evaluation of discourse from traditional grammar and the language as rule perspective is to evaluate the correctness of form to see whether language rules are violated or not. Judgment about the meaning of discourse may be made at the same time, but they are usually holistic, impressionistic and, consistent with the conduit metaphor, made independently of the evaluation of form. The implications for evaluation from Halliday's view are much different. The emphasis shifts from what the learner cannot do to what the learner can do. This view encourages us to evaluate discourse as making meaning using linguistic resources in context. How does the writer relate form and meaning? (ibid.: 258)

This perspective offers the same reasons for the use of SFL that are proposed in this chapter in relation to translation assessment, namely that SFL, which theorizes the correlation between grammar, semantics and register can make a significant contribution to improving translation assessment.

Translation: Product vs. process

In order to assess a translation systematically, one needs to understand the process through which a translation was produced. As shown in Figure 3 below, the translator produces a target text (TT) based on his or her own understanding of a source text (ST). This understanding is based on the translator's language skills, text analysis skills, cultural and background knowledge. When it comes to the production of a target text, he or she makes choices in such a way as to convey the multi-dimensional meaning of the ST in an appropriate form of the TT. In the choice-making process, the negotiation of meaning is inevitable. That is, although the translator understands all different kinds of meaning, it may be impossible to convey every aspect of meaning in the TT because grammatical resources that are responsible for different aspects of meaning work differently from language to language. Therefore, the translator needs to decide which aspects of meaning are most important, considering the context that determines the register of the TT.

Skopos theory, which focuses on the purpose of the translation, argues that the translator should adopt translation methods and strategies to produce a TT that fulfils its functional roles (Reiss & Vermeer 1984). The functional roles of the TT are often determined by a "translation brief" that states the TT's purpose and other relevant information (Nord 1997: 30). In other words, the brief is the source for determining the context of the TT, which the translator depends on to decide whether to realize an "overt" or "covert" translation (House 1997: 66). An overt translation is a translation in which it is made explicit or obvious that

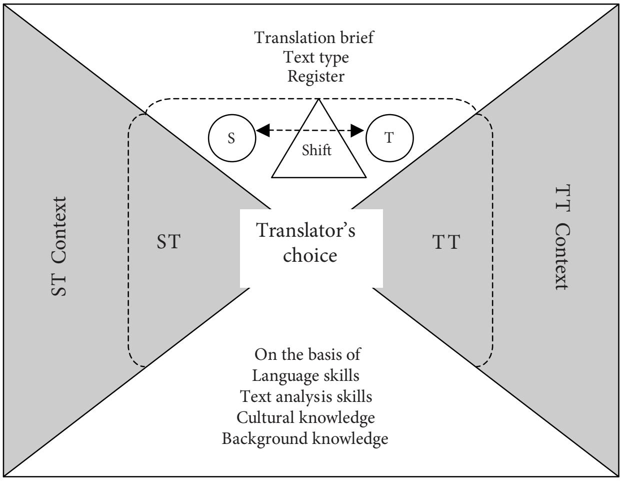


Figure 3. Translation product, process and skills

what is being produced is a translation, while a covert translation "is a translation which enjoys the status of an original source text in the target culture" (ibid.: 69). The function of a covert translation is "to recreate, reproduce or represent in the translated text the function the original has in its linguacultural framework and discourse world" (ibid.: 114). However, the distinction between "overt" and "covert" is, as House points out, a cline rather than a pair of irreconcilable opposites. The relevance of these notions to translation assessment can be found in the fact that translators are required to produce texts that suit a certain context (register). Therefore, any translation assessor must understand beforehand the contextual information and judge the extent of covertness or overtness that would be necessary for the translation. They also need to assess how appropriately linguistic resources have been used in the translation.

The fact that there is a cline explains why the notion of translation shift is essential. Figure 4 shows the continuous process of meaning negotiation, which was briefly explained above. The process of negotiation takes place through translation shift. The term "translation shift" originates in Catford's *A Linguistic Theory of Translation* (1965: 73–83), and it means "departures from formal correspondence in the process of going from the SL to the TL" (ibid.: 73). Depending on the degree of covertness of the translation, the translator may have to decide how far the translation choices should move away from word-for-word equivalence. A literal translation tends to be closer to it, and a free translation tends be further away from it. Again, translation assessors using any assessment system need to consider how meaning has been negotiated within the given situational and cultural contexts.

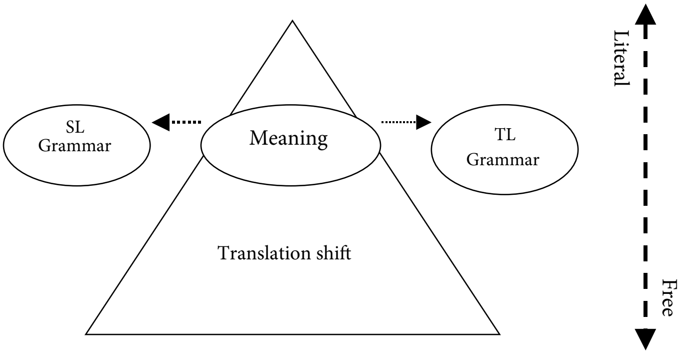


Figure 4. The meaning negotiation process through translation shift

Matthiessen (2001: 79) explains the difference between free and literal translation in relation to the environment of translation:

> The narrower the environment, the more "literal" the translation – e.g. word for word translation (rather than clause-based translation) or translation of wording [lexicogrammar] rather than translation of meaning [semantics]. In the default case, "free" translation is probably preferred as the most effective form of translation. However, freedom is a matter of degree. Perhaps one of the freest types of translation is the translation of comic strips. Ingrid Emond used to translate Donald Duck from Italian to Swedish and she told me she enjoyed this task because the translation could be quite free as long as it made contextual sense – and as long as it was in harmony with the pictorial representation of the narrative. And there are of course contexts of translation … where "literal" translation has value – e.g. context in linguistics or translation studies where we try to indicate how the wording of a particular language works.

The concepts and notions in SFL and translation studies discussed above have meaningful implications for translation assessment. Firstly, a translation must be treated as discourse that fulfils its functions within a specific context. As a consequence, secondly, what the translation assessor should do is not just focus on whether or not there are any grammatical errors in the translation but, more importantly, whether or not the translation as a text or discourse serves its purpose within the context. The next section, which presents the SFL-based meaning-oriented assessment criteria with sample texts, will demonstrate how these implications of SFL theory can be addressed.

Meaning-oriented assessment criteria

The meaning-oriented assessment criteria presented here are devised to address the limitations of the NAATI assessment criteria as a formative tool within the institution. The new assessment criteria are still within the framework of the NAATI criteria, and deducted points are still subtracted out of 45 (the full mark for each translation is 45 points on the NAATI translation exam). A range of deductions in points is suggested. It is inevitable and essential that the grader will need to determine the appropriate extent of translation shift in the different modes of meaning, considering the contextual factors such as the translation brief and register. In this process it cannot be guaranteed that different graders will be in agreement all the time, in the same way that it is not uncommon for different graders to give different scores for an essay. However, if graders can identify their differences according to the categories as suggested below, any discussions to narrow the gaps in the marking would be more efficient than discussions based on the graders' personal preferences or impressions. The decision in relation to the scales, such as 1–2, 1–3 and 3–5, is based on an analysis of points deducted in translation examinations graded by the author and other graders over a number of years.

In the meaning-oriented assessment criteria (see Table 1), translation errors are categorized into major and minor errors. Major errors are those that influence one or more aspects of meaning, while minor errors are simple mistakes that have little impact on the delivery of ST meaning. Major errors are analyzed on the basis of different aspects of meaning (Experiential, Logical, Interpersonal, and Textual), and whether the error has an impact on the accurate delivery of the meaning of the ST (Accuracy) or on the natural delivery of the meaning in the TT (Naturalness). These categories will be illustrated with examples in the section below.

Unlike the NAATI criteria, the present criteria do not specify possible forms of errors, such as additions, omissions, and inadequate equivalence, because what is important is to judge whether a mistake has something to do with accurate and natural delivery of different aspects of meaning. Additions and omissions can be employed as legitimate translation strategies in certain circumstances. Thus such categories are potentially misleading student translators to think any addition or omission is wrong, which in turn tends to lead to the production of a literal translation heavily influenced by the source text structure.

Table 1. Meaning-oriented assessment criteria

			Lexis	Clause	Text
Major	Experiential	Accuracy	1-2 pts	2-3 pts	
		Naturalness	1-2 pts	2-3 pts	
	Logical	Accuracy		1-3 pts	
		Naturalness		1-3 pts	
	Interpersonal	Accuracy	1-2 pts		3-5 pts
		Naturalness	1-2 pts		3-5 pts
	Textual	Accuracy		1-2 pts	3-5 pts
		Naturalness		1-2 pts	3-5 pts
Minor	Graphological mistakes such as spelling			0.5	
	Minor grammar mistakes that do not impact meaning			0.5	

Illustration with sample texts

This section will illustrate the above meaning-oriented assessment criteria with sample texts (see Appendix 1). It will explain why an error or issue was identified in a particular way, and show how many points were deducted and why. For the purposes of illustration, a pair of sample texts, that is an English source text and a Korean target text, will be used. For ease of demonstration, the following translation is a composite text which includes instances of erroneous translations by different students enrolled in *Translation Practice* in 2007. Sections in bold highlight errors or problems, while italicized sections highlight inevitable or justifiable translation shifts. The English text was given to the students in a class, and they were required to translate it in a period of one hour. A translation brief provided with the text outlined the source of the ST, the intended target readers of the TT, and the place of publication. Therefore, students had to decide on an appropriate point on the cline between overt and covert translation based on the contextual information provided.

The English text is titled *The Indian Exception,* and the translation brief stated that it is an excerpt from an article from the printed version of *The Economist* and requests a translation into Korean for an equivalent magazine in Korea. In terms of field, the socio-semiotic function of this text is reporting. The text deals with Australia's new uranium policy to lift a ban on exporting uranium to India. In terms of tenor, the institutional role is expert (reporter) to educated people who are assumed to have an interest in and knowledge about current international affairs. This adapted reporting text expresses a clear opinion about the situation toward the end, using a modal finite *would*, which indicates high possibility. In terms of mode, it is a written and monologic text published in an international weekly magazine. Theme analysis at the clause level shows a tight coherent pattern in which new information introduced in the previous discourse is picked up as Theme in the following discourse.

Considering the ST register and the translation brief given, the TT is required to have the same register and functions as the ST. Therefore, the TT to a large extent needs to be a covert translation so that the target reader can understand what is going on in Australia without experiencing serious difficulties in reading the translated article. In order to render a covert translation, the translator has to make translation shifts, taking into consideration the expected register of the TT at the context level and how it is realized at the lexico-grammatical level. In this case, the translator needs to reflect on certain patterns of lexical choices in journalistic texts that deal with current issues. For example, in Korean, such journalistic texts use nominalization to a large extent (field) and, to maintain the high level of formality (tenor), they also use words made up of Chinese characters. Also pronouns and definite articles are used much less in Korean than in English and their cohesive function is often performed through ellipsis or repetition of nominal groups (mode). The translator needs to take these features into account when selecting lexical choices in order to produce a target text of the expected register.

Experientially inaccurate translations

Example 1 shows the English ST and the composite student Korean TT with the meaning-oriented grading scheme error deductions for experiential meaning for each sentence.

The first sentence of the TT has failed to accurately deliver the experiential meaning of the ST. While the main experiential meaning of the ST is that *Australia's outback deserts* (participant 1) *make up for* (process) *what they lack in water* (participant 2) *in uranium* (circumstance: how), the TT rendered the second participant into *problems caused by the lack of water* and the process, *make up for*, into *solve*. As a consequence, the TT says that in Australia's outback deserts people solve the problems caused by the lack of water with uranium and, as macro-Theme (that is, Theme at the text level), it provides a substantially different orientation in relation to the remaining sentences. Considering that the experiential meaning error occurs in a sentence that is important textually (i.e. in macro-Theme), 3 points were deducted instead of 2.

Sentence 7 also contains a serious experiential error. Although the error has occurred at the local level of process, namely *lift a ban* has been rendered *ban*, it is as serious as the error in Sentence 1 because it delivers the opposite message to the target reader from that of the ST and renders the rest of the TT contradictory.

S. no.	English source text	Korean target text	Back translation	E	
				A	N
1	**What Australia's outback deserts lack in water they make up for in uranium.**	**호주 오지 사막에서는 물부족으로 오는 문제는 우라늄으로 해결하고 있다.**	**In Australia's outback deserts (people) solve the problems caused by the lack of water with uranium.**	3	
7	But on August 16th, Australia's prime minister said he would **lift a ban on selling uranium to India,** *which refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more.*	그러나 지난 8월 16일, 호주 총리는 인도에 **우라늄을 수출하는 것을 금지할 수도 있다**.는 가능성을 시사했다. *인도는 핵확산 방지 조약에 서명하기를 거부하고, 핵무기를 시험했던 적이 있으며 앞으로의 핵무기 실험 가능성도 배제하지 않고 있다.*	But on August 16th, Australia's prime minister said he would **ban exporting uranium to India**. India refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more.	3	
10	Howard first **flagged** the change of Australia's nuclear policy during a visit to New Delhi in early 2006.	하워드 총리가 호주의 핵 정책에 처음으로 변화를 **가져온 것은** 2006년 초 뉴델리 방문 기간 중이다.	It was during a visit to New Delhi in early 2006 when Howard first **brought in** the change of Australia's nuclear policy.	2	

Example 1. Experientially inaccurate translations

Therefore, 3 points were deducted. The error in Sentence 10 is also an error of inaccurate rendering of the process, namely *flagged* is rendered *brought in*. However, 2 marks were deducted because the information in Sentence 10 is less critical to the overall meaning of the text than that of Sentence 7 and, therefore, the impact of the error is not as significant.

Experientially unnatural translations

Example 2 presents three sentences that contain parts that cause the TT to sound unnatural. The target reader might understand the meaning but would certainly know that it is an inadequate word-for-word translation.

The examples here are all related to differences in the transitivity systems of Korean and English. That is, in Korean, it is rare for an inanimate object to be a participant in most process types (namely material, mental, behavioral, verbal, and possessive relational), while such an object can be a participant in any process

S.	English source text	Korean target text	Translation shift	E	
no.			needed	A	N
2	**They contain** almost 40% of the world's known low-cost reserves of the nuclear fuel.	호주 사막은 세계에 알려진 저가 핵연료의 40%에 가까운 양을 비**축하고 있다**.	**In Australian deserts**, almost 40% of the world's known low-cost reserves of the nuclear fuel **is reserved**		1
4	And **ore** from Australia's three operating mines **supplies** about a quarter of the world's uranium-oxide exports.	또한 현재 가동 중인 호주의 탄광 세 곳에서 나는 광석은 세계 산화 우라늄 수출량의 4분의 1 가량을 **공급한다**.	And ore from Australia's three operating mines **accounts for** about a quarter of the world's uranium-oxide exports.		1
13	**Uranium mining has always divided Australians**, but more seem to be leaning towards an expansion of the industry in response to global warming.	**우라늄 광산업은 언제나 호주인들을 분열시켰지만** 근래에는 지구 온난화에 대한 관심으로 우라늄 채굴 사업의 규모를 늘리자는 방향으로 의견이 모아지고 있는 듯하다.	**Australians' opinions about uranium mining have been always divided**, but more seem to be leaning towards an expansion of the industry in response to global warming.		2

Example 2. Experientially unnatural translations

type in English. Therefore, in Sentence 2, when an inanimate object is used as a participant in a possessive relational clause or, in Sentences 4 and 13, when an inanimate object is used in a material clause, a translation shift is inevitable in order to produce a natural translation. Consistent with NAATI's suggestion that an issue related to natural rendering should be regarded as less serious than one related to accurate rendering, 1 mark was deducted for Sentences 2 and 4. Given the fact that it is extremely rare to use an inanimate object in a material clause, and the resulting possibility of hindering the target reader's comprehension, 2 marks were deducted for Sentence 13.

Logically inaccurate translation

Example 3 presents a sentence that demonstrates a logically inaccurate translation.

In the ST, the clause that starts with *guaranteeing* is an example of a non-finite clause, in which logico-semantic relations may not always be clear. The

S.	English source text	Korean target text	Back translation	L	
no.				A	N
9	India will first have to sign a safeguards agreement with Australia, **guaranteeing** that none of its uranium will be diverted to weapons.	인도는 우선 호주와 안전 보장 협정을 맺어야 하며, **협정은** 인도로 판매된 우라늄이 일정 무기로 변환되지 않을 것임을 **보장하는 것이다**.	India will first have to sign a safeguards agreement with Australia, **and the agreement will be what guarantees** that none of its uranium will be diverted to weapons.	2	

Example 3. Logically inaccurate translation

relation between the two clauses can be analyzed as an extension, which means the second clause builds up the experience of *India will have to guarantee* following the experience of *India will first have to sign*. Alternatively it can be analyzed as an enhancing relation in the sense that *India will have to sign … in order to guarantee*. However the logico-semantic relation in the translation is that of elaboration, which explains what the agreement is about. This relation is different from any possible analysis of the source text. Therefore, the TT misrepresented the logical link between the two clauses, for which 2 points were deducted, as the error also leads to a misrepresentation of the experiential meaning, changing the participant of the process, *guaranteeing*, from *India* to *the agreement*.

Logically justifiable translation shift

Example 4 presents a sentence that demonstrates a logically justifiable translation shift. It was discussed in Example 1 above in relation to its experiential meaning error, for which 3 points were taken off.

This example also includes a translation shift of logical meaning in that one sentence was translated into two sentences. It could have been translated into one sentence, but this would have resulted in the TT structure being too complicated. This is because for one sentence *which refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more* has to be translated before *India* in the TT. In addition, the logico-semantic relation of the second sentence is an elaboration of *India,* and so does not change the relationship between the counterparts of the ST. Therefore, it was regarded as a justifiable translation shift.

S.	English source text	Korean target text	Back translation	L	
no.				A	N
7	But on August 16th, Australia's prime minister said he would **lift a ban on selling uranium to India**, *which refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more.*	그러나 지난 8월 16일, 호주 총리는 인도에 **우라늄을 수출하는 것을 금지할 수도 있다**는 가능성을 시사했다. *인도는 핵확산 방지 조약에 서명하기를 거부하고, 핵무기를 시험했던 적이 있으며 앞으로의 핵무기 실험 가능성도 배제하지 않고 있다.*	But on August 16th, Australia's prime minister said he would **ban exporting uranium to India**. *India refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more.*		

Example 4. Logically justifiable translation shift

S.	English source text	Korean target text	Back translation	I	
no.				A	N
14	However, should India test another bomb, public outrage **would** kill uranium exports in a flash.	그렇지만 만일 인도가 또 다시 핵 무기 실험을 하게 되면, 호주 대중의 노여움으로 인도로의 우라늄 수출은 그 즉시 중단**될 지도 모른다**.	However, if India tests another bomb, *due to Australian public outrage, the uranium export* **might** *be stopped immediately.*	2	

Example 5. Interpersonally inaccurate translation

Interpersonally inaccurate translation

In the following Example 5, a sentence is presented to demonstrate an interpersonally inaccurate translation, and explain why one lexical mistranslation can be treated as a more serious error than others.

One might treat the error of translating *would* into *might* as a simple lexical error, or overlook this kind of error because it does convey the experiential meaning. However, considering this is the last sentence of the text, and so has textual significance as macro New, and indicates the paper's opinion about the situation through the use of the modal finite, *would*, it is rather a serious issue. Therefore, the lexical error was regarded as a serious interpersonal error, for which 2 points were deducted.

S.	English source text	Korean target text	I	
no.			A	N
3	It is big business for Australia: exploration companies are at present **spending** ten times more **money** searching for deposits than they did three years ago.	우라늄은 호주에게는 큰 사업으로 우라늄 탐사 기업들은 현재 저장된 우라늄을 찾는데 3년 전보다 10배에 가까운 **돈을 쓰고 있다**.		1
5	Until now all this **has gone** to countries that have signed the Nuclear Non-Proliferation Treaty (NPT).	지금까지 모든 우라늄은 핵확산 방지 조약 가입국으로만 **팔려갔다**.		1

Example 6. Interpersonally inadequate translations

Interpersonally inadequate translations

Example 6 presents two sentences that demonstrate interpersonally inadequate translations. There is no back-translation for this example because it would not successfully illustrate the interpersonal issues involved. This is because in this example a translation shift is necessary to meet certain expectations required in these two sentences, as will be explained below.

Sentences 3 and 5 accurately deliver the experiential meaning but fail to make formal lexical choices, which are expected in a Korean magazine that is an equivalent to *The Economist*, as explained above. The lexical choices in this example would be suitable for informal talk. More appropriate choices would be the Korean equivalent of investment for *spending* and the Korean equivalent of exported for *has gone*. One point was deducted for each error, as each was regarded as a relatively less serious issue compared to the error in Example 5.

Textually inaccurate translations[7]

Example 7 presents two sentences that demonstrate textually inaccurate translation. Among textual meaning issues, the issue of cohesion and coherence is most critical in that it often leads to an inaccurate rendering of experiential meaning, as well. *This* in Sentence 6 refers to Sentence 5, *Until now all this has gone to countries that have signed the Nuclear Non-Proliferation Treaty (NPT)*, rather than *by signing the agreement*. The lack of cohesion also led to the experiential meaning error in this case and therefore 2 points were deducted.

[7] A discussion of natural delivery of translation in relation to Theme can be found in Kim (2007b and 2008).

S.	English source text	Korean target text	Back translation	T	
no.				A	N
6	**This** ensures, in theory, that they will use it to produce electricity rather than bombs.	**조약을 맺음으로** 이론상으로는 수입된 자원이 폭탄 제조에 쓰이기 보다는 전기 생산에 쓰일 것이라는 것을 보장한다.	**By signing the agreement**, in theory, (they) guarantee that the imported resource will be used to produce electricity rather than bombs.	2	
8	**The** sales will be subject to “strict conditions”.	**우라늄 판매는** 앞으로 “엄격한 조건”에 한해 이루어질 것이다.	**Uranium sales** will be subject to “strict conditions”.	1	

Example 7. Textually inaccurate translation

The error in Sentence 8 is also related to cohesion. While in the ST it is clear that the sales refers to the sales of uranium to India, the TT appears to suggest uranium sales in general. The issue may have occurred because Korean very rarely uses articles. A strategy to address this issue is to add the necessary information with *sales*, such as 대인도 우라늄 판매 (Uranium sales to India).

As demonstrated by the translation errors and shifts presented above, the meaning-oriented criteria proposed here are useful for analyzing the nature of mistranslation, (un)justifiable omission, and unnatural translation. The examples also serve to illustrate the reason for this type of assessment, as well as the process of deciding on deduction points. Some examples are provided below, along with a comparison of the error category of the meaning-oriented criteria and a possible error category of the NAATI criteria for each error.

As shown in Table 2, “mistranslation” and “non-idiomatic usage” seem to be the categories of the NAATI criteria that are suitable for most of the errors. There are a few problems with this. Firstly, “mistranslation” is overused so much that it does not mean anything but simply indicates that the translation is wrong. However, the meaning-oriented criteria enable the grader to explain what aspect of meaning is mistranslated and why. When trained with such an analytical tool, students can develop skills to assess translations of their own and others.

Secondly, the category of “non-idiomatic usage” chosen for Examples 2 and 6 does not represent the reasons as to why these two translations sound awkward. In fact, the sources of the awkwardness in the two examples are different. Example 2 sounds awkward because it does not take into account the limited use of inanimate subject in expressing some experiences in the target language, while the issue in Example 6 was caused because of the lack of consideration of tenor (interpersonal meaning) required in the target text.

Table 2. Comparative assessment of translation errors

Ex.	ST	TT (deduction point)	NAATI criteria	Meaning-oriented criteria
1	lift a ban on selling uranium to India	ban exporting uranium to India (3)	Mistranslation	EXPERIENTIAL (A) Misrepresentation of the Process (what happened)
2	They (Australian desserts) contain	In Australian deserts (1)	Non-idiomatic?	EXPERIENTIAL (N) Inanimate subject is hardly ever used as possesser in Korean
3	guaranteeing that none of its uranium will be diverted to weapons	and the agreement will be what guarantees (2)	Mistranslation	LOGICAL (A) Misrepresentation of the logical relation (elaboration vs. extending/enhancing)
5	would	might (2)	Mistranslation	INTERPERSONAL (A) Misrepresentation of modality (probability)
6	spending ten times more money	돈을 쓰다 (1)	Non-idiomatic?	INTERPERSONAL (N) Inadequate formality
7	This	By signing the agreement (2)	Mistranslation	TEXTUAL (N) Cohesion

Thirdly, it is hard to justify why the deduction point of the mistranslation of one word in Example 5 is greater than any other mistranslations of words if it is simply labeled as a mistranslation of a word. However, when different aspects of meaning are considered in assessing translation products in relation to the register analyzed for the ST and TT, it becomes clearer and easier to judge the seriousness of an error. As explained in Example 5 earlier, the mistranslation of *would* is graded as a more serious error than that of other individual words because it plays an important interpersonal role in a textually significant sentence.

This chapter so far has shown how the meaning-oriented criteria can be used for formative assessment of translations, and has demonstrated that SFL-based text analysis linking three layers of language (lexico-grammar, semantics, and context) provides both an efficient tool for translation assessment and the technical terms needed to explain the subtle and complicated concept of translation quality. The following section will discuss the pedagogical efficacy of the text analysis underlying the meaning-oriented assessment criteria.

Applications in a pedagogical context

Assessment of text through text analysis serves as an analytical tool to systematically compare the lexical and grammatical resources of the two languages with reference to different modes of meaning within context. Therefore, it enables students and teachers to identify translation errors, as well as translation choices or strategies, in different dimensions of meaning. It also enables them to explain why they are analyzed as such, and how critical analysis is in the particular translation assignment, referring to the text's contextual information using the evidence of wording and grammar.

In terms of selecting testing materials, text analysis is also helpful in selecting a variety of texts that impose different translation challenges so that they can be handled more systematically, with a particular emphasis on one aspect of meaning at a time. In addition, it has been of great help in providing constructive feedback on individual students' performances in translation assignments and exams.

As students learn how to analyze translation errors, they start to analyse their own error patterns and develop strategies to avoid them, and gradually move away from the source text structure to be creative in producing a target text. Eventually, this approach helps them to become autonomous learners and their own quality controllers because they do not have to rely on the teacher's intuition-based feedback. It also stimulates their interest in research as they see the relevance of theory to practice.

In the following section, the pedagogical efficacy of text analysis will be discussed based on the results of students' surveys (conducted on two different occasions) and the NAATI recommendation ratios from 2004 to 2008. It is important to discuss the pedagogical efficacy of using the meaning-oriented assessment criteria on the basis of evidence, since the criteria are presented here as a formative assessment tool designed to help students analyze their own translation issues and develop their own translation strategies. In fact, the decision to continue to use text analysis for formative assessment was made on the basis of the measurement of both qualitative data [students' learning journals and quantitative data (the surveys and NAATI ratios)]. Significant changes were observed when text analysis was incorporated into translation assessment.

Survey results

At the end of the second semester of 2006, a survey was conducted of Korean students who were taught to apply text analysis in the second-semester translation practice course.[8] The same survey was conducted again at the end of the first semester of 2008. The surveys focused on three questions: the degree of difficulty in following the new approach; its usefulness in developing critical thinking on translation issues; and its usefulness in enhancing translation skills. The survey results will be presented together with students' learning journals, where relevant. Writing learning journals is a routine activity in many classes in the Macquarie University programs. Students are guided to reflect on their own learning, ask questions or make suggestions in their learning journals. Data discussed in this section was obtained from some of the learning journals submitted in the last week of the first semester of 2008 by Korean students enrolled in *Introduction to Text Analysis* and *Translation Practice*. In the final entry in their journals, students reported on highlights in the learning process during the text analysis course during the semester.

A main concern about applying text analysis in teaching translation and self assessment was whether or not the application of SFL-based analysis would be too difficult or challenging for students, given that they are expected to deal with a new linguistic paradigm, and must learn new concepts and terminologies that they did not learn in previous educational contexts. Therefore, the first question was concerned with the level of appropriateness of teaching. It is presented here, with results from both surveys:

1. The application of SFL-based text analysis in this course was at an appropriate level for me.

As shown in Table 3, the majority of students surveyed in both periods answered that it was at an appropriate level for them: in the first survey, 87.5% of them agreed, and in the second survey, 16% of them strongly agreed. The slight increase in the "strongly agree" response may be attributed to the increased portion of contrastive analysis between the two languages, drawing on a systemic functional description of Korean Theme (Kim 2007c), that was incorporated in classes by the time of the second survey. The contrastive analysis was extremely limited in 2006 due to the lack of resources to describe Korean from the systemic functional point of view.

The second survey question was concerned with the role of text analysis as a tool for critical thinking. Interestingly, this is the question on which the students agreed most strongly on both occasions (see Table 4). It is presented here, with results from both surveys:

2. The meaning-oriented (experiential, logical, interpersonal and textual) analysis of translation issues helped me think critically about translation issues.

[8] These results are also reported in Kim (2007b).

Table 3. Students' responses to Question 1

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree
2006/2 (8)		87.5%	12.5%		
2008/1 (12)	16%	67%	16%		

Table 4. Students' responses to Questions 2

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree
2006/2 (8)	31.5%	62.5%			
2008/1 (12)	25%	75%			

One of the reasons why they responded very positively to the question seems to be the fact that they learned how to explain complex translation issues in the text analysis class and to apply their knowledge in the translation practice class, as stated in the following excerpts from two students' learning journals:

> The text analysis techniques we learnt from this course provide basis to analyze texts in functional point of view so that we could systematically explain *why* we have to make translation choices, *why* certain translation shifts cannot be avoided due to language features and what translation shifts are inappropriate.
>
> To my surprise, the title for my very first journal was 'choices that translators must make'. I had thought it was only from week 6 that I'd started to grasp the idea of a 'translation shift' but it looks like I knew choices (shifts) already. It was there in my head, and I just didn't know how to put it into words. *I think the whole subject, TRAN819 (Introduction to Text Analysis) benefited me in terms of teaching me words and confidence to explain some things that had existed in my head ever since I had started learning translation and interpreting.* … Also, I think knowing these concepts accelerates the process of becoming an excellent translator. They say a human being is born with a brain, an ability to conceptualize, and as human beings invented language our creativity increased as language allowed us to develop concepts in our head. Without language we would lose the means of communication in our inner world. Only indescribable feelings and basic concepts would be present in our head. *In the same way, SFG gave me language with which I can develop my ideas about translation.*

With the use of SFL, students can articulate what aspects of their translation have improved, as shown in the following excerpts.

> Another thing which I have learned from this course is that *experiential* meaning is my weak area where I was/am the most frequently making errors. In the past, I did not critically and carefully read source texts, which led me to producing totally different target texts.
>
> My view on *interpersonal* meaning was only limited to the appropriate level of lexical and politeness choice depending on the age and the type of the target audience. However this course has broadened my understanding of the interpersonal meaning. *Mood person* and *modality* were the two aspects that I did not fully appreciate the importance of. I didn't think of the consequence of replacing an interactant with a non-interactant in translating.

Other comments on how text analysis is useful as a tool for critical thinking include the following excerpts from students' learning journals.

> On the whole, I learnt how to think critically regarding reading texts through the lectures and tutorials. Before learning this analytical method I had not realized that I did not grasp the exact meaning or fully understand the writer's intention or purposes after reading texts. … it helped improve my reading skills in a great way.
>
> I was surprised to know how much this course had helped me see the translated text critically. While doing the final assignment I could see what looked fairly truthful translated text was full of translation shifts and I could identify many different types of translation shifts by applying text analysis skills. I could see that the purpose of this course is to examine the translation we do more systematically and scientifically using the text analysis skills we have learnt so that we can improve our translation.

The third question was about whether students perceived any improvement in translation competence and skills. On both occasions, the response to this question was less positive than the previous two questions. The questions is presented here, with results from both surveys (see Table 5).

3. The application of text analysis to translation helped me improve my overall translation competence and skills.

This response may reflect the fact that a period of time is required for the learner to internalize new knowledge and skills, as a student said in her journal:

Table 5. Students' responses to Question 3

	Strongly agree	Agree	Not sure	Disagree	Strongly disagree
2006/2 (8)	25%	25%	50%		
2008/1 (12)	8%	58%	25%	8%	

> Broadening our lexical basis, extending the cultural understanding and our exposure to the subject matter are important aspects in translation practice. However translation skills that could take a lot of experience over a long period of time without a help of formal teaching can be achieved through this formal academic teaching in a much shorter time frame. It may not appear to have immediate effect in our actual translation partly because it is fairly new concept for many of us and partly because we need more practice in the actual application of the theory. However I can certainly claim that this course has provided deep understanding on many aspects of text that need to be considered and how we can apply them in our translation.

Students also described how their translation competence has improved by reflecting on what has changed in their translation process, as shown in the following excerpts:

> I can think about what a good translation is and how to deal with a text in order to make better translation though this text analysis course. Whenever I translate source text after learning this course, I try to identify field, mode and tenor before starting actual translation. Although I learnt these concepts in advanced writing course, I did not apply these to translation practice and actually I did not know the_importance of these concepts to translation. Before I learned the text analysis, I used to translate a text without thinking and just interpret meaning of words and clause complexes. I thought natural translation is a good translation but I did not think about what natural translation is. However, after learning text analysis I constantly asked to myself *why* my translation is natural or unnatural. I can say that before learning text analysis I just started translating without any preparation but after learning text analysis I can start translation step by step. This text analysis course gives me a big change of *translation process* and makes me to think about a text constantly. All of this makes me to build up my translation strategies.
>
> Another big fruit of this course is learning about *translation shift*. In the past, I used to find Korean equivalence of English words in a dictionary. Although some expressions which came from the dictionary seemed to be unnatural, I had to use it because there was no other way to solve the problem. … When I first learnt this concept I felt like I found a treasure, which I did not expect to have. I learnt that it is definitely helpful to my translation but I also learnt that I have to *be responsible* for making translation shift.

As stated in a number of the extracts above, students found text analysis particularly helpful for explaining aspects of translation quality and analyzing the linguistic resources responsible for them. The development of the skills to analyze translation issues and to view them critically naturally enables them to assess their own translations and others' and improve their translation performance, as will be shown with the quantitative data of NAATI recommendation ratios based on the exam results for *Translation Practice*.

NAATI recommendation ratios

Text analysis and meaning-oriented assessment criteria have been used for formative assessment in *Translation Practice* every semester except three (Semesters 1 and 2, 2004 and Semester 1, 2006). As Figure 5 shows, the ratios of students who were recommended for NAATI accreditation (for translation from English into Korean) significantly increased from around 10% in 2004 to over 30% (Semester 1) and 40% (Semester 2) in 2005 when text analysis started to be used as a tool to assess students' translations in the particular course. The ratios continued to improve from 45% (Semester 2) in 2005 to over 60% (Semesters 1 and 2) in 2007 with more integration of text analysis in the meaning-oriented assessment criteria in 2007. Given the fact that all the exam papers are graded by both the internal lecturer and an external marker who used the NAATI criteria, and that the same entry conditions applied every semester, these were very surprising results particularly given that no other language group in the course has shown such consistent improvement over the period.

Figure 6 compares the recommendation ratios of the other classes where other language students studied the translation practice subject without application of SFL-based text analysis and meaning-oriented assessment criteria. One language group (A) has shown relatively similar recommendation ratios between 20% and 40%, while the other group (B) has shown some occasional improved ratios but not consistent improvements. Two facts, namely that other language groups have not shown consistent improvements and that the texts used for the *Translation Practice* exam are always approved by NAATI, confirm that the texts used in later periods have not been easier to translate and therefore cannot account for the improvements.

My experience over several years as a grader confirms that translation errors identified by using the meaning-oriented criteria suggested here are not fundamentally different from those identified by the NAATI criteria, although they are more useful in assessing textual meaning and more helpful in deciding a deduction point. The data presented above suggest that when students are taught how to make translation choices on the basis of wording and grammar in relation to the context of the translation, they become more confident in making informed choices when translating and learn to control their own translation quality. Their confidence leads to better performance even within the limited time of one semester.

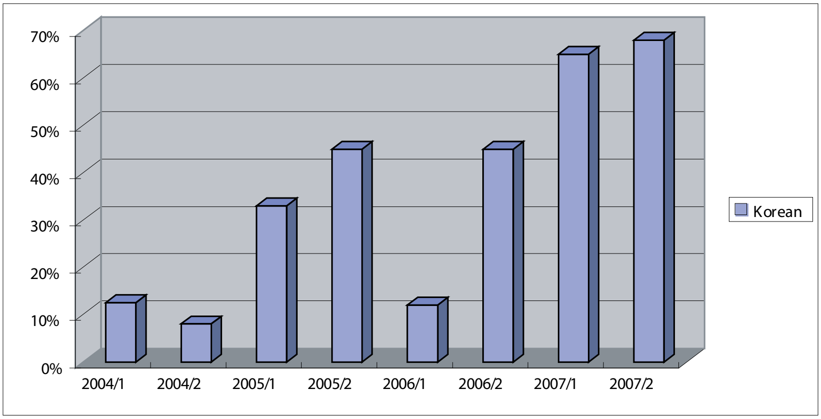
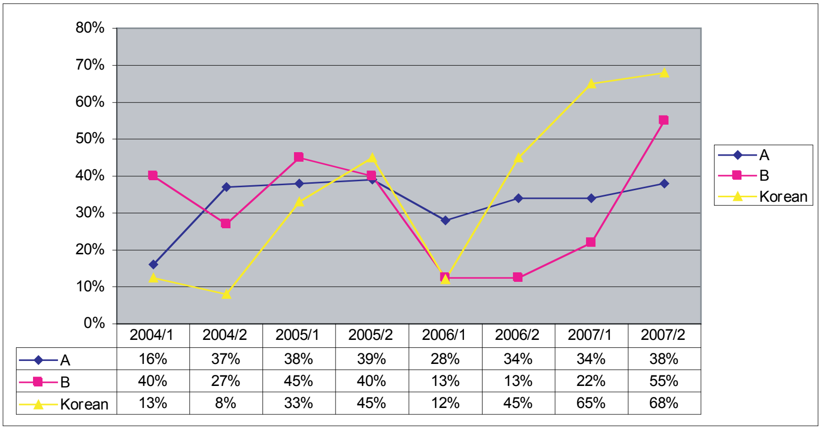


Figure 5. NAATI recommendation ratios of Korean group



	2004/1	2004/2	2005/1	2005/2	2006/1	2006/2	2007/1	2007/2
A	16%	37%	38%	39%	28%	34%	34%	38%
B	40%	27%	45%	40%	13%	13%	22%	55%
Korean	13%	8%	33%	45%	12%	45%	65%	68%

Figure 6. NAATI recommendation ratios comparison

Limitations and further study

This chapter presented meaning-oriented assessment criteria as an approach to assessing different aspects of meaning in translation within context on the basis of the evidence of lexical and grammatical choices. The pedagogical efficacy of this approach was also discussed. It can be viewed as a bottom-up approach in that it deducts points for an inadequate choice at word and clause levels, but it is also a top-down approach in that the judgment of inadequacy comes from a register analysis that encompasses domains of context.

The study discussed the benefits of using the meaning-oriented assessment criteria as a formative tool following a brief discussion of some drawbacks of the NAATI assessment criteria as a formative tool. These issues, namely the lack of clear guidelines on the definition of criteria and deduction points, and the lack of criteria to assess multi-faceted meaning, may be the result of the fact that there has been a limited amount of rigorous research undertaken to explore the complex features of translation that need to be assessed. Therefore it is not surprising that these drawbacks have been identified as critical issues that need to be addressed in order to improve the NAATI criteria as a summative tool. Turner (2008), who was a co-developer of the 2005 edition of the NAATI Manual, suggests that the criteria need to be improved to provide holistic guidance to graders, and detailed and consistent feedback to test candidates. It is a critical problem that must be addressed as a matter of urgency if the field of translation is to be widely recognized as a profession. Brisset (1990) says:

> Can you qualify a translator as 'professional' if he doesn't have the means to talk about his work in *technical terms*? [...] You must be able to read a text to translate it. Reading can be intuitive or it can be based on analysis that draws on a range of concepts and procedures. The purpose of theory is, among other things, to provide the translator with the mastery of these concepts and procedures. And above all, to teach the translator to name his tools, the way any technician learns the name of his tools and the tasks that he carries out.
> (Brisset 1990: 378, English translation by H. Slatyer; my emphasis)

There is substantial potential for the meaning-oriented assessment criteria to make a contribution to improving the NAATI criteria as a summative tool. However, in order to take the meaning-oriented assessment criteria beyond the personal level of use and develop them to the level required for industry standards, the limitations of the present study need to be addressed. Whether or not the meaning-oriented approach to assessment is valid and reliable remains to be determined. A follow-up study to investigate this question is in the conceptual stage. NAATI graders in different languages will be asked to test the criteria with actual translation tests.

A major difficulty anticipated in carrying out the follow-up study is that not all languages dealt with by NAATI have been described in terms of SFL. Although there have been an increasing number of attempts to describe languages other than English from a SFL perspective (cf. Caffarel et al. 2004), it is true that resources are not yet sufficient. Therefore, the follow-up study will be undertaken in the language pairs of English and other languages which have been studied in SFL. The aim of the study will be to find out how efficient the approach is for those who do not have linguistic backgrounds, and also to determine to what extent the approach can solve the existing issues of grading, such as inter-rater reliability.

The fact that linguistic description from a systemic functional perspective is not available for all the NAATI languages is certainly a challenge in applying SFL to advance the field of translation assessment, but it should not be a reason to give up altogether. Instead it can serve as a practical stimulus and encouragement for linguists to work on a number of languages that have not been well investigated.

Mohan and Slater (2004) insist that "SFL has major implications for the assessment of discourse" (ibid: 255). I would argue that it also has major implications for translation assessment: translation is discourse, and serves a certain function within a context. It should therefore be assessed as such, rather than as a series of sentences in isolation from their context. The meaning-oriented assessment criteria drawn from SFL theory proposed here is just one step toward the meaningful, interdisciplinary collaboration between translation assessment and SFL.

References

Baker, Mona. 1992. *In Other Words*. London and New York: Routledge.

Bassnett-McGuire, Susan. 1991. *Translation Studies* (Revised edition). London: Routledge.

Bowker, Lynne. 2000. "A Corpus-Based Approach to Evaluating Student Translations." *The Translator* 6(2): 183–209.

Bell, Roger T. 1991. *Translation and Translating: Theory and Practice*. London and New York: Longman.

Brisset, Annie. 1990. "La théorie: pour une meilleure qualification du traducteur." In *La Traduction au Canada: les acquis et les défis*, M. Cormier (ed.), 235–243. Montréal: CTIC.

Caffarel, Alice, Martin, James R. and Matthiessen, Christian M. I. M. (eds). 2004. *Language Typology: A Functional Perspective.* Amsterdam/Philadelphia: John Benjamins.

Cao, Deborah. 1996. "Translation Testing: What Are We Measuring?" In *Translation and Meaning* (Part 3), Barbara Lewandowska-Tomaszczyk and Marcel Thelen (eds), 525–532. Maastricht: Universitaire Pers Maastricht.

Catford, John C. 1965. *A Linguistic Theory of Translation*. London: Oxford University Press.

Eggins, Susan. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter.

Firth, John R. 1957. *Papers in Linguistics 1934–51*. London: Oxford University Press.

Halliday, Michael A. K. 1978. "Is Learning a Second Language like Learning a First Language All Over Again?" In *Language Learning in Australian Society*, D. E. Ingram and T. J. Quinne (eds), 3–19. Melbourne: International Press and Publications.

Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar*. London: Edward Arnold.

Halliday, Michael A. K., and Matthiessen, Christian M. I. M. 2004. *An Introduction to Functional Grammar*. London: Arnold.

Hatim, Basil and Mason, Ian. 1997. *The Translator as Communicator*. London and New York: Routledge.

Hatim, Basil and Mason, Ian. 1990. *Discourse and the Translator*. London and New York: Longman.

Hatim, Basil and Munday, Jeremy. 2004. *Translation: An Advanced Resource Book*. London and New York: Routledge.

Hönig, Hans. 1998. "Positions, Power and Practice: Functionalist Approaches and Translation Quality Assessment." In *Translation and Quality*, Christina Schäffner (ed.), 6–34. Clevedon: Multilingual Matters.

House, Juliane. 1977/1997. *Translation Quality Assessment*. Tübingen: Gunter Narr Verlag.

House, Juliane. 2001. "How Do We Know When a Translation is Good?" In *Exploring Translation and Multilingual Text Production: Beyond Content*, Erich Steiner and Colin Yallop (eds), 127–160. Berlin and New York: Mouton de Gruyter.

Kim, Mira. 2003. *Analysis of Translation Errors Based on Systemic Functional Grammar: Application of Text Analysis in English/Korean Translation Pedagogy.* Minor dissertation, Macquarie University, Sydney.

Kim, Mira. 2007a. "Translation Error Analysis: A Systemic Functional Grammar Approach." In *Across Boundaries: International Perspectives on Translation Studies*, Dorothy Kenny and Kyongjoo Ryou (eds), 161–175. Newcastle upon Tyne: Cambridge Scholars Publishing.

Kim, Mira. 2007b. "Using Systemic Functional Text Analysis for Translator Education: An Illustration with a Focus on the Textual Meaning." *Interpreter and Translator Trainer* 1 (2): 223–246.

Kim, Mira. 2007c. *A Discourse Based Study on THEME in Korean and Textual Meaning in Translation*. Unpublished Doctoral Dissertation, Macquarie University, Sydney.

Kim, Mira. 2008. "Readability Analysis of Community Translation: A Systemic Functional Approach." *FORUM: International Journal of Interpretation and Translation* 6 (1): 105–134.

Malmkjaer, Kirsten. 1998. "Linguistics in Functionland and Through the Front Door: A Response to Hans G. Hönig." In *Translation and Quality*, Christina Schäffner (ed.), 70–74. Clevedon: Multilingual Matters.

Malinowski, Branislav. 1935. *Coral Gardens and Their Magic*. London: Allen & Unwin.

Martin, James. R. 1992. *English Text: System and Structure*. Amsterdam/Philadelphia: John Benjamins.

Martin, James. R. 2001. "Language, Register and Genre." In *Analysing English in a Global Context*, Anne Burns and Caroline Coffin (eds), 149–166. London: Routledge.

Matthiessen, Christian. M. C. 2001. "The Environments of Translation." In *Exploring Translation and Multilingual Text Production: Beyond Content*, E. Steiner and C. Yallop (eds), 41–124. Berlin: Mouton de Gruyter.

Mohan, Bernard and Slater, Tammy. 2004. "The Evaluation of Casual Discourse and Language as a Resource for Meaning." In *Language, Education and Discourse*, J. A. Foley (ed.), 255–269. London and New York: Continuum.

Munday, Jeremy. 1997. *Systems in Translation: A Computer-Assisted Systemic Analysis of the Translation of Garcia Marquez*. Unpublished Doctoral Dissertation, University of Bradford, Bradford.

Newmark, Peter. 1987. "The Use of Systemic Linguistics in Translation Analysis and Criticism." In *Language Topics: Essays in Honor of Michael Halliday*, R. Steele and T. Threadgold (eds), 293–304. Amsterdam/Philadelphia: John Benjamins.

Nord, Christiane 1997. *Translating as a Purposeful Activity*. Manchester: St. Jerome.

Reiss, Katharina and Vermeer, H. J. 1984. *Grundlegung einer Allgemeinen Translationstheorie*. Tübingen: Niemeyer.

Sager, Juan C. 1989. „Quality and Standards – the Evaluation of Translations.“ In *The Translator's Handbook* (2nd edition), Catriona Picken (ed.), 91–102. London: Aslib.

Snell-Hornby, Mary. 1992. "The professional Translator of Tomorrow: Language Specialist or All-round Expert?" In *Teaching Translation and Interpreting: Training, Talent and Experience*, C. Dollerup and A. Loddegaard (eds), 9–22. Amsterdam: John Benjamins.

Steiner, Erich. 2002. "Grammatical Metaphor in Translation – Some Methods for Corpus-based Investigation". In *Informational Structure in a Cross-Linguistic Perspective,* H. Hasselgard, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds), 213–228. Amsterdam: Rodopi.

Steiner, Erich. 2004. *Translated Texts: Properties, Variants, Evaluations*. Frankfurt/M.: Peter Lang Verlag.

Steiner, Erich. 2005. "Halliday and Translation Theory – Enhancing the Options, Broadening the Range, and Keeping the Ground." In *Continuing Discourse on Language: A Functional Perspective*, Ruqaiya Hasan, Christian M. I. M. Matthiessen and Jonathan Webster (eds). 481–500. London: Equinox.

Trosborg, Anna. 2002. "Discourse Analysis as Part of Translator Training." In *The Role of Discourse Analysis for Translation and in Translator Training*, Christina Schäffner (ed.), 9–52. Clevedon/Buffalo/Toronto/ Sydney: Multilingual Matters.

Appendix 1

S. no.	English source text	Korean target text	E		L		I		T	
			A	N	A	N	A	N	A	N
1	**What Australia's outback deserts lack in water they make up for in uranium.**	**호주 오지 사막에서는 물부족으로 오는 문제는 우라늄으로 해결하고 있다.**	3							
2	**They contain** almost 40% of the world's known low-cost reserves of the nuclear fuel.	호주 사막은 세계에 알려진 저가 핵연료의 40%에 가까운 양을 **비축하고 있다**.		1						
3	It is big business for Australia: exploration companies are at present **spending** ten times more **money** searching for deposits than they did three years ago.	우라늄은 호주에게는 큰 사업으로 우라늄 탐사 기업들은 현재 저장된 우라늄을 찾는데 3년 전보다 10배에 가까운 **돈을 쓰고 있다**.						1		
4	And **ore** from Australia's three operating mines **supplies** about a quarter of the world's uranium-oxide exports.	또한 현재 가동 중인 호주의 탄광 세 곳에서 나는 광석은 세계 산화 우라늄 수출량의 4분의 1 가량을 **공급한다**.		1						
5	Until now all this has **gone** to countries that have signed the Nuclear Non-Proliferation Treaty (NPT).	지금까지 모든 우라늄은 핵확산 방지 조약 가입국으로만 **팔려갔다**.						1		
6	**This** ensures, in theory, that they will use it to produce electricity rather than bombs.	**조약을 맺음으로** 이론상으로는 수입된 자원이 폭탄 제조에 쓰이기 보다는 전기 생산에 쓰일 것이라는 것을 보장한다.							2	
7	But on August 16th, Australia's prime minister said he would **lift a ban on selling uranium to India**, *which refuses to sign the NPT, has tested nuclear weapons and does not rule out testing more.*	그러나 지난 8월 16일, 호주 총리는인도에 **우라늄을 수출하는 것을 금지할 수도 있다**는 가능성을 시사했다. *인도는 핵확산 방지 조약에 서명하기를 거부하고, 핵무기를 시험했던 적이 있으며 앞으로의 핵무기 실험 가능성도 배제하지 않고 있다.*	3							

S.	English source text	Korean target text	E		L		I		T	
no.			A	N	A	N	A	N	A	N
8	The sales will be subject to "strict conditions".	**우라늄 판매는** 앞으로 "엄격한 조건"에 한해 이루어질 것이다.							1	
9	India will first have to sign a safeguards agreement with Australia, **guaranteeing that none of its uranium will be diverted to weapons.**	인도는 우선 호주와 안전 보장 협정을 맺어야 하며, **협정은** 인도로 판매된 우라늄이 일정 무기로 변환되지 않을 것임을 **보장하는 것이다**.			2					
10	Howard first **flagged** the change of Australia's nuclear policy during a visit to New Delhi in early 2006.	하워드 총리가 호주의 핵 정책에 처음으로 변화를 **가져온 것은** 2006년 초 뉴델리 방문 기간 중이다.	2							
11	Australia is also **keen** to build a solid regional relationship with India similar to those it already has with Japan and China.	호주는 또한 이미 호주와 굳건한 관계를 맺고 있는 일본, 중국과 마찬가지로 인도와도 탄탄한 지역 관계 맺는데 **관심이 있다**.					1			
12	Relations with India soured after Australia strongly criticised its nuclear weapons test in 1998.	호주와 인도의 관계는 1998년 인도가 핵무기를 실험한 것에 대해 호주 정부가 강력히 비난한 후로 소원해졌다.								
13	**Uranium mining has always divided Australians**, but more seem to be leaning towards an expansion of the industry in response to global warming.	**우라늄 광산업은 언제나 호주인들을 분열시켰지만** 근래에는 지구 온난화에 대한 관심으로 우라늄 채굴 사업의 규모를 늘리자는 방향으로 의견이 모아지고 있는 듯 하다.		2						
14	However, should India test another bomb, public outrage **would** kill uranium exports in a flash.	그렇지만 만일 인도가 또 다시 핵무기 실험을 하게 되면, 호주 대중의 노여움을 사게 되어 인도로의 우라늄 수출은 그 즉시 중단**될 지도 모른다**.					2			

Assessing cohesion

Developing assessment tools on the basis of comparable corpora

Brian James Baer and Tatyana Bystrova-McIntyre

Translation scholars have long noted that assessment, a key component in translator training, is performed in a generally arbitrary manner. The use of corpora to document differences within language pairs, however, can provide an empirical basis for the formulation of assessment tools. The present study, based on data collected from Russian and English comparable corpora organized by text type, offers a case study in the development of an assessment tool designed to evaluate three isolatable, but nevertheless frequently ignored, features of textual cohesion: punctuation, sentencing, and paragraphing. Because novice translators tend to focus on the level of the word or phrase, ignoring textual elements occurring above the sentence level, focusing assessment on such textual elements can encourage novices to consider the target text globally, as a professional product composed of various features above and beyond lexis. A granular tool for assessing punctuation in Russian>English texts is provided, which can be replicated for other language pairs as well.

Introduction

Katharina Reiss noted in 1971 that in translation assessment "the standards most observed by critics are generally arbitrary, so that their pronouncements do not reflect a solid appreciation of the translation process" (2000: xi). Unfortunately, relatively little has changed despite the unprecedented increase in translation training programs throughout the world. As Colina pointed out in 2003: "translation evaluation and testing today is done on an asymptomatic basis… Furthermore, the numerous translation textbooks on the market rarely devote any time to a systematic study of error evaluation and grading guidelines" (2003: 128). In addition to the specific problems with translation test and evaluation methods described by Nord (1991) and Hatim and Mason (1997), which concern the pedagogy of assessment (i.e., how the assessment task is conceived, presented, and evaluated), many evaluators base their assessments on their own subjective, anecdotal experience with the language pair in question, or on the incorporation of dominant translation norms, which may differ significantly from the current stylistic norms of the target language (Toury 1999: 204).

The same holds true for assessments of translator performance that occur in the context of translation research. As Rui Rothe-Neves points out:

> …the researcher has to rely on what is known about translation norms, and that requires appropriate empirical investigations *prior* to error counting. Typically, however, there is no such concern; the researcher's knowledge and experience as a translator or as a translation teacher will be used to provide the parameters against which to assess translation *well-formedness* (as e.g., Nord 1999). Thence the objective of every empirical research is subverted, that is, to generate theory grounded on collected data. The problem here is that data is not based on tenable parameters. (2007: 133)

In other words, we need more objective data on both linguistic and stylistic norms of the given languages in order to develop more objective assessment criteria. As Reiss puts it, "From a pedagogical perspective…the development of objective methods of evaluating translations would have advantages, because it would be an excellent and even more attractive way of honing an awareness of language and of expanding the critic's linguistic and extra-linguistic horizons" (2000: xi).

Corpora studies present a powerful instrument for analyzing linguistic differences and stylistic norms in an objective and penetrating way. By taking copious, systematic measurements across a wide field of real-world writing, these studies support assessments based on empirical data rather than assumptions. They highlight more or less subtle differences, such as punctuation use, between language pairs, which in turn allows translators to create more natural-sounding translations and better communicate the original writers' intents. They may also challenge the personal preferences of "seasoned" translators, and improve the speed and efficacy of translator training, a key component of which is assessment.

Using corpora in translation assessment

The rapid development of electronic resources and computer programming in the last few decades has had a considerable impact on different areas of linguistics. The ability to store, retrieve, and analyze large collections of texts enables researchers to do things that were not considered possible before. Often defined as a "collection of electronic texts assembled according to explicit design criteria," a corpus is typically compiled with the goal of "representing a larger textual population" (Zanettin 2002: 11), and provides linguists with a "much more solid empirical basis than had ever been previously available" (Granger 2003: 18). Olohan (2004: 1) notes in the introduction to her book *Introducing Corpora in Translation Studies*, referencing Graeme Kennedy (1998: 1), that, while linguistics has used electronic corpora for more than three decades, "the use of corpora […] in translation studies has a short history, spanning no more than ten years." Therefore, the use of corpora in translation studies and translation practice represents a fruitful area of development in translation-related areas, including testing and assessment.

Translation evaluation has been a debated issue in translation studies and practice (cf., House 1997; Nord 1991; Reiss 2000; Schäffner 1998; Williams 2001). Subjectivity in evaluating translations has been among the most often cited criticisms of the process of evaluating translations in various areas (cf. Faharzad 1992; Hönig 1998; Horguelin 1985). The difficulty, as Bowker (2001: 184) points out, lies in developing objective evaluation methods. The use of corpora in this process can make "the task of translation evaluation somewhat less difficult by removing a great deal of the subjectivity" because it provides an evaluator with "a wide range of authentic and suitable texts" to verify the choices evaluators come across when assessing translations (Bowker 2001: 345; Bowker 2000b: 184).

In the area of translation pedagogy, productive applications of a corpus-based approach to translation evaluation include assessing students' translations in general (Bowker 2001; Bowker 2003), recording and analyzing students' errors (Uzar 2003), and developing students' self-assessment and peer-assessment skills (Bowker 2000b; Bowker 2003; Uzar 2004; López-Rodríguez et al. 2007). Bowker, who has devoted a great deal of attention to the use of corpora in translation pedagogy, suggests that a corpus-based approach is a practical and objective approach to translation evaluation, especially for specialized translation assignments in areas in which translator trainers are not actively familiar (2001: 345). Translator trainers can use a corpus as "a benchmark against which [they] can compare students' translations on a number of different levels" (2001: 245). In support of Bowker's point, Uzar (2004: 159) notes that a corpus-based approach can help translation evaluators fine-tune their assessment by comparing several translations of the same texts and pinpointing "what is adequate, appropriate, inadequate or inappropriate." Uzar also mentions that such an approach can help evaluators with the categorization of error types.

As Bowker points out, the valuable characteristics of a corpus-based approach lie in its broad scope, the authenticity of texts, and in the fact that data are in machine-readable form. Moreover, the availability of computational tools and methods for both quantitative and qualitative analyses (e.g., concordances or frequencies) allow for more empirical and, thus, more objective research (2001: 346). Bowker notes that using corpora in evaluating students' translations can help trainers offer more objective and constructive feedback to their students (2001: 361). In addition, corpora provide "a common evaluative framework" for both the student and the evaluator (2001: 361), and make the students more receptive to the trainer's feedback because "they can see for themselves that it is based on corpus evidence and not merely on the subjective impressions or incomplete understanding of the translator trainer" (Bowker 2003: 180). Of course research is needed to determine whether students are indeed more receptive to such feedback.

Bowker's evaluation corpus consists of four types of sub-corpora (2001: 350–354):

- Comparable source corpus, which is a corpus of comparable source language texts used to gauge "normality" of the source text (ST); it is an optional part of Bowker's evaluation corpus.
- Quality corpus, which is a hand-picked corpus aimed at helping translation evaluators to understand the subject matter.
- Quantity corpus, which is a larger corpus that provides a more reliable quantitative analysis.
- Inappropriate corpus, which is a corpus of texts that are not appropriate for the given translation assignment based on such parameters as their production date, generality or specificity of a domain, etc. The inappropriate corpus is used to determine possible sources of students' errors.

While an evaluation corpus may be time-consuming to compile, it has clear advantages, and may be considered a potentially promising tool for inclusion into translator training curricula. Students may themselves be responsible for compiling corpora. The use of corpora in a translation classroom can "raise students' interest in and awareness of specialized language" and thus contribute to their becoming independent learners (Bowker 2001: 362). In this article, however, we focus on the advantages and challenges of using corpora as the basis for the design of an assessment tool to evaluate textual cohesion.

Analyzing textual cohesion

Since the "textual turn" in Translation Studies, translation scholars and trainers have recognized global textual features, such as cohesion, to be of central importance (Neubert & Shreve 1992) for it is cohesion that creates "text" out of individual sentences. Moreover, studies documenting translations done by novices and experts point to cohesion as a fundamental distinguishing trait. Because novices tend to translate at the level of word, phrase, and sentence, their translations often lack cohesion – which is by definition an inter-sentential quality – and so appear awkward and unfocused. Nevertheless, few studies have been done to isolate cohesive features of translation. As Le notes, "How coherence works [at the macro level, i.e., between units larger than the sentence] is generally overlooked and never demonstrated" (Le 2004: 260). Not surprisingly then, until only recently, the American Translators Association (ATA) error marking framework had no explicit category of cohesion and few other categories that explicitly recognized phenomena above the level of the sentence (American Translators Association website). And so, while we know that expert translator behavior is marked by a global, top-down approach to text creation (Jääskeläinen 1990; Tirkkonen-Condit and Jääskeläinen 1991; Kussmaul 1995), we often conduct assessments in a bottom-up fashion, concentrating, like novice translators themselves, on the sub-sentential level.

This situation can in part be explained by the fact that the qualities that constitute cohesion are generally difficult to pinpoint and isolate. As Wilson (Wilson, quoted in Callow 1974: 10–11) noted regarding the first translation of the bible into Dagbani: "For a native speaker it was difficult to express what was wrong with the earlier version, except that it was 'foreign.' Now, however, a comparison … has made clear that what the older version mainly suffers from are considerable deficiencies in 'discourse structure,' i.e., in the way the sentences are combined into well-integrated paragraphs, and these in turn into a well-constructed whole." Moreover, the construction of cohesion in translated texts may be complicated, as Mona Baker (1992: 125) points out, by a tension between syntax and thematic patterning, requiring recasting not for the sake of semantics, understood in a limited sense, but for the sake of cohesion. And so, Le notes, "Translators are torn between the apparent need to respect sentence and paragraph boundaries and the risk of sounding unnatural in the target language" (2004: 267).

Based on data collected from Russian and English comparable corpora organized by text type, we examine the feasibility of an assessment tool that treats three easily isolatable – but nevertheless frequently ignored – features of textual cohesion: punctuation, sentencing, and paragraphing. Since assessments impact learning priorities in academic and professional settings, an assessment tool that focuses on these important, though often-overlooked, textual features, encourages novice translators to consider the target text globally, as a product involving a variety of features above and beyond lexis, for which they are professionally responsible. Specifically, it can serve as an introduction to global aspects of text, such as discourse organization and textual cohesion.

In addition, because punctuation, sentencing and paragraphing can be studied with an untagged corpus (i.e., a corpus of raw texts, not tagged with additional linguistic information), employing rather simple statistical analyses, translator trainers can have their students themselves design an assessment tool based on empirical data collected from bilingual corpora. This process, including the presentation and discussion of the proposed tools, is a fairly simple way to sensitize students to the concept that the most reliable assessment criteria are not the sole possession of the expert translator-cum-trainer. They are based on empirical data, which are available to all those willing to collect and to analyze it. Moreover, once students have designed corpora to study punctuation and segmenting, they can then use those corpora to investigate other linguistic and textual phenomena of interest to them and of special significance in their language pair(s). In this way, introducing corpora studies into the process of translator training may produce a generation of translators who can also generate empirical data – a powerful skill that may speed their progress from journeymen to master translators, while in the process helping to bridge the age-old gap between theory and practice.

Methods

Designing the corpora

This study involved the analysis of punctuation, sentencing, and paragraphing. For our study of punctuation, two untagged comparable corpora of Russian and English editorials were used for the analysis. The editorials for the corpora were taken from leading daily Russian and American newspapers – *Izvestia* and *The New York Times.* Both newspapers provide free on-line access to their issues, which made it much easier to compile the corpora. All the editorials were published in 2005 and thus represent a recent state of events in both languages. The editorials for the corpora were selected randomly, regardless of their content. Each corpus consists of 20,000 words (titles and names of the authors are not included in the word count). The similarity of the two corpora contributed to the overall reliability of the study.

Since the corpus compilation was performed manually and with limited time and resources, the length of each corpus is only 20,000 words. It should be noted that the statistical results would have been more reliable if larger corpora had been used. A further investigation of punctuation use in larger English and Russian corpora is therefore desirable, although a 20,000 word corpus should reflect the general tendencies in a given text type, particularly in regard to punctuation marks, which of course appear in every sentence.

For the study of sentencing and paragraphing, our corpus was developed further, with new text-types added. In addition to editorials, international news articles (from the same newspapers, *The New York Times* and *Izvestia*) and contemporary literary texts were included as additional corpora. All the editorials and international news articles were published in 2005–2006. The articles were selected randomly, regardless of their content. The same holds true for the literary corpora. The publication dates of the randomly selected excerpts of books available online range from 2004 to 2006. The number of different texts in each category was 20 (i.e., 120 texts, totaling 116,140 words). The increased number of text types improves the reliability of the data elicited from the corpus.

Data: Punctuation marks

The study involved the systematic counting of all the punctuation marks per 1,000 words, including commas, colons, semicolons, dashes/em-dashes, hyphens/n-dashes, and parentheses. With respect to commas, which may also be used to represent numbers in writing, it was decided to count both the total number of commas as well as the number of non-numerical commas by subtracting the number of numerical commas from the total number of commas. (Numerical commas are commas used to represent numbers in writing [e.g., Russian: 13,4%; English: $70,000]).

A number of statistical methods have been applied to the collected data. Descriptive statistics were obtained to analyze the overall characteristics of punctuation usage in the English and Russian samples. Significance testing was performed to identify statistically significant differences between English and Russian punctuation patterns.

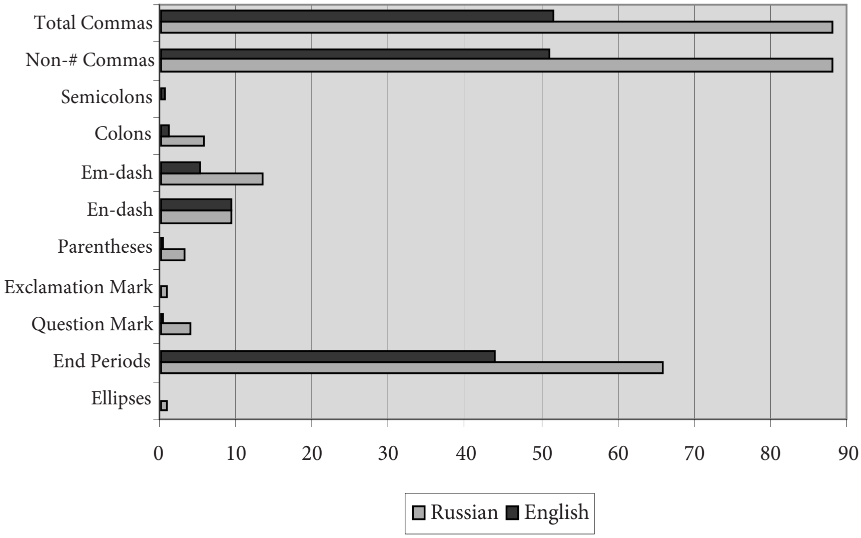
It should be noted that the present quantitative study looks into punctuation usage in only one text type (editorials) and so may not be considered valid for other text types in English and Russian. The genre of the editorial has particular characteristics due to its nature as a commentary or opinion, not always written by professional reporters. Still, common features of Russian and American English editorials justify a comparative study of punctuation usage in this text type. In the future, comparing the variation of punctuation usage among different text types, as well as deriving an average for punctuation usage in a larger sample containing multiple written text types, would be desirable. In addition, using a larger corpus would improve the external reliability of the study. Combined with the comparison of Russian and English style-guides, this study, however, should provide sufficient grounds for suggesting certain strategies for translating Russian punctuation into English and designing an assessment tool for punctuation use in translations.

Results: Punctuation

In order to investigate the differences in the use of punctuation marks in English and Russian, we conducted a comparative quantitative analysis of punctuation. The general account of the results for the comparative analysis of Russian and English punctuation is presented in Graph 1. The graph is the summary of the average use, per 1,000 words, for all punctuation marks, including end punctuation marks. The differences in the use of commas, colons, em-dashes, and parentheses are noticeable from the graph. With respect to end punctuation, we can see that Russian, on average, uses non-period punctuation marks, such as exclamation marks, question marks, and ellipses, more frequently.

The results of the descriptive statistical analysis of the use of commas, semicolons, colons, dashes/em-dashes, hyphens/n-dashes, and parentheses are presented in Table 1. The results for commas are given for the total number of commas in both corpora and for the number of non-numerical commas.

On average, the Russian editorials used 88.00 total commas per thousand words vs. only 51.40 total commas for the English editorials, a difference significant at p < .001; in the case of non-numerical commas, the numbers were 87.90 vs. 50.75, respectively (p < .001). In Russian texts, the average use of colons per 1,000 words was 5.70 vs. 1.00 in English, with p < .001. The average use of dashes/em-dashes per 1,000 words was 13.55 vs. 5.15 in Russian and English, respectively, a difference significant at p < .001. No significant difference was found in the use of semicolons and n-dashes in the two corpora.



Graph 1. English vs. Russian punctuation: Average use per 1,000 words

Table 1. The average numbers of commas, semicolons, colons, dashes/em-dashes, hyphens/n-dashes, and parentheses (per 1,000 words), their standard deviations, and the results of significance testing for the corresponding populations

	Russian (SD)	English (SD)	T-test
Total Commas	88.00 (13.53)	51.40 (10.32)	p < .001
Non-# Commas	87.90 (13.59)	50.75 (10.24)	p < .001
Semicolons	0.30 (0.66)	0.65 (0.81)	Ns
Colons	5.70 (3.50)	1.00 (0.97)	p < .001
Dashes/Em-Dashes	13.55 (6.11)	5.15 (3.27)	p < .001
Hyphens/En-Dashes	9.40 (3.78)	9.25 (3.37)	Ns
Parentheses	3.15 (3.25)	0.5 (0.76)	p < .01

Results: Punctuation

The results of the analysis reveal that the use of commas, colons, dashes/em-dashes, and parentheses in the Russian editorials occurs with significantly greater frequency than in the English editorials, while the use of semicolons and hyphens/n-dashes is not significantly different. Significant differences in the use of commas, colons, dashes/em-dashes, and parentheses in English and Russian imply different grammatical and stylistic principles underlying the use of punctuation in those languages, and so would support the development of more nuanced strategies for translating punctuation than simply preserving the ST punctuation in the TT, as well as the development of a discrete-item assessment for translating punctuation. The more frequent use of commas, colons, and dashes/em-dashes in Russian, if preserved in English, may seem inappropriate to an English reader, and may contradict the punctuation norms of the English language, producing to a greater or lesser extent the disorienting effect of "translationese."

While it seems natural that translators should take into account the norms of the target language (TL) when translating linguistic features of the ST, this is often not the case when it comes to punctuation. As Ishenko (1998: 155) points out, translators often "tend to automatically copy any graphic features" of the ST to the target text (TT). In fact, Schwartz suggests that translators treat punctuation marks as "false grammatical cognates" (2006: 93). Consider, for example, a selection from Hugh Alpin's translation of Ivan Turgenev's short story "Faust":

> Отречение, отречение постоянное - вот ее тайный смысл, ее разгадка: не исполнение любимых мыслей и мечтаний, как бы они возвышенны ни были, - исполнение долга, вот о чем следует заботиться человеку; не наложив на себя цепей, железных цепей долга, не может он дойти, не падая,

> до конца своего поприща; а в молодости мы думаем: чем свободнее, тем лучше, тем дальше уйдешь. (Ivan Turgenev, "Faust" 1856)

> Renunciation, constant renunciation—that is its secret meaning, its solution: not the fulfillment of cherished ideas and dreams, no matter how exalted they might be—the fulfillment of his duty, that is what ought to concern a man; unless he has put chains upon himself, the iron chains of duty, he cannot reach the end of his life's journey without falling; whereas in our youth we think: the freer, the better; the further you'll go. (Turgenev 2003)

The direct borrowing of Russian punctuation suggests a failure to distinguish between a more or less neutral, norm-governed use of punctuation and a more individual, artistically-motivated use of punctuation. Anecdotal evidence suggests that novice translators, who tend to stick closely to the syntax of the ST, are especially unable to make this distinction, and so simply borrow the Russian punctuation.

The results of our analysis reinforce the idea that the norms of the TL should be taken into account when dealing with the translation of punctuation from Russian into English. Special attention should be paid to the use of commas, colons, dashes/em-dashes, and parentheses. When devising strategies for translating these punctuation marks from Russian into English, it is important to note that these punctuation marks are used with statistically greater frequently in Russian than in English. In addition, the linguistic function of these punctuation marks appears to be different in the two languages. A separate study is required to isolate the concrete differences in the linguistic nature of Russian and English commas, colons, and dashes/em-dashes. In any case, as this quantitative comparative research shows, punctuation marks are used in editorials with greater frequency in Russian than in English.

In order to determine the extent to which these results are specific to the text type of the editorial, descriptive statistics were calculated for selected literary texts and editorial corpora of comparable length (6,773 words). For this preliminary analysis, works by two contemporary authors – Tatyana Tolstaya ("Perevodnye kartinki," 2001) and John Updike (*Seek My Face*, 2002) – were chosen. Both authors are respected as stylists, but are not considered avant-garde. The following table compares the use of punctuation in Tatyana Tolstaya's essay and the excerpt from Updike's novel *Seek My Face* (the excerpt from Updike's novel is of comparable length). The descriptive statistics presented in Table 2 show that Tolstaya tends to use such punctuation marks as commas, colons, and em-dashes with much greater frequency.

To see if there is a difference in the frequency of punctuation marks between different text types in Russian, we compared the use of punctuation marks in Tatyana Tolstaya's essay *Perevodnye kartinki* (6,773 words) to the corpus of randomly

Table 2. Use of punctuation marks in Russian and English literary corpora of comparable length

	Tolstaya	**Updike**
Commas	1,073	667
Non-# Commas	1,073	667
Semicolons	27	26
Colons	84	12
Dashes / Em-dashes	180	56
Hyphens / En-dashes	86	92

Table 3. Use of punctuation marks in Russian literary and newspaper corpora of comparable length

	Tolstaya	***Izvestia***
Total Commas	1,073	627
Non-# Commas	1,073	626
Semicolons	27	1
Colons	84	17
Dashes / Em-dashes	180	82
Hyphens / En-dashes	86	52

selected *Izvestia* editorials of the same length. The numbers for Tolstaya were generally much higher. The results are presented in Table 3.

Next, the frequency of punctuation marks in different text types in English was compared. The descriptive statistics for the excerpt from Updike's novel *Seek My Face* (6,773 words) and the corpus of randomly selected *New York Times* editorials (6,773 words) are presented in Table 4. As with Tolstaya, the numbers for Updike tend to be much higher for all punctuation marks.[1]

This quantitative comparative research using these relatively small corpora suggests that punctuation marks are used with greater frequency in Russian than in English in two text types – editorials and literary texts. Also, punctuation marks are used with greater frequency in the selected literary texts than in the newspaper editorials in both Russian and English. Tests of statistical significance were, however, not performed due to the sample size of n = 2. Further statistical analysis of larger corpora is required.

[1] This also reflects the general tendency to use "open punctuation" in English-language newspapers, eliminating all unnecessary punctuation, in particular, commas, and "closed punctuation" in literary publications.

Table 4. Use of punctuation marks in English literary and newspaper corpora of comparable length

	Updike	***NYT***
Total Commas	667	312
Non-# Commas	667	309
Semicolons	26	7
Colons	12	8
Em-dashes	56	26

Table 5. Comparative study: Sentence and paragraph length (in words) across text-types (word/sent – average number of words per sentence, char/sent – average number of characters per sentence, words/para – average number of words per paragraph, char/para – average number of characters per paragraph)

	Editorials		**Literary**		**International News**	
	English	**Russian**	**English**	**Russian**	**English**	**Russian**
Words/sent	**25** (3)	**16** (4)	**20** (6)	**16** (5)	**25** (4)	**15** (3)
Char/sent	**149** (20)	**112** (33)	111 (36)	108 (35)	**149** (25)	**108** (37)
Words/para	73 (15)	72 (20)	113 (97)	73 (33)	**46** (7)	**60** (14)
Char/para	**443** (83)	**508** (138)	625 (525)	484 (215)	**278** (42)	**431** (90)

Data: Sentencing and paragraphing

In our study of sentencing and paragraphing, we looked at sentence and paragraph length in our corpora composed of English and Russian editorials, literary texts, and international news articles described above. Table 5 summarizes the averages and standard deviations for the given groups (the items in bold are statistically significant).

Results: Sentencing and paragraphing

We can see that the average number of words per sentence is significantly higher for English for all three text types. This challenges the stereotype that Russian use longer sentences. The average number of characters per sentence is significantly higher for English editorials and English international news, which, again, challenges the stereotype of long sentences in Russian. The number of characters may be more representative of length, since Russian is a more synthetic language, and English, a more analytical one.

For paragraphs, the words-per-paragraph count was significantly higher for Russian international news. English international news reports revealed a tendency for more concise, focused paragraphs, often 1–2 sentences long. For literary texts, the result was not statistically significant (high SD caused by several authors having extremely long paragraphs (e.g., Updike, 728 words per paragraph)).

In terms of characters per paragraph, Russian paragraphs were found to be 1.6 times longer than English ones. Especially striking are the differences across the text-type sub-corpora, particularly for English, with English international news reports having the shortest paragraphs, and literary texts, the longest. The analysis of standard deviation (which can tell us how spread out the examples in a set are from the mean) showed that the length of sentences and paragraphs in literary texts, both in English and Russian, have the highest standard deviations, compared to editorials and international news, which is of course not surprising.

Since the analysis of our bilingual corpus showed differences in sentence and paragraph length across text-types and languages, a translator trainer may calculate averages (and, if desired, standard deviations) for their students' translations of the same text, and compare the results among the group, looking for things that stand out or appear similar. Modern tools (e.g., Excel, Access, SAS) can support much of this work, even eliminating the need to know the formulas for calculating these statistical measures. However, since only few conclusive remarks can be made based on the corpus analyses described above (e.g., that literary texts tend to have a higher variance in sentence length, or that Russian editorials are less consistent in terms of their sentence length), developing a specific assessment tool based on these results is problematic and, perhaps, unnecessary.

Discussion

The findings of our comparison of corpora of English and Russian editorials to corpora of English and Russian literary texts suggest that, when discussing the use of punctuation in general, it is important to keep in mind that there are two general categories of punctuation:

- Conventional, i.e., obligatory or prescribed by the accepted norms of the language. For example, according to English norms, a comma is needed after a lengthy modifying phrase or clause at the beginning of a sentence: "Some thirty-eight years later, the book bobbed up again in my life when Macmillan commissioned me to revise it for the college market and the general trade" (from E. B. White, "Introduction" to the 1979 edition of *The Elements of Style*).

– Emphatic, i.e., grammatically unnecessary or even inappropriate but characteristic of a writer's personal style. In such cases, the writer might overuse, misuse, challenge, or violate the existing stylistic norms in order to draw the reader's attention to something otherwise unemphasized, to create rhythm in prose or poetry, to structure the focus of the sentence, or to create a certain visual representation from a text, etc. For example, May cites a beautiful example of such creative use of punctuation by Abram Tertz (also known by his pseudonym, Andrei Sinyavsky) where he uses numerous square brackets in his writing about prison life "as a means of depicting the walls of the cell" (May 1994: 130).

Conventional and emphatic use of punctuation can be understood as the poles on a continuum of punctuation usage, ranging from mandatory, grammar-driven usage to creative usage for stylistic effect; punctuation marks used in the latter way are referred to in Russian as *avtorskie znaki*, or 'authors' punctuation.' Examples of fully norm-governed punctuation would include the use of punctuation with direct and indirect citations, enumerations, etc. At the other end of the continuum, we would have 'author's punctuation marks', which cannot be explained by any language norms (Dziakovich 1999). Most writers, however, use punctuation from across this continuum, combining its norm-governed side with aesthetics, manipulating and exploiting the norms to produce a stylistic effect. The choice of the most appropriate translation strategy will certainly depend on whether the punctuation in question is used conventionally or emphatically.

When punctuation is used emphatically, it is reasonable for the translator to preserve or compensate for the stylistic intention of the author. May complains that "as far as punctuation is concerned, translators assume the role of editor," brushing up and clarifying the text (May 1997: 1). According to May, translators should take an "interpretative or creative approach" to translating punctuation (1997: 10). For a translator, it is essential to recognize and respect the artistic use of punctuation. However, this is sometimes easier said than done. Creative interpretation of punctuation is difficult due to:

1. Interference of editors in the final product resulting from the editors' misunderstandings of the artistic uses of punctuation, as well as their own views on how to make a product successful in the target market (May 1994: 12).
2. Different expectations of the TT readers as to the use of punctuation, which influence their judgments about the quality of a translated work. In other words, readers may interpret creative use of punctuation as a "mistake" on the translator's part. Moreover, insufficient knowledge of the writer's overall style and persona on the part of readers and of translators for that matter may hinder their understanding of his/her stylistic intentions (May 1994: 139).

3. The difficulty in determining whether the punctuation is used conventionally or artistically due to the complex use of punctuation by a given author.
4. Finally, the difficulty in choosing appropriate artistic means in the TL to convey the effect of the original punctuation.

A more granular assessment tool for punctuation can help novices make the distinction between norm-driven and artistically-motivated punctuation and, no less importantly, can ensure that those performing the assessment are themselves aware of the differences in punctuation usage between the languages, something a catch-all category of "punctuation" cannot do.

In order to design a targeted assessment tool based on these findings, it is necessary to understand why the use of certain punctuation marks is statistically greater in Russian than in English. At this point, the results of this study make it possible to suggest that the use of commas, colons, dashes, and parentheses in Russian is, in general, more grammar-driven than in English. English seems to rely more on the concept of "inseparability" of certain segments of meaning. Hannay (1987) notes that inseparability comes into play when two things in question are components of a unit. According to Hannay, a unit may be predicational (a predicate and its arguments), referential (a head noun phrase and any restrictive modification), or message-based (an optional topical segment and a focus). For example, in English, we cannot separate a verb from its object (e.g., He wanted to learn who is going to the party) or a restrictive clause from its head noun (e.g., The girl who brought the cookies is my niece). In Russian, however, a comma would be obligatory in both cases (*Он хотел узнать, кто пойдёт на вечеринку. Девушка, которая принесла печенье, — моя племянница*). In the first case, it would be required since every clause in Russian must be separated from other clauses (unless there is a coordinating conjunction) regardless of the verb-object relationship between them. In the latter case, the same rule applies due to the fact that Russian, strange as it may seem to native speakers of English, does not distinguish between restrictive and non-restrictive clauses. This example may be seen to confirm different principles underlying English and Russian punctuation, with Russian punctuation being more grammar-oriented and English, more style-oriented.

However, further research of occurrences of different punctuation marks in English and Russian is needed to support this idea and to articulate further the reasons for punctuation differences in English and Russian. To that end, we examined authoritative style guides for (American) English and Russian: Strunk and White's *Elements of Style* and *Pravila russkoi orfografii i punktuatsii* (Rules of Russian Orthography and Punctuation), and Sandra Rosengrant's *Russian in Use.* Comparing these style guides revealed differences in the underlying principles for using the colon, the comma, and the em-dash (*тире*[2]) in Russian and English, all of which were found to be more frequent in the Russian editorials analyzed above. The results are presented in Tables 6–8.

In addition, Russian appears to tolerate the use of em-dashes more than English, as suggested by the appearance of multiple em-dashes in a single sentence, as in the example: "*У них есть надобность — хранить железо, а надобности других людей — редколлегия, работа — они видали в гробу.*"

While the frequency of usage of semi-colons and parentheses differed among the various corpora-semi-colons occur with greater frequency in English editorials and parentheses with greater frequency in Russian editorials-we could find no difference in the functions of these punctuation marks. The semi-colon, used in both languages to separate closely-related sentences, may appear with less frequency in the Russian corpus where that function is also performed by the comma (see 1 in Table 8). As for parentheses, it appears that Russian simply exhibits a greater tolerance, or preference, for this means of setting off parenthetical information than does English.

Table 6. Usage of the colon in Russian and English

Russian	**English**
1. To introduce a list	
Прилетели музыканты - тоже все лауреаты: Юрий Башмет и Игорь Бутман со своими коллективами, детский хореографический ансамбль из Чечни "Зия", а еще профессора-итальяноведы, преподаватели русского языка и переводчики.	*Long Islanders can […] choose to de-vote half or all of a monthly electric bill to buying power from marketers that sell energy from renewable sources: wind farms, hydro-electric plants and biomass operations, […].*
1a. To introduce a list (separating a verb from its object)	
Читаем: повар, водитель, охранник, помощник посла.	*— In English, a colon should NOT separate a verb from its object or a preposition from its object.*

[2] The Russian *тире* and *дефис* roughly correspond to the English em-dash and hyphen/en-dash, respectively. Technically, there are three dashes in English: the hyphen, which is the shortest of the three, is used in compound words; the en-dash, which is slightly longer is used to present a numerical range; and the em-dash, which is the longest of the three is described in the chart above.

Table 6. (*continued*)

Russian	**English**
2. After an independent clause when the following clause interprets or amplifies the preceding clause	
Тут все как на футбольном поле: понятно, что надо делать, главное — исполнить.	*Electricity use, in fact, is climbing rapidly on Long Island: it is up more than 20 percent since 1997 […].*
3. When the following clause explains the reason for a state or an action	
У Сергея Иванова объяснение простое: недоработки командиров.	— Not typical in English
4. When the first clause has such perception verbs as *видеть, слышать*, etc. (no conjunction)	
И тысячи юных дурочек искренне верят: стоит укоротить нос, увеличить губы и накачать силиконом грудь, как счастье и им улыбнется и они тоже станут персонажами светских хроник, за которыми день и ночь охотятся папарацци.	— Not typical in English
5. To introduce direct speech	
«[…] вы читали мой текст?» А они говорили: «Нет, не читали, потому что знаем — это греховный рассказ, это театр абсурда».	— Not as common in English: a colon may be used when the quotation SUPPORTS or contributes to the preceding clause *The squalor of the streets reminded him of a line from Oscar Wilde: "We are all in the gutter, but some of us are looking at the stars."* (S&W)
6. After a salutation in a formal letter	
Not typical in Russian, where the exclamation is more frequently used: *Уважаемый господин Президент!*	*Dear Mr. Montague:* (S&W)
7. To separate hour from minute in a notation of time	
10:45 «События. Время московское». *10.45 «События. Время московское».*	*The train departs at 10:48 PM.* (S&W)

Table 7. Usage of the em-dash in Russian and English

Russian	**English**
1. Between the subject and the predicate when both are nouns in the nominative case, unless the predicate is negated, often used together with a comma	
*Россия, по Достоевскому, — страна, созданная для страдания.**	— Not used as the copula in English because English has an explicit verb “to be” in the present tense.
2. Between the subject and the predicate when either is an infinitive	
Быть сотрудником посольства за рубежом — конечно, не самая “народная профессия”.	— Not used as the copula in English because English has an explicit verb “be” in the present tense.
3. Between the subject and the predicate when either is a numeral	
Из 547 тысяч американских сержантов 241,5 тысячи — простые сержанты, 168 тысяч — штабс-сержанты, 100 тысяч — сержанты 1-го класса, 26 тысяч — мастер-сержанты и 10,6 тысячи — сержант-майоры.	— Not used in English
4. Before or after a summarizing word in an enumeration (in this case, the word “это” is often used to link the subject and the predicate in Russian constructions)	
Умный человек сказал: у престижа в современном мире три составляющие — количество нобелевских лауреатов, число спортивных наград и успехи в космонавтике.	*In the meantime, you have to feel sorry for the people in places like Arizona—the residents, the immigrants and the border police.*
5. To join clauses and homogeneous parts of a clause when the following clause or part contains an abrupt opposition or a sudden connection to the preceding part	
А за углом дежурит эвакуатор, “зазеваешься — он хвать и тикать!”	— Acceptable in English although no examples of such usage were found in the English corpora.
6. To join clauses not connected by conjunctions, when the following clause contains the result, summary or explanation of the preceding clause	
В ответах президента звучала серьезная уверенность — темпы экономического роста сохранятся на достаточный отрезок времени и это позволит решать социальные проблемы.	*In this and other ways, the administration is manipulating information—a tacit, yet devastating, acknowledgement, we believe, that an informed public would reject privatizing Social Security.*
7. To join two clauses when a subordinate conjunction is omitted	
Шокировать публику — так уж по полной программе!	— Not used in English

* All the examples in the tables are from the corpus of editorials, unless otherwise indicated

Table 7. (*continued*)

Russian	**English**
8. To mark a parenthetical clause or construction in the middle of a sentence	
В ряду примеров он привел случай, когда "идиот-командир — извините, другого слова не подберу — после бани вывел солдат на улицу и полчаса пересчитывал их на морозе."	*Donations for Acehnese relief from the rest of Indonesia—where Aceh is not popular—have run high.*
9. Emphatic use	
Спасать своих рядовых Райанов должно и государство, и общество. Иначе — не получится, не научимся мы это делать.	*The rebels announced a unilateral ceasefire, but this was not matched by the military—long indifferent to how its actions turn Acehnese citizens against the government.*
10. To mark direct speech	
"Да-да, мы поняли, и лобио больше не будет, кстати, ваши казино закроем", — отвечают в Москве.	— Not used in English

Table 8. Usage of comma in Russian in English

Russian	**English**
1. To join independent clauses with no conjunctions	
Они вышагивают по подиумам, их берут замуж заморские принцы, они тусуются в светских компаниях, а телекамеры охотно и подробно фиксируют эту сладкую жизнь.	In English, a conjunction must be present: *Working together in times of human disaster can help build confidence between the two sides, and foster a feeling of solidarity among ethnic groups.*
2. To join independent clauses with conjunctions и, а, но, и…и, ни…ни, или…или, то…то	
Осень — пора плодово-овощных революций, но всему должно быть хоть какое-то объяснение.	In English, a comma is optional. Comma not used: *Breast cancer was controllable if caught in the early stages but Lynn may have waited too long.* Comma used: *They would have sympathized, but that was not the same thing.*
3. To join main and subordinate clauses	
В "Кредо сержанта," которое каждый сержант армии США знает наизусть, говорится: "Я — сержант, лидер солдат.	In English, a comma is used with NON-restrictive clauses* only

* A non-restrictive clause "is one that does not serve to identify or define the antecedent noun" (Strunk and White, p. 6)

Table 8. (*continued*)

Russian	**English**
4. To separate items in a list, unless they are connected by one conjunction “и” or “или”	
Он снимал Марчелло Мастроянни, Софи Лорен, самого Феллини...	In English, a comma is optional, but may be needed even before the single conjunction “and” or “or.” *“In keeping with his chilling comment that Western democracy is a “blind alley,” Mr. Hu has already made it clear that the government is ready to crack down on journalists, scholars and protesters who cross his unmarked line.” And how lovely it was, a bike ride around the forest preserve on a Sunday in May with our mountain bikes, water bottles, and safety helmets.*
5. To set off comparative constructions (with some exceptions)	
Никогда еще после распада СССР отношение политического истеблишмента США к России не менялось так резко, как в 2004 году.	— Not needed in English *And while I was still looking toward him there was another roll of drums, suddenly silenced, and then the thud of the ax, first once, then again and a third time: a sound as domestic as chopping wood.*
6. To set off parenthetical and descriptive constructions (with some exceptions)	
В сцене прибытия нового комбата именно сержант-майор Пламли, седовласый верзила, командует офицерам батальона “Стройся!”	*The Indonesian province of Aceh and the country of Sri Lanka, united today by the ravage of a tsunami, previously had in common histories of man-made destruction.*
7. To set off parenthetical words (e.g., *например, видимо, таким образом,* etc.)	
Много лет кремлевские обитатели возмущались тем, что Москва подвергается критике за то, что легко сходит с рук, например, Китаю.	*For example, as a teenager in the early 1950’s I belonged to the Peter Pan Magic club, which was sponsored by the Parks Department.*
8. To set off participial clauses and clauses with verbal adverbs	
Учитывая обогащенность вызывающего всеми сокровищами культурного знания, нас ждет впечатляющее зрелище.	Not obligatory unless the participial clause is long (typically seven or more words)
9. To set of a direct address	
Граждане, ау!	Used less often than in Russian, where it is common in personal and business letters

Table 8. (*continued*)

Russian	English
10. To set off interjections	
Граждане, ау!	*Oh, how sad, how perfect.*
11. After confirmations, negations, and question words (*да, конечно, нет, что*, etc.)	
Много лет кремлевские обитатели возмущались тем, что Москва подвергается критике за то, что легко сходит с рук, например, Китаю.	*For example, as a teenager in the early 1950's I belonged to the Peter Pan Magic club, which was sponsored by the Parks Department.*
12. Note: In Russian, a comma is not used to set off the introductory phrase of a sentence	
	In Aceh, where at least 100,000 people have died so far from the tsunami, rebels have fought since 1976 to free the province, which was an independent nation for centuries, from Indonesian rule.

Framework for error marking in punctuation use (R>E)

Once the analysis of the statistical discrepancies has been accomplished, an optimally granular assessment tool can be fashioned to help explain errors to those being assessed and to guide the work of assessors. Note that in the sample assessment instrument developed on the basis of the empirical data gathered (see Table 9), the errors described in relation to each of the punctuation marks are listed in descending order from obligatory usage to more stylistically-driven usage.

Conclusions

It should be noted here that the proposed assessment of punctuation, sentencing, and paragraphing challenges the oppositions that have long organized discussion of translation assessment: global versus discrete-item assessment and formative versus summative assessment. Rather than mutually-exclusive categories, we recognize them as points on a cline. For example, while an assessment such as the ATA Certification exam may be primarily summative, the moment an examinee requests feedback, it becomes the basis of a formative assessment, in that it should provide information that will improve future performance. Similarly, in the translation training classroom, most assessments are both summative – they help to determine the trainee's final grade – and formative – they are designed to help improve the trainee's performance. Along the same lines, it is certainly

Table 9. Sample framework for error marking in punctuation use (R>E)

Colon			Comments/Suggestions:
	–	Used to introduce a list, separating a verb from its object	
	–	Used to introduce a clause that explains the reason for a state or action	
	–	Used when the first clause has a perception verb (such as 'to see' or 'to hear') without a conjunction	
Em-Dash			Comments/Suggestions:
	–	Used instead of hyphen or en-dash	
	–	Used hyphen or en-dash in place of em-dash	
	–	Used to introduce direct speech	
	–	Used between subject and predicate	
	–	Used to join two clauses when a subordinate conjunction is omitted	
	–	Over-used (i.e., multiple em-dashes in a single sentence)	
Comma			Comments:
	–	Used as a decimal point with fractions	
	–	Used to join independent clauses with no conjunction	
	–	Used with a restrictive clause	
	–	Used to set off comparative constructions	
	–	Over-used (used close punctuation in a venue where open punctuation is the norm)	
Parentheses			Comments:
	–	Over-used to set off parenthetical information	
Question Marks			Comments:
	–	Over-used as a rhetorical device (i.e., to introduce explanations)	

useful for trainees to see precisely where a translation error occurred, when it is possible to localize the error. However, it is also important for novice translators to recognize global textual features and to see their translations as professional products – goals more effectively achieved in global assessments. Moreover, this assessment can be part of a larger tool or can be used on its own, for, as Koby and Baer (2005) pointed out, it may be advisable in the course of translator training to focus the attention of novice translators on certain, typically overlooked, textual features by isolating them in an assessment.

Note

We also collected data on the use of end punctuation; the results showed a significantly greater number of end punctuations marks in the Russian corpus, indicating a greater number of sentences. This suggests the need to study the larger issue of segmentation in English and Russian and the relationship of segmentation to cohesion.

References

American Translators Association. [cited 7 Jan. 2009]. Available from http://www.atanet.org/certification/aboutexams_error.php.

Pravila russkoi orfografii i punktuatsi. [cited 5 Sept. 2008]. Available from http://www.gramota.ru/spravka/rules/.

Bowker, Lynne. 2000a. "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources." *International Journal of Corpus Linguistics* 5 (1): 17–52.

——— 2000b. "A Corpus-Based Approach to Evaluating Student Translations." In *Evaluation and Translation. Special Issue of the Translator* 6 (2). Carol Maier (ed.), 183–210.

——— 2001. "Towards a Methodology for a Corpus-Based Approach to Translation Evaluation." *Meta* 46 (2): 345–364.

——— 2003. "Corpus-Based Applications for Translator Training." In *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies.* Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds.), 169–183. Amsterdam & New York, NY: Rodopi.

——— 2004. "Corpus Resources for Translators: Academic Luxury or Professional Necessity?" *TradTerm* 10: 213–247.

Bowker, Lynne and Peter Bennison. 2003. "Student Translation Archive: Design, Development and Application." In *Corpora in Translator Education.* Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), 103–117. Manchester, UK & Northampton, MA: St. Jerome.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Callow, Kathleen. 1974. *Discourse Considerations in Translating the Word of God*. Michigan: Zondervan.

Colina, Sonia. 2003. *Translation Teaching: From Research to the Classroom. A Handbook for Teachers*. Boston: McGraw Hill.

Dziakovich, E. V. 1999. "O Vyrazitelnykh Sredstvakh Punktuatsii." *Russkii iazyk v shkole* September-October: 76–78.

Granger, Sylviane. 2003. "The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies." In *Corpus-Based Approaches to Contrastive Linguistics and Translation Studies.* Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson (eds.), 17–29. Amsterdam & New York, NY: Rodopi.

Halliday, Michael and Ruqaiya Hassan. 1976. *Cohesion in English*. London: Longman.

Hannay, Mike. 1987. "English Comma Placement: A Functional View." In *One Hundred Years of English Studies in Dutch Universities.* G. H. V. Bunt, E. S. Kooper, J. L. Machenzie and D. R. M. Wilkinson (eds.), 81–92. Amsterdam: Rodopi.

Hatim, Basil and Ian Mason. 1990. *The Translator as Communicator*. London and New York: Routledge.

Hönig, Hans G. 1998. "Positions, Power and Practice: Functionalist Approach and Translation Quality Assessment." In *Translation and Quality*, Christina Schäffner (ed.), 6–34. Clevedon: Multilingual Matters.

Horguelin, Paul. 1985. *Pratique de la Révision*. Montréal: Linguatech.

House, Juliane. 1997. *Translation Quality Assessment: A Model Revisited*. Tübingen: Gunter Narr Verlag.

Ishenko, Michael. 1998. "Translating Punctuation Marks: Punctuating and Formatting Issues in English-Russian Translation." *Proceedings of the 39th Annual Conference of the American Translators Association*: 155–174.

Jääskeläinen, Riita. 1990. *Features of Successful Translation Processes: A Think-Aloud Protocol Study.* Licentiate Thesis. University of Joensuu, Savolinna School of Translation Studies.

Kussmaul, Paul. 1995. *Training the Translator.* Philadelphia and Amsterdam: John Benjamins.

Le, Elisabeth. 2004. "The Role of Paragraphs in the Construction of Coherence – Text Linguistics and Translation Studies." *International Review of Applied Linguistics in Language Teaching* 42 (3): 259–275.

López-Rodríguez, Clara Inés, Bryan J. Robinson and María Isabel Tercedor-Sánchez. 2007. "A Learner-Generated Corpus to Direct Learner-Centered Courses." In *Translation and Meaning.* Marcel Thelen and Barbara Lewandowska Tomaszczyk (eds.), 197–211. Maastricht: Zuyd University.

López-Rodríguez, Clara Inés and María Isabel Tercedor-Sánchez. 2008. "Corpora and Students' Autonomy in Scientific and Technical Translation Training." *The Journal of Specialized Translation* 9: 2–19.

May, Rachel. 1994. *The Translator in the Text: On Reading Russian Literature in English*. Evanston, IL: Northwestern University Press.

——— 1997. "Sensible Elocution: How Translation Works in and Upon Punctuation." *The Translator* 3 (1): 1–20.

Neubert, Albrecht and Gregory M. Shreve. 1992. *Translation as Text*. Kent, OH: Kent State University.

Nord, Christiane. 1991. *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis.* Amsterdam&Atlanta: Rodopi.

——— 1999. "Translating as a Text-Production Activity." *On-Line Symposium on Innovation in Translator and Interpreter Training*. <http://www. fut.es/~apym/symp/nord.html> (accessed 1/28/2009).

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London; New York: Routledge.

Reiss, Katharina. 2000. *Translation Criticism – The Potentials and Limitations: Categories and Criteria for Translation Quality Assessment*. Trans. Erroll F. Rhodes. Manchester (UK): St. Jerome Publishing.

Rosengrant, Sandra. 2006. *Russian in Use: An Interactive Approach to Advanced Communicative Competence*. New Haven: Yale University Press.

Schäffner, Christina (ed.). 1998. *Translation and Quality*. Philadelphia: Multilingual Matters.

Schwartz, Mirian. 2006. "Marks of Punctuation as False Grammatical Cognates." In *Translating Russia: From Theory to Practice* [Ohio Slavic Papers 8]. Brian James Baer (ed.), 93–102.

Strunk, William and E. B. White. 2005. *The Elements of Style*. New York: The Penguin Press.

Tirkkonen-Condit, Sonja and Riita Jääskeläinen. 1991. "Automatised Processes in Professional vs. Non-Professional Translation: A Think-Alound Protocol Study." In *Empirical Research in Translation and Intercultural Studies.* Tirkonnen-Condit, Sonja, (ed.), 89–109. Tübingen: Gunter Narr.

Toury, Gideon. 1999. "The Nature and Role of Norms in Translation." In *The Translation Studies Reader*. Lawrence Venuti (ed.), 198–211. London: Routledge.

Turgenev, Ivan. 2003. *Faust*. Trans. Hugh Alpin. London: Hesperus.

Uzar, Rafal. 2004. "A Toolbox for Translation Quality Assessment." In *Practical Applications in Language and Computers.* Barbara Lewandowska-Tomaszczyk (ed.). Vol. 9, 153–162. Frankfurt: Peter Lang.

Uzar, Rafal and Jacek Walinski. 2001. "Analyzing the Fluency of Translators." *International Journal of Corpus Linguistics* 6: 155–166.

Williams, Malcolm. 2001. "The Application of Argumentation Theory to Translation Quality Assessment." *Meta* 46 (2): 326–344.

Zanettin, Federico. 2002. [cited 30 Apr. 2008]. "Corpora in Translation Practice." *First International Workshop on Language Resources for Translation Work and Research Proceedings.* Available from http://www.ifi.unizh.ch/cl/yuste/postworkshop/repository/proceedings.pdf (accessed 1/28/2009).

Assessing software localization

For a valid approach

Keiran Dunne

The purpose of this chapter is to begin a critical dialogue on localization quality and assessment of the quality of localized products. Building on the author's previous work on localization quality management, this chapter examines tools and methods currently used in localization quality assessment. Problems inherent in current product-based approaches suggest that localization quality assessment should focus less on localized end products than on the customer's requirements and expectations with regard to such end products. Identifying and documenting client needs, preferences, and expectations in a client quality requirements specification during the project planning phase and measuring compliance with such requirements offers a valid basis on which to empirically measure the quality of localized products.

> There is no point in using exact methods where there is no
> clarity in the concepts and issues to which they are to be applied.
> John Von Neumann and Oskar Morgenstern

> Not everything that can be counted counts,
> and not everything that counts can be counted.
> Albert Einstein

Introduction

The past two decades have witnessed the emergence of localization as a professional business service and its subsequent growth into a multi-billion dollar industry. Nevertheless, "localization remains a little-known and poorly understood phenomenon outside of the relatively closed circle of its clients and practitioners," and even among localization stakeholders, "there exists no consensus as to what precisely constitutes localization" (Dunne 2006a: 1). These observations raise a number of fundamental questions concerning localization quality assessment. If no consensus exists as to what constitutes localization, how can we assess the quality of localized products? In the absence of standards, scholarship, or empirically validated best practices, how can clients, practitioners, and educators find a common framework within which to discuss, evaluate, measure, and improve localization quality?

The goal of this chapter is to provide preliminary answers to these questions and to begin a critical dialogue on localization quality and assessment of the quality of localized products. This chapter will examine tools and methods currently used in the assessment of localized software quality. In so doing, we will explore the notions of quality on which current approaches to localization assessment are implicitly based, as well as the limitations inherent in current approaches. Finally, we will examine possible solutions to some of the issues raised. Because an examination of the assessment of the various types of localization is beyond the scope of this chapter, our discussion will reflect current market conditions. Since the Windows family of products collectively accounts for 90.5% of operating system usage (W3C Schools 2008), and since 87% of companies outsource most or all of their translation and localization work (DePalma and Beninatto 2003: 11), we will focus on the localization of 32-bit Windows desktop applications in the context of outsourced localization projects.

Localization: A process of adaptation

Before undertaking a discussion of software localization quality or quality assessment, we must first define what we mean by the term "localization." Our working definition is as follows:

> The process by which digital content and products developed in one locale (defined in terms of geographical area, language and culture) are adapted for sale and use in another locale. Localization involves: (a) translation of textual content into the language and textual conventions of the target locale; and (b) adaptation of non-textual content (from colors, icons and bitmaps, to packaging, form factors, etc.) as well as input, output and delivery mechanisms to take into account the cultural, technical and regulatory requirements of that locale.
>
> (Dunne 2006a: 4)

The critical point for the purposes of our discussion is that localization is less a process whereby new products are *created* than a process whereby existing products are *adapted*. That being the case, what is – or should be – the scope of localization quality assessment? Should assessment focus on the localized product unto itself, without reference to the original source-language version (i.e., from the perspective of a target-language end-user)? Or should assessment focus on the quality of the adaptation by comparing critical characteristics of the original and localized versions of the product? Or should assessment do both? Can localization be understood and can localization quality be measured without reference to the source materials that serve as input to the localization process?

To contextualize these questions, we will begin by briefly examining what software is, how it is designed and developed, and how its quality is defined and evaluated.

What is software and how is it developed?

Software refers to programs that allow humans to solve real-world problems and achieve real-world objectives using computers. Software development begins with an idea or with "here's what I want" statements (e.g., "I want to be able to use my mouse like a pen to scribble notes or drawings, and save them for future reference"), which are then translated by developers from the natural (human) language in which users express them into a formalized set of instructions in a language that the computer can understand (see Figure 1). The processes whereby developers transform natural-language requests into software code running on a computer are known collectively as the Software Development Life Cycle (SDLC, see Figure 2).

The first and arguably foremost challenge of software engineering is representing the *idea* of the software in the form of a conceptual model that serves

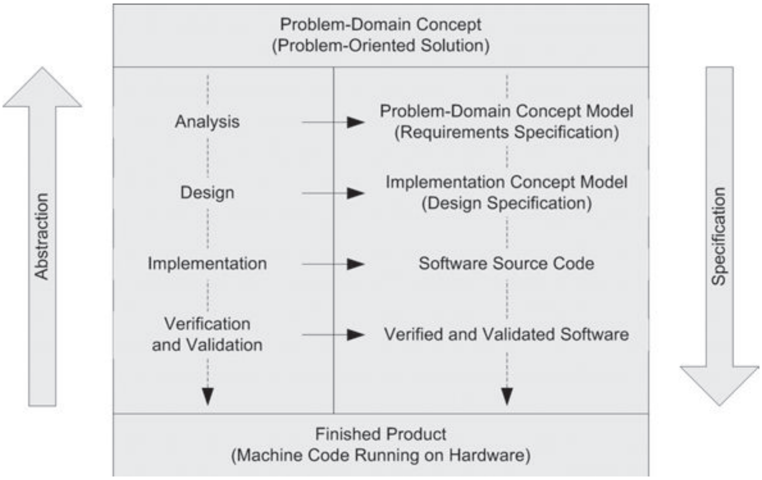

Problem-Domain Concept
(Problem-Oriented Solution)
Abstraction
Analysis
Problem-Domain Concept Model
(Requirements Specification)
Design
Implementation Concept Model
(Design Specification)
Implementation
Software Source Code
Verification
and Validation
Verified and Validated Software
Specification
Finished Product
(Machine Code Running on Hardware)

Figure 1. Software development comprises a set of processes for moving from the level of the abstract to the concrete, with an increasing degree of specification at each step.

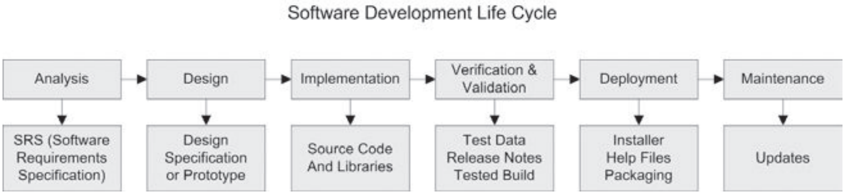

Software Development Life Cycle
Analysis
Design
Implementation
Verification & Validation
Deployment
Maintenance
SRS (Software Requirements Specification)
Design Specification or Prototype
Source Code And Libraries
Test Data Release Notes Tested Build
Installer Help Files Packaging
Updates

Figure 2. A generic model that illustrates the various phases and outputs of the Software Development Life Cycle.

as the foundation for subsequent development efforts. Thus, the goal of the first phase of the SDLC, analysis, is to elicit, analyze and capture users' needs as formal requirements. Analysis defines the context of the specific domain problem that the system should solve. The output of the analysis phase is a document or document-like artifact called the Software Requirements Specification (SRS). The SRS provides the baseline specification against which the conformance of the final product is ultimately assessed. For this reason, it is critically important that the SRS be as clear, complete, and correct as possible. It is also worth noting that when defining project scope, "anything not explicitly included is implicitly excluded" (PMI 2000: 56). Consequently, if the requirements specification of the original product does not address localization, then localization quality is by definition outside the scope of the original project.

Analysis is followed by design. Whereas analysis provides the requirements of what shall be, design provides a structure of how those requirements will be met. During the following phase, implementation, programmers work from the design specification and implement the design in the form of a software program. When the programming has been completed, the application is debugged, and it is subjected to verification and validation testing. Verification is defined as "confirmation, through the provision of objective evidence, that specified requirements have been fulfilled," whereas validation is defined as "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (ISO/IEC 2008: 8). Once the program has been verified and validated, it is installed or delivered to the customer. It then enters the maintenance phase, during which it is modified to meet evolving user needs; episodic upgrades are offered as new functions are demanded or become available.

What is software quality? How is it evaluated and measured?

From the point of view of the developer and of the users who provide the requirements, the quality of a software product is understood as the degree to which a

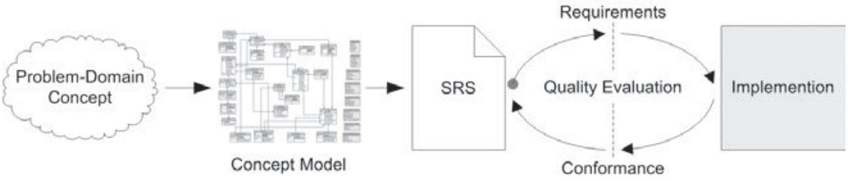


Figure 3. Software quality is assessed during development by testing conformance

program successfully conforms to the initial model of itself, i.e., to the concept that the stakeholders had in mind before the program was developed. Quality is evaluated during the Verification and Validation phase of the SDLC by measuring the degree to which the product satisfies the formalized criteria laid out in the SRS (see Figure 3). Quality measurement thus ultimately reflects the extent to which requirements are accurately and adequately captured, communicated, and expressed in the final product. It is important to remember that these requirements are contextually – and thus *culturally* – bound. In other words, software is both the mirror and product of the culture in which it is created (Marcus 2005). It follows that localization should in theory address the dimensions of culture (Hall and Hall 1990; Hofstede 1991) inherent in the program being localized.

Localized software quality: A problem of perspective

Having examined what software is, how it is designed and developed, and how its quality is evaluated and measured, let us now turn our attention to localized software. There is at present no *de jure* or *de facto* localization process quality or product quality standard. As DePalma and Beninatto have observed, "Neither formal nor ad hoc industry associations have succeeded in developing generally accepted vendor certification metrics, quality assurance standards, generic requests for proposals (RFP), or even standard practices" (2003: 5). Indeed, in the outsourced localization project model, notions of quality tend to vary dramatically depending on when the product is evaluated, who conducts the evaluation, and the criteria (or lack thereof) on which the evaluation is based. To explore the various perspectives on product quality in a typical outsourced localization project, we will examine the localization process, the points in the process at which quality is assessed, and the tools and methods used. We will first consider quality assessment from the perspective of the localization vendor and project teams who perform the localization work. We will then consider quality assessment from the perspective of the client reviewer, i.e., the person or team that evaluates the final product. Finally, we will discuss approaches to assessment in general and propose some possible solutions to issues in current approaches.

Localized software quality and quality assessment: The vendor's (or practitioner's) perspective

In professional practice today, the quality of localized software products is evaluated by performing "quality assurance" or "testing" (Luong et al. 1995: 61–87, 181–195; Urien, Howard and Perinotti 1993: 87–91; Esselink 2000: 145–164; Symmonds 2002: 328–330; Rätzmann and De Young 2003: 239–305; Chandler 2005: 197–233; Lommel and Ray 2007: 24–26; Smith-Ferrier 2007: 487–536). In the literature, as in practice, the terms "quality assurance" and "testing" tend to be used interchangeably to refer to the process of inspection, detection, and correction of defects in the target version of the software with respect to the source version. In the interest of clarity, we will refer to this process as "localization testing."

Localization testing typically focuses on three categories of quality characteristics: linguistic, cosmetic, and functional (Esselink 2000: 150–154; Kirimoto 2005; LISA 2008; Lionbridge Technologies 2009; Tek Translation 2009). The goal of linguistic testing is to ensure that all translatable text has in fact been accurately translated (and that any graphics which require modification have in fact been modified). The goal of cosmetic testing is to ensure that all text and graphics display correctly and completely in the target version of the application. The goal of functional testing is to ensure that localization has not broken anything or introduced any functional problems into the software. Functional testing (ideally) replicates the testing procedure that was performed on the source version of the software, and is designed to test whether "the functionality and feature set of the localized application mirror that of the source" (Esselink 2000: 152).

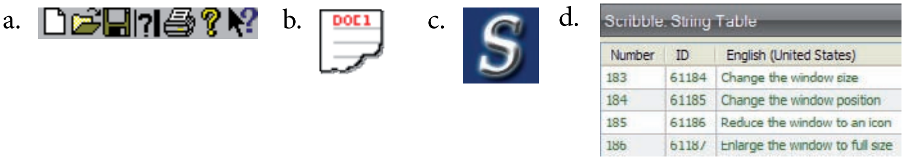
One might wonder on what basis a distinction is (or can be) drawn between linguistic, cosmetic, and functional characteristics in software, in which text is literally embedded in its context. The categorization of localized software quality attributes in terms of linguistic, cosmetic, and functional characteristics reflects the sequence of steps in a software localization project, as well as the type and scope of testing performed at each step of the process.

Software localization processes

A localization project begins with the handoff of source materials from the client to the localization vendor. From the vendor's perspective, the scope of localization is generally limited to the culturally-dependent contents of the graphical user interface (GUI) that may require localization, which are collectively referred to as *resources* (see Figure 4). Resources in a typical desktop application include the following:

- **Bitmaps:** graphic files in bitmap (*.BMP) format, which typically include toolbar button images (see Figure 4a).
- **Icons:** small images that represent, and provide shortcuts to, programs, files, or documents (see Figures 4b and 4c).
- **Menus:** lists of commands or options that display at the top of the main program window.
- **Dialog boxes:** secondary windows that display information and/or request input from the user. Common examples include the "Open," "Save As," and "Print" dialog boxes.
- **String tables:** collections of strings in tabular format. A string is a series of characters, i.e., text that is stored and manipulated as a group (see Figure 4d). Strings can take the form of menu items, command button captions, dialog box title captions, mouseover text, status messages, error messages, and so forth.

In theory, any program that is to be localized should first undergo internationalization, a process whereby all culturally-dependent GUI content is separated from the functional core of the application and stored in resource files, such as satellite DLLs. For example, if an application is internationalized by storing all of the localizable content in a satellite (external) DLL file, then localization merely requires the creation of parallel DLL files to support additional languages and locales. A U.S. English application internationalized in this way and subsequently localized into Russian and Japanese would have three DLLs: one for the source



Scribble. String Table

Number	ID	English (United States)
183	61184	Change the window size
184	61185	Change the window position
185	61186	Reduce the window to an icon
186	61187	Enlarge the window to full size

Figure 4. Typical resources include a toolbar bitmap, a document icon, a program icon, and a string table.[1]

[1] These resources are derived from a sample application called Scribble developed by the author using Microsoft Visual Studio sample files. Microsoft Download Center, "101 Visual Basic and C# Code Samples." http://www.microsoft.com/downloads/details.aspx?familyid=08e3d5f8-033d-420b-a3b1-3074505c03f3&displaylang=en

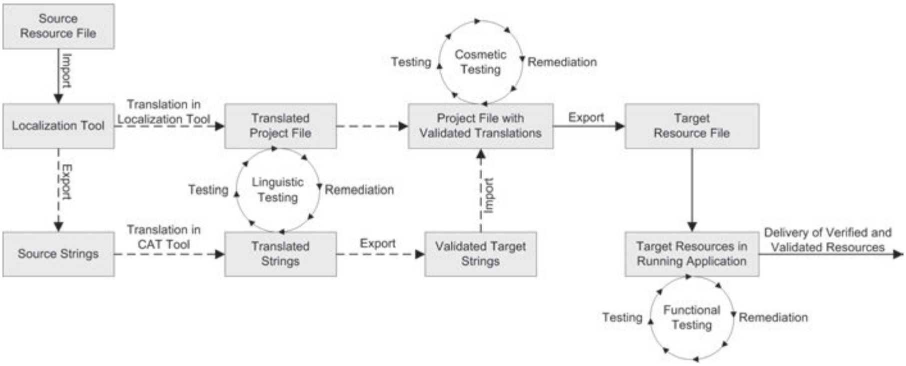


Figure 5. The GUI localization process. Translation of strings can be performed in a localization tool or in a CAT tool; the dotted lines illustrate the sequence of steps in these two approaches.

resources, one for the Russian resources, and one for the Japanese resources. If a program has been properly internationalized and no strings are hard-coded (i.e., embedded in the source code), the vendor works only with resource files and does not touch (and probably does not even have access to) the functional code of the application. The translation of strings in menus, dialog boxes and string tables represents the bulk of the work required to localize resource files. The typical steps of the localization process are as follows (see Figure 5):

1. Receipt of source-language resource files from client.
2. Translation of all translatable strings. Translation in a localization tool is preferable, as such tools display menus and dialog box strings in context. However, the translation of user interface strings can also be (and often is) performed in a Computer-Assisted Translation (CAT) tool. For the sake of simplicity we will assume that translation is performed in a localization tool.
3. Linguistic testing of translated strings (i.e., verification of the completeness, accuracy, and consistency of the translations). If the translation has been performed in a CAT tool, the validated target-language strings are exported from the CAT tool, and then imported into the localization tool.
4. Cosmetic testing of the translated project file in the localization tool. This testing, which is static in nature, seeks to detect and correct errors in the visual aspects of the user interface before generating the target-language resource files.
5. Generation of target versions of the resource files from the localization tool once the localized materials been verified and validated.
6. Functional (dynamic) testing of the localized resources in the running application to detect and correct linguistic, cosmetic, and/or functional defects.

7. Delivery. After all detected problems have been corrected, the localized resources are deemed fit for the intended use and are delivered to the client.

Having looked at the context within which localization testing is performed, let us turn our attention to linguistic testing, cosmetic testing, and functional testing to examine in greater detail how each of these types of testing assesses localized product quality.

Linguistic testing

Linguistic testing of the user interface involves the comparison of source- and target-language strings in order to detect and correct the following types of defects (discussed in detail below):[2]

- Inconsistent or missing ellipsis in target menus and/or dialog boxes
- Inconsistent number of accelerators in source vs. target strings
- Accelerator assignments that do not reflect the conventions of the target platform
- Inconsistent number of hotkeys in source vs. target strings
- Duplicate hotkeys in target menus and/or dialog boxes
- Invalid ampersand position in target hotkey assignments
- Inconsistent number of control characters (\n, \t, etc.) in source vs. target strings
- Inconsistent leading and/or trailing spaces in source vs. target strings
- Inconsistent number and/or type of dynamic variables in source vs. target strings
- Spelling errors, typos, grammatical errors, and/or punctuation errors
- Incomplete and/or inconsistent translation

Since these concepts may not be familiar to the reader, we will examine each in turn.

[2] Reflecting the fact that the distinction between linguistic, cosmetic, and functional characteristics is often impossible to maintain in practice, the first six items in this list arguably fall within the scope of functional testing. However, since these items are encoded in strings (linguistic signs), they also fall under the purview of linguistic testing, and it is for this reason that we are considering them in this section.

Inconsistent or missing ellipsis in target menus and/or dialog boxes

In software, an ellipsis is the set of three dots that follows a command in a menu or dialog box (see Figure 6). The presence of an ellipsis indicates that the application requires additional input from the user in order to execute the associated command. For example, when a user selects the "Print..." command in the "File" menu of a given application, the document is not sent directly to the printer. Instead, the application displays the "Print" dialog box. Ellipsis in source strings should be retained in the corresponding target strings.

Inconsistent number of accelerators in source vs. target strings

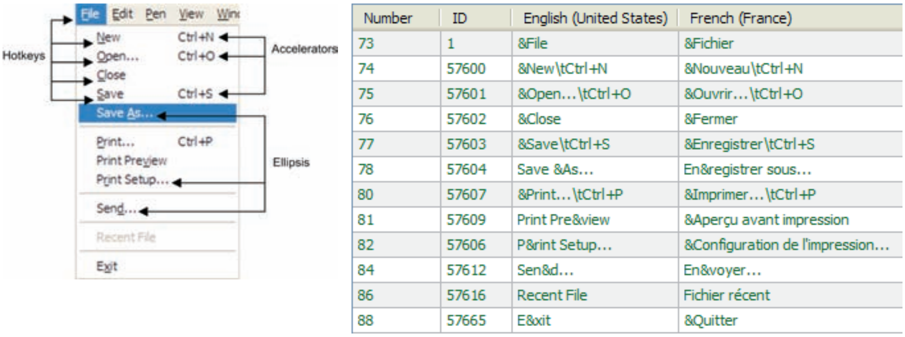
Software typically contains two types of shortcuts: accelerators and hotkeys. Accelerators enable the user to execute standard commands in Windows by simultaneously pressing the Ctrl key plus a specific letter on the keyboard (see Figure 6). The source and target versions of the application should offer the same number of accelerators.

Accelerator assignments that do not reflect the conventions of the target platform

To avoid confusing end-users, accelerator assignments in the target-language version of an application should be consistent with those used on the target-language version of Windows on which the application will run. For example, the "Bold" command shortcut is Ctrl+B in the English version of Windows, but Ctrl+G in the French version of Windows (*gras* being the French equivalent of "bold").

Inconsistent number of hotkeys in source vs. target strings

Hotkeys are application-specific shortcuts that enable the user to access commands in menus or in dialog boxes by simultaneously pressing the left-hand Alt key on the keyboard plus a specific letter. Hotkeys are typically assigned to the first letter of the corresponding command. Hotkey designations are visible in the running application as underlined letters (see Figure 6, left-hand image). Hotkeys are assigned during localization by placing an ampersand [&] in front of the letter to be used as the shortcut. It is not always desirable, nor possible, to maintain the same hotkey assignments in the source and target versions of an application. For instance, in the "Exit" command in Figure 6, the letter "x" is not found in the



Number	ID	English (United States)	French (France)
73	1	&File	&Fichier
74	57600	&New\tCtrl+N	&Nouveau\tCtrl+N
75	57601	&Open...\tCtrl+O	&Ouvrir...\tCtrl+O
76	57602	&Close	&Fermer
77	57603	&Save\tCtrl+S	&Enregistrer\tCtrl+S
78	57604	Save &As...	En&registrer sous...
80	57607	&Print...\tCtrl+P	&Imprimer...\tCtrl+P
81	57609	Print Pre&view	&Aperçu avant impression
82	57606	P&rint Setup...	&Configuration de l'impression...
84	57612	Sen&d...	En&voyer...
86	57616	Recent File	Fichier récent
88	57665	E&xit	&Quitter

Figure 6. Scribble's "File" menu in the running application with ellipsis, hotkeys, and accelerators (left-hand image) and the corresponding strings (right-hand image).

corresponding French string (*Quitter*) and thus cannot be designated as a hotkey. Instead, the target hotkey is assigned to the letter "Q" (see Figure 6). In Asian languages, the Roman-alphabet hotkeys of the source strings are maintained in the localized versions, but are placed in parentheses following the target strings. All strings that contain hotkeys in the source version should also contain hotkeys in the target version.

Duplicate hotkeys in target menus and/or dialog boxes

A given hotkey can be used only once in a given vertical menu or in a given dialog box (such as the "File" menu shown in Figure 6). Likewise, a given hotkey can be used only once across the top-level menu items that display in the main program window. Assigning the same hotkey to two different strings in a given dialog box, top-level menu, vertical menu, or sub-menu will corrupt the functionality of the hotkey.

Invalid ampersand position in target hotkey assignments

Accented or special characters should not be designated as hotkeys when other alternatives are available (Esselink 2000: 110). Likewise, hotkeys should not be assigned to the lower-case letters g, j, p, q, or y. These letters all contain "descenders," i.e., parts that fall below the line. Descenders occupy the space in which the underline would normally display. If a hotkey is assigned to a lower-case letter that contains a descender, the underline will not be visible in the running application and the user will not know that a hotkey is assigned to that string.

Inconsistent number of control characters (\n, \t, etc.) in source vs. target strings

Control characters are non-printing characters that were originally designed to control teletype machines, and which are now used to control the formatting and display of text on the screen and during printing. Examples of control characters include \t (tabulator), \n (line feed) and \r (carriage return). Applications typically contain control characters in menus (see Figure 6) and string tables. Where a control character is present in a source-language string, a control character must also be present in the corresponding target-language string.

Inconsistent leading and/or trailing spaces in source vs. target strings

Leading and trailing spaces refer to white space (one or more empty spaces) at the beginning or end of a given string. It is assumed that the source-language materials have been subjected to verification and validation testing prior to being provided to the vendor for localization. Thus, any white space in source strings should be replicated in the corresponding target strings.

Inconsistent number and/or type of dynamic variables in source vs. target strings

So-called "printf" variables are one example of dynamic variables commonly encountered in software localization projects. Printf format specifiers are placeholders that store variable data and specify the format in which the data should be output (to a monitor or printer, for example) using the printf (print formatted) function in C++, Java, Perl, and other programming languages. Printf format specifiers commonly encountered during localization include the following: %c (character); %d (signed decimal integer that can be negative); %f (floating-point number); %s (string); and %u (unsigned decimal integer that cannot be negative). Suppose we are localizing a software application whose resources include the following strings:

> This program requires the file %s, which was not found on this system.
> Free Disk Space: %lu MB Free on %c:

The first sample string above contains one printf format specifier, "%s," which is a placeholder for the name of a file that is required by the program, but which is not found on the user's system. The name of the file is stored and retrieved by the application as a string, thus the use of the string format specifier, "%s." The second sample string above contains two printf format specifiers, “%lu” and “%c.” The first, “%lu,” is a placeholder for a decimal integer, whereas the second, “%c,” is a placeholder for a character. Thus, “%lu” is a placeholder for the amount of free disk space, and “%c” is a placeholder for the drive letter of the disk in question.

Format specifiers must be retained, as they are, in the target strings. Inverting the order of the percentage symbol and the associated character(s) that comprise a given format specifier will break the functionality of the variable, as will the introduction of space between the percentage sign and the character or characters that follow. In addition, if a string contains more than one occurrence of a given format specifier, such as “%s” (string), the sequence of the variables in the target string cannot be changed.

FormatMessage format specifiers are another type of dynamic variable that function in much the same way as printf format specifiers. However, they usually follow a numeric format (%1, %2, etc.). The precise format specifiers and number thereof must be identical between source and target strings. However, the sequence of these variables can be changed within a given string, which facilitates recasting in translation.

Spelling errors, typos, grammatical errors, and/or punctuation errors

Spelling, grammatical, and punctuation errors are typically tested using spelling and grammar checking utilities integrated in CAT and/or localization tools. The main problem in testing for grammatical errors is that it is not always possible to determine the part of speech of homographs when working on decontextualized strings. “In understanding text, a reader must not only be able to integrate information within sentences but also make connections across sentences to form a coherent discourse representation,” as Rayner and Sereno observe (1994: 73). However, it is not always possible to make such connections while translating a software “text.” Due to their non-linear structure and lack of narrative thread, software programs cannot be “read” in the same way as prose. Thus, some grammatical errors may not be identified as such at this stage of testing, but are generally caught during cosmetic or functional testing (see Figure 5).

Incomplete and/or inconsistent translation

The goal of this facet of linguistic testing is threefold, namely to verify that (a) all strings that should be translated are in fact translated; (b) all non-translatable strings, such as trademarked names, have in fact remained untranslated; and (c) a given user interface term or command is translated using the same target-language equivalent in each occurrence. Much of this testing can be performed automatically using consistency checking utilities available in certain CAT tools. It is worth noting that even if each user interface term is indeed translated consistently across the application, the target version can only be as consistent as the source version is to begin with.

Automatic linguistic testing versus manual linguistic testing

The tests discussed thus far evaluate linguistic "quality" based primarily on the completeness of the translation and on the relative degree of formal equivalence of certain surface attributes of the source and target strings, such as ellipsis, hotkeys, accelerators, control characters, leading spaces, trailing spaces, and dynamic variables. Most commercial software localization tools enable users to automatically test for standard linguistic defects such as those described above. The relative quality of target resources can be quantified by measuring defects in the target text, expressed either in terms of total number or relative frequency. In practice, the goal of these automated tests is less to measure quality than to ensure that no errors are introduced during translation that would adversely impact functionality. These automatic tests are performed to eliminate as many tangible variables as possible from the assessment process prior to the evaluation of translation by a human being (ideally a localization translator with knowledge of the subject domain).

Manual linguistic testing: Translation quality in localization

Having examined those aspects of linguistic testing that are typically performed automatically, let us now turn our attention to manual linguistic testing and examine the tools, methods, and criteria whereby translation quality is evaluated during software localization projects. Manual linguistic testing of the user interface entails the comparison of target-language strings to the original source-language versions of the strings by a human being. The goal of manual linguistic testing is to detect and correct linguistic defects introduced during localization, with a focus on mistranslations and errors of meaning. Two commercial metrics are currently used to assess translation quality during localization projects: the SAE J2450 Translation Quality Metric and the LISA (Localization Industry Standards Association) QA Model.

SAE J2450 Translation Quality Metric

The SAE J2450 Translation Quality Metric is a "surface vehicle recommended practice" issued by the Society of Automotive Engineers. SAE J2450 provides a framework for measuring translation quality based on the number and severity of errors in the target text relative to the original. The J2450 metric includes seven different categories of translation error types, which are defined primarily in terms of terminological accuracy and grammatical correctness. Errors of style are explicitly excluded from the metric, as are formatting errors. Error types are further categorized based on severity level, e.g., "serious" or "minor." A weight is assigned to each error type and severity level in the form of a numerical value. (For specific details on error categories and weights in the SAE J2450 Translation Quality Metric, see Kingscott 2007: 5.) A quality translation is one that presents fewer errors, and thus a lower weighted numeric score.

Translation quality assessment performed using the J2450 metric reflects a measurement of the *relative equivalence of surface linguistic attributes* of the source and target texts. As the J2450 standard notes, the definitions of error categories "depend upon the surface form of the translation deliverable and are generally divorced from the meaning" (SAE International 2001: 3). Although "meaning is accommodated in the notion of a 'serious' versus a 'minor' occurrence of an error type" (SAE International 2001: 3), it is implicitly assumed that the source-text meaning is unambiguous and correct, and that the *meaning itself* is functionally equivalent in both the source and target contexts of use.

The presumption of functional equivalence in the SAE J2450 metric, along with the exclusion of the end-user and context of use as operational variables in the assessment of translation quality, are justified by the nature of the problem domain (the communication of automotive service information across languages), the context of use (performing automotive service operations), as well as the profile of the target customer (the service technician), all of which are explicitly specified in the metric (SAE International 2001: 3). However, the presumption of functional equivalence and the exclusion of the end-user and the context of use as operational variables in quality assessment are not necessarily valid for other domains, text types, customers, or contexts of use. Consequently, using J2450 to measure the quality of translated materials other than automotive service information, such as software user interface strings, requires an expansion of the metric to account for variables that affect perceived quality of the product in question, such as style, tone, register, or formatting, to cite but a few possible variables. Using the SAE J2450 metric "as is" in a localization project would be anachronistic.[3]

LISA QA Model

The LISA QA Model is a database-driven application that provides a framework for measuring translation and localization quality. The LISA QA Model includes the seven error categories of the SAE J2450 Translation Quality Metric, to which it adds a translation quality characteristic that is particularly important in localization, namely consistency. A localized program should employ the same translation of a given term, command, etc., consistently across the user interface, and between the interface, the Help, and the documentation. Consistency minimizes the risk of confusing users. Moreover, in some cases, proper functionality may depend on consistent translation. (SAE J2450 addresses consistency errors as part of terminological errors.)

The LISA QA Model also expands the scope of quality assessment beyond the translation of service information by including a list of seven pre-defined testing tasks typically performed during localization projects. Errors are pre-assigned to each task from among a list of 26 pre-defined error categories. (For specific details on error categories and tasks defined by the LISA QA Model, see Melby 2005: 14–22.) The Model also provides three pre-defined severity levels with assigned weights (Critical: weight = 10; Major: weight = 5; Minor: weight = 1). Project tasks, error categories, severity levels, and severity level weights are assigned on a per-project basis.

Like SAE J2450, the LISA QA Model measures errors, and thus translation and localization quality, in terms of the equivalence of the target version of the product relative to the source. Indeed, the LISA QA Model v3 product documentation explicitly advises users that "*only errors which are introduced in the localization process should be listed as such in the LISA QA Model database*" (LISA 2004: 7, emphasis in the original). In both SAE J2450 and the LISA QA Model, the quality

[3] Presumably companies that use J2450 to evaluate translations of materials other than automotive information do expand the metric, but the author has been unable to confirm this. Nevertheless, the J2450 metric enjoys widespread use among providers and purchasers of localization services. In June 2007, John Guest, Business Manager at Microsoft, asked for suggestions about software localization metrics on the business networking site, LinkedIn. A number of industry experts, including Daniel Gray, Jeff Allen, Uwe Muegge and Peter Reynolds replied. The J2450 metric was the subject of much of the discussion, with Muegge noting that "SAE J2450 . . . even though developed for the automotive industry, is widely used in the translation and localization field" (Guest 2007).

of the source version of the product, and by extension the correctness of the concept model on which it is based, are implicitly assumed.

However, SAE J2450 distinguishes itself from the LISA QA Model in one fundamentally important way: SAE J2450 specifies the problem domain, namely the communication of automotive service information across languages. In so doing, the J2450 standard eliminates the problem domain as a variable in the assessment of quality, and by extension implicitly constrains other important variables such as context of use, target audience, and user expectations. In contrast, the LISA QA Model does not specify the problem domain beyond the parameters of *country* and *language* of the product. This likely reflects a desire to create a model that can be generalized for various project and product types. However, because the LISA QA Model does not define context of use, target customer or user expectations as operational variables in the assessment of localized product quality, these factors are implicitly treated as de facto *constants* that have no bearing on quality or quality assessment. This implicit assumption that formal equivalence of the product translates into functional equivalence of meaning for all users in all contexts of use is not without problems, as we shall see.

In any event, the SAE J2450 Translation Quality Metric and the LISA QA Model both claim to provide an objective means of evaluating translation quality, both posit the quality of the source materials as a given, and both measure translation quality based on the relative accuracy, equivalence, and consistency of the target text relative to the source. However, a close examination of the application of these three criteria in linguistic testing suggests that they do not in fact allow for objective measurement.

Accuracy, equivalence, and consistency in manual linguistic testing: What constitutes an error?

In order to assess the accuracy of translation in a localization project, the reviewer must first be able to ascertain with certainty the meaning of the source strings. After all, how can one evaluate a translation if the meaning of the source text is unclear or ambiguous? It bears repeating that because the software has been verified and validated during development, the quality of the source is assumed, and is outside the scope of linguistic testing. Thus, the referential quality-as-equivalence approach to assessment assumes that the text to be translated is denotative and unambiguous – in sum, that *comprehensibility* is not a variable. This is a problematic assumption.

Homographs

For example, some software applications use homographs – words with the same lexical form but different meanings – as discrete single-word strings, e.g., *Archive* meaning "to file" (v.) and also "the storage system" (n.), "the collection of stored items" (n.), and perhaps "the location of the stored items" (n.). Typically, a homograph in English corresponds to two or more target-language equivalents. It is generally not possible to assess the accuracy of the translation of a homograph without knowing the intended part of speech, the precise context in which the string appears, and the specific function with which it is associated. When working in a WYSIWYG ("what you see is what you get") environment such as a commercial localization tool, the localizer and the tester can disambiguate such strings, provided that the strings display in menus or dialog boxes. However, if the work is being performed in a non-WYSIWYG environment, or if the strings are stored in a format that does not enable in-context representation, such as string tables or XML files that have not been authored following best internationalization practices, the localizer's (or tester's) ability to determine the appropriate target-language equivalent is seriously undermined (Dunne 2006b: 102–105).

Telegraphic style

The use of telegraphic style in software presents similar problems. From a developer's point of view, omitting relative pronouns and other words that are "understood" may seem desirable to achieve economy of style and/or to limit translation costs, which are typically calculated on a per-word basis. However, what is implicitly understood by an expert user or developer may not be clear (or may not occur) to a translator, localizer, or tester working at the level of individual strings in the nether world of a CAT or localization tool. When telegraphic style is used in error message and status message strings, the meaning is likely to be lost on those who are not intimately familiar with the workings of the system. Consider the following message string, which consists of three words followed by a colon and a dynamic variable:

> Problem checking variables: {0}

In order to ascertain whether the translation of this string is accurate, the reviewer must first determine what exactly the string means. Is it a status message or an error message? Does the string mean that (a) the application is currently conducting a problem-check on the following variables: {0}; (b) a problem was found while checking the following variables: {0}; (c) etc.? As Campbell (2005) points out, telegraphic style can effectively block comprehension: "outsiders may be unable to get the meaning even if they are able to undo the clause reductions – primarily because there are many possible underlying clauses" (2005: 17). In the case of the sample string above, definitively determining the intended meaning is difficult, if not impossible, without clarification from the development team.

Noun stacking and hidden plurals

Noun stacking (i.e., modifier + noun + noun + noun, etc.) presents similar challenges. As Campbell notes, stacked nouns are "inherently ambiguous" because "they can be understood several ways depending on how the syntax is reconstructed" (2005:16). Hidden plurals are another example of an inherently ambiguous syntactic structure. For example, does the phrase "user data" mean (a) data about *a* user; (b) data about *the* user; (c) data about users in general; (d) user-generated data; or perhaps (e) something else? A definitive determination of the intended meaning of this phrase as written would be difficult, if not impossible, without further clarification. As the above examples demonstrate, extreme economy of expression may result in phrases and sentences whose meaning cannot be determined unequivocally, and thus which cannot be translated as written without clarification from the client or developer. In the absence of clarification, there is no objective, valid foundation on which to assess the accuracy of translation of ambiguous strings such as these.

Lexical ambiguity

The ability to see strings in context can facilitate disambiguation. However, working in a WYSIWYG environment does not automatically eliminate risks of misapprehension. For example, let us consider the case of the Webroot Spy Sweeper v5.5 installation program. Upon launch, the installer displays a "Welcome" screen, which contains three command buttons whose captions are "Install Help," "Next >" and "Cancel" respectively. As their name suggests, command buttons cause a command to be executed when they are pressed. Thus, when the "Next >" command button is pressed, the installer displays the next screen in the installation sequence, and when the "Cancel" command button is pressed, the installer aborts the setup process. Following this logic, it would appear to a localizer or to a linguistic tester working in a static environment such as a CAT or localization tool that the command button caption "Install Help" is telegraphic form of the phrase, "Install the Help." In fact, when the "Install Help" command button is pressed, the system does not install the Help but rather displays a Help file that contains an installation Quick Start guide.

Thus, the string "Install Help" does not mean "install the Help" but rather "display the 'install' Help," in other words, the Help that explains the installation process. In this case, the word "install" is not a verb, but rather a *noun* used as a pre-positioned modifier. However, that fact will not necessarily be apparent to translators, localizers, or testers unless they realize that "install" is commonly used as a noun in the realm of IT, and/or they actually click the "Install Help" command button. Given that this stage of the localization process and this level of linguistic testing are conducted in a static environment such as a CAT or localization tool, and not in the dynamic environment of the running application, it is likely that "install" would be misapprehended as a verb instead of a noun. However, if the error was not detected during linguistic testing, it would probably be identified by an alert reviewer during subsequent functional testing (see Figure 5).

The Spy Sweeper "Install Help" example sheds light on another problematic facet of "accuracy" in localization. Clicking the "Install Help" command button causes a file to open whose name is not "Install Help" but rather "Spy Sweeper 5.5 Quick Start." That being the case, a translator, localizer, tester or end-user could legitimately ask whether the source string itself is accurate to begin with. Indeed, from the point of view of an end-user, the button caption should arguably be "Display the Quick Start Guide" or some other such string that more clearly identifies the command that is executed when the button is clicked. In this case, the production of a "functional" translation – instead of one that is merely equivalent or accurate – would require the translation of what is *meant*, not what is *written*. However, as noted above, linguistic testing evaluates quality based on equivalence and accuracy of the target relative to the source, whose quality is *assumed*. In other words, the goal of linguistic testing is to ensure that the target strings accurately reflect the source strings as written – not in terms of what they *should* say, but in terms of what they *actually* say. Thus, the fact that the "Install Help" command button caption could likely confuse users of the source-language version of the application, as it did the author of this chapter, is outside the ostensible scope of linguistic quality assessment. In this case, the successful application of the quality-as-equivalence approach would result in the production of a target version that is as effective in confusing users as the source version. In localization, this problem is known as "Garbage In, Garbage Out."

Dynamic numerical variables

The foregoing paragraphs provide a glimpse into the difficulty of merely ascertaining the intended meaning of individual software strings. Problems of lexical and semantic ambiguity are amplified when strings contain dynamic variables,

Table 1. Dynamic string variables can cause problems when working into inflected languages such as Russian.

This text contains	{0}	vowels	and	{0}	consonants.	Noun ending:
В этом тексте	0	гласных	и	0	согласных.	ых
	1	гласная	и	1	согласная.	ая
	2, 3, or 4	гласные	и	2, 3, or 4	согласные.	ые
	5–20	гласных	и	5–20	согласных.	ых

The pattern repeats across multiples of 10 (21, 22–24, and 25–30; 31, 32–34 and 35–40; etc).

i.e., placeholders for other strings that are inserted to create composite phrases, sentences, or meanings. For example, let us assume that we are localizing a small program that counts the number of consonants and vowels in a given text, and then displays the results in the form of the following string:

> This text contains {0} vowels and {1} consonants.

In this example, the variable {0} represents the number of vowels, whereas the variable {1} represents the number of consonants. The string as written assumes that the numerical value neither the variable {0} nor the variable {1} will be equal to "1," since a substitution of either {0} = 1 or {1} = 1 will result in a grammatically incorrect sentence in English (e.g., "This text contains 1 vowels and 1 consonants."). The sample string above also assumes that the plural form is invariable, i.e., the same lexical form is used for the plural of "vowels" and "consonants" regardless of the numerical quantity, provided that quantity is greater than 1. How then can – or should – one evaluate the "quality" of a translation of this string into an inflected language? In Russian, for example, the declensions of the nouns *vowel* and *consonant* must change depending on the numerical quantity of the noun: one form is used for a numerical quantity of 1; another for a numerical quantity of 2, 3 and 4; and a third for a numerical quantity 5 to 20 (and also for zero). Consequently, a total of nine different target-language combinations would be required to render all permutations of the output of the above example in grammatically correct Russian. Further complicating things is the fact that these permutations follow a repeating pattern across numerical values (see Table 1).

In this example, an accurate translation is impossible, barring modifications to the program code, since the syntactic assumptions about plural formation inherent in the source-language string are not valid in Russian. The localized Russian version of the program does in fact function as designed; in other words, it can be said to be *functional* insofar as the number of vowels and consonants in a given Russian text can indeed be successfully counted and the final tally of each displayed as output in place of the dynamic variables {0} and {1}, respectively.

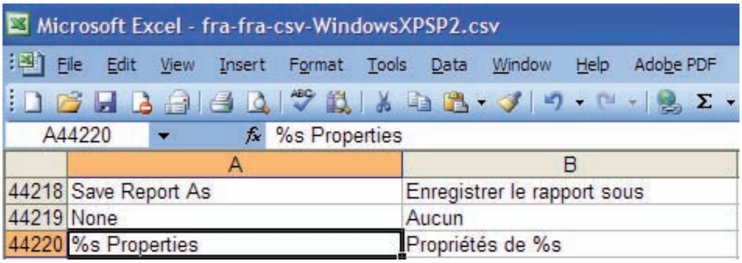
However, from the user's perspective, the program is a black box whose "functionality" is evaluated based on the correctness of the output. In this regard, the localized Russian version can be said to be *dysfunctional* insofar as the linguistic format in which the output is expressed violates the grammatical norms of the target locale.

Dynamic string variables

The previous example illustrates the ways in which the substitution of numerical variables can undermine the possibility of achieving an accurate translation in software localization. Let us now consider the challenges posed by the substitution of string variables. In Microsoft Windows, when the user right-clicks on an object and selects the "Properties" item in the context menu, the operating system displays a dialog box whose title is formed by combining the name of the object plus the string "Properties." For example, if a user right-clicks on the desktop (i.e., the Display) and selects "Properties" in the context menu, the composite string that comprises the title of the dialog box is "Display Properties."

In Windows' resources, the generic "Properties" dialog box title is represented as the string "%s Properties," where "%s" is a string variable, or placeholder, for the name of the selected object (see Figure 7). As was the case with the numerical variable substitution example discussed above, the rules governing this string variable substitution are based on and valid for rules of English grammar, and may not be valid for the target language. For example, translation of this string into French requires recasting using a prepositional phrase. In other words, "%s Properties" is translated as "Properties of %s," i.e., *Propriétés de %s* (see Figure 7). However, in keeping with conventions of usage, translation into French should specify not only the *name* of the object, but also the *type* of object. Thus, if the user right-clicks on the shortcut to a program, such as Transit NXT, and chooses the "Properties" item in the context menu, the operating system should – in theory – specify that the properties are those of the *program* Transit NXT (*Propriétés du logiciel Transit NXT* as opposed to *Propriétés de Transit NXT*). Likewise, if the user right-clicks on a document, such as Détritus.doc, and chooses the "Properties" item in the context menu, the operating system should specify that the properties are those of the *document* Détritus.doc (*Propriétés du document Détritus.doc* as opposed to *Propriétés de Détritus.doc*). However, expressing properties in terms of both the type and the name of the selected object would require modifications to the Windows source code, which is by definition outside of the scope of localization.

The translation of the string "%s Properties" into French also causes grammatical problems. The simple recasting and translation of the source string



%s Properties

	A	B
44218	Save Report As	Enregistrer le rapport sous
44219	None	Aucun
44220	%s Properties	Propriétés de %s

Figure 7. The generic "Properties" dialog box title string in the en-US/fr-FR bilingual Microsoft Windows XP Service Pack 2 Glossary of Translated User Interface Terms.

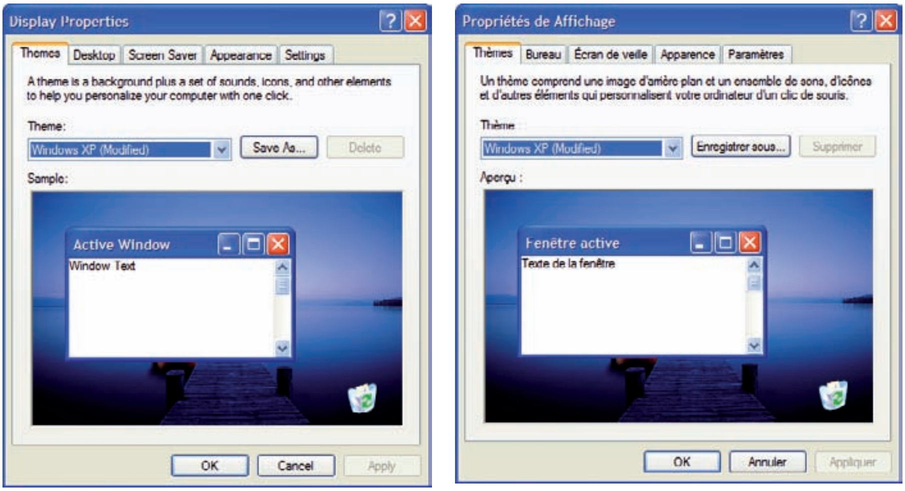
"Properties of %s" as *Propriétés de %s* fails to account for the ellipsis of the definite article in the English string. However, the use of the definite article is mandatory in French when referring to a specific object, such as the display. Thus, French translations of the string "%s Properties" should include both the preposition *de* and the definite article, which could be masculine singular (*le*), feminine singular (*la*) or plural (*les*) depending on the selected object. Further complicating things is the fact that the singular form of the definite article changes to *l'* if noun begins with a vowel. In addition, French grammar requires contraction when the preposition *de* is followed by a masculine or plural definite article: *de* + *le* = *du* and *de* + *les* = *des*.

Consequently, when a user right-clicks on the Desktop (i.e., the Display) in the French Multilingual User Interface of Windows and selects *Propriétés* in the context menu, the composite string that displays as the dialog box title is *Propriétés de Affichage*, which is grammatically incorrect in French (see Table 2; see also Figure 8, right-hand image).

The correct form of the target-language composite string is in fact *Propriétés de l'affichage*, as shown in Table 2, but the implicit English-language syntactic assumptions governing this specific string substitution do not enable the formation of grammatically correct French-language equivalents when the name of the selected object begins with a vowel. From the perspective of the linguistic tester, the translations of the individual strings "Display" and "Properties" meet the quality inspection criteria of accuracy and equivalence. However, accuracy, equivalence, and consistency are not sufficient in and of themselves to ensure the quality of strings formed by dynamic variable substitution if the grammatical and syntactic rules governing the string combination do not enable the generation of linguistically correct output.

Table 2. Problems caused by dynamic string variable substitution when working from English into French.

Language	String	Value of variable %s	Composite formed by insertion of variable	Correct form of composite
English	%s Properties	Display	Display Properties	Display Properties
French	Propriétés de %s	Affichage	Propriétés de Affichage	Propriétés de l’affichage



A theme is a background plus a set of sounds, icons, and other elements to help you personalize your computer with one click.

Theme: Windows XP (Modified)

Sample:

Un thème comprend une image d'arrière plan et un ensemble de sons, d'icônes et d'autres éléments qui personnalisent votre ordinateur d'un clic de souris.

Thème : Windows XP (Modified)

Aperçu :

Figure 8. “Display Properties” dialog in English (left-hand image) and French (right-hand image) Multilingual User Interface of Windows XP Professional. Microsoft product screen shots reprinted with permission from Microsoft Corporation.

Tabbed dialog boxes

Tabbed dialog boxes also call into question the extent to which linguistic quality can be assessed in terms of equivalence between source and target versions of strings. A tabbed dialog box is a composite user interface object in which multiple forms, or “boxes,” are superimposed; string labels on respective tabs permit identification of, and navigation among, the various options. Dialog box tabs are based on the metaphor of paper file folders, which are used to store and organize papers in a file cabinet. Following the file folder metaphor, tab labels should theoretically be labeled using nouns. However, software dialog box tabs are often labeled using verbs and hanging adjectives. Such tabs arguably violate the conceptual model on which they are based, and render “accurate” translation problematic. For example, one tab in the “Preferences” dialog box shown in Figure 9 is labeled with the hanging adjective “Global.” In this case, the adjective “Global” modifies the

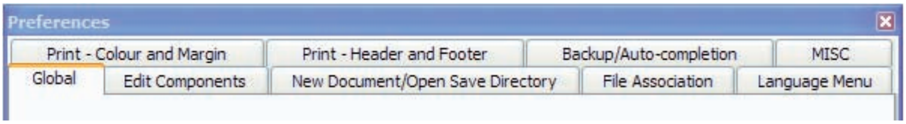
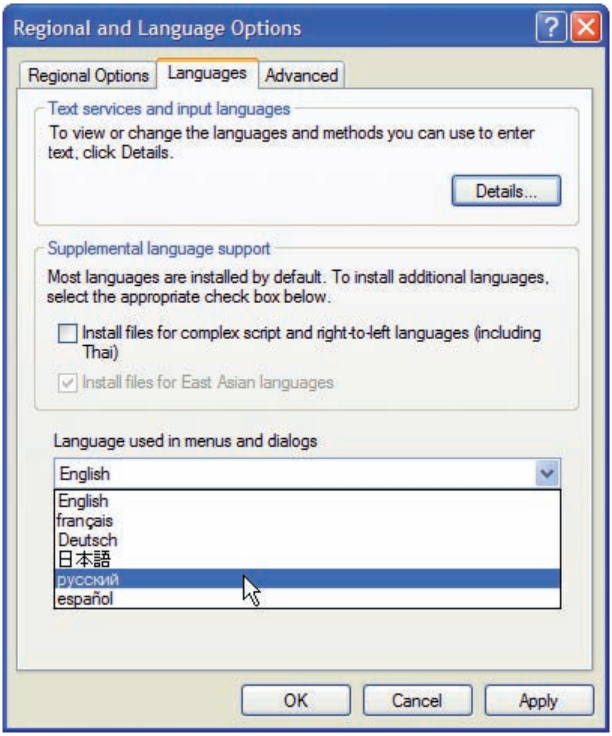


Figure 9. A tabbed dialog box in Notepad++ v5.3.1. © 2009 Don HO, SourceForge.net.

noun “Preferences,” which displays in the dialog box title bar. The user of the running application, who sees the tab label caption in context, understands that the hanging adjective “Global” is in fact a telegraphic form of the plural noun phrase “Global Preferences.” The meaning of the “Global” label can only be correctly understood in context, and in relation to the title of the dialog box. However, the localizer and linguistic tester may or may not be able to see the strings in context, depending on the ways in which the resource files have been authored and/or the tools being used to perform the work. If the localizer and linguistic tester cannot see or do not realize that “Global” modifies “Preferences,” they will default to transposing the lexical form of the source strings onto the corresponding target versions. If the morphological rules governing noun-adjective syntax in the target language do not mirror those of the source language, transcoding the part of speech of the source labels in translation will not suffice to produce functional target-language equivalents. This problem can be rectified during localization by explicating the referent and translating the tab label caption “Global” as “Global Preferences.” However, such explication may or may not be possible depending on the amount of space available on the tab and/or the extent to which the tab can be resized to account for expansion.

Alphabetized lists

Alphabetized lists that are not re-sorted after translation provide another example in which a translation can be both accurate and equivalent to the corresponding source text but still be defective. If the sorting order in the target version is determined by the alphabetical order of the source strings, then accuracy of translation and equivalence of sequence will not suffice to produce a translation that conforms to the rules of usage of the target language. Figure 10 presents an example of a dialog box in which the names of six languages have been translated and written in the corresponding scripts, and capitalized (or not) based on the conventions of the individual target languages. Hence, “English” and “Deutsch” are capitalized, whereas “français” and “español” are not. However, the list is not sorted by the alphabetical order of the translated language names, but by the alphabetical order of the names of the languages in English.



Regional and Language Options

Text services and input languages

To view or change the languages and methods you can use to enter text, click Details.

Supplemental language support

Most languages are installed by default. To install additional languages, select the appropriate check box below.

□ Install files for complex script and right-to-left languages (including Thai)

☑ Install files for East Asian languages

Language used in menus and dialogs: English

Figure 10. Multilingual User Interface (MUI) language selection in Windows XP Professional. Microsoft product screen shot reprinted with permission from Microsoft Corporation.

This cursory examination of linguistic testing, while not exhaustive, nonetheless suggests that objective measurement of linguistic quality in localization is more of an ideal toward which to strive than a goal that can be effectively achieved in practice. Let us now turn our attention to the other two facets of localization testing, namely cosmetic testing and functional testing.

Cosmetic testing

Cosmetic testing focuses on the visual aspects of the user interface and is designed to ensure that everything that should display in the localized version does in fact display in its entirety in the localized version. A typical software localization project comprises two rounds of cosmetic testing. The first round is performed in the localization tool following the translation of the user interface strings (see Figure 5). Because the localized resources have not yet been integrated into the application at this stage of the localization process, this round of cosmetic testing is generally static in nature. A second round of testing is thus performed later in the process in the localized version of the running application. Cosmetic testing is designed to confirm the following (Esselink 2000: 151–152):

- The localized version displays the same number of user controls as the original, i.e., menus, menu options, dialog boxes, combo boxes, etc.
- All strings display in their entirety in localized dialog boxes, without truncation.
- The tab order in localized dialog boxes matches that of the original version.
- Accented characters, extended characters, and special characters display correctly.
- The main application window, main menu, menu items, status bar messages, tool tips, and dialog boxes display correctly and in their entirety at all screen resolutions.

In current professional practice, the primary focus of cosmetic testing is arguably to ensure that translation-related expansion does not lead to truncation of any strings. Translation from English into Western European languages can cause string length to expand by more than 30% (Esselink 2000: 67). Individual words or strings can expand by 300% or more. For instance, the source-language string “Pop-up blocker” in Internet Explorer’s “Tools” menu is only 14 characters long (including spaces), whereas the corresponding French equivalent, *Bloqueur de fenêtres publicitaires intempestives*, is 48 characters long (including spaces). In this case, translation-related expansion exceeds 340%. In addition, expansion tends to be inversely proportional to string length: the shorter the string, the fewer the number of syntactic and stylistic options available to the translator to limit expansion, and the more acute the problems of expansion tend to be.

Design elements such as dialog boxes may require twice as much space in localized versions. If the software being localized has not been authored with localization in mind, and if strategies have not been implemented in the design to accommodate expansion, truncation may be widespread, requiring extensive effort to resize user interface controls and layouts to enable the translated strings to display in their entirety. If no translation-related expansion is tolerated, it may be impossible to formulate an equivalent that is idiomatic, accurate – or in a worst-case scenario, even comprehensible – in the target language.

Mainstream localization tools offer automatic truncation testing features, which can be performed on the fly during the translation process (if the strings are translated in the localization tool and not in a CAT tool; see Figure 5), or after the translated strings are imported into the localization tool. On the other hand, symmetrical layout and display are generally verified by performing manual cosmetic testing, as are bitmaps and icons.

Like linguistic testing, cosmetic testing frames notions of quality largely in terms of formal equivalence between the source and target versions. However, cosmetic testing does expand the scope of localization assessment beyond the dichotomy of source-target equivalence insofar as it focuses on the format and presentation of locale-dependent data (referred to as "regional settings") and the layout of the user interface. Thus, cosmetic testing verifies that date formats (MM/DD/YYYY, DD/MM/YYYY, etc.), time formats (12-hour vs. 24-hour clock; etc.), decimal separators (comma vs. period), and other culture-dependent data presentation formats are consistent with the conventions of the target locale. In this regard, it is important to note that locales and languages do not necessarily coincide. For example, although Mexico and Spain are both Hispanophone countries, the decimal separator in Mexico is a period, whereas in Spain it is a comma. It is for this reason that requests for localization of software into "international Spanish" are not just anachronistic, but actually impossible to perform. Any assessment of the compliance of localized software with target locale conventions presupposes that the intended *locale* has been specified, and not merely the target language.

Cosmetic testing also aims to ensure that the layout of the user interface is acceptable to target users. It must be noted, however, that the adaptability of the target layout is largely constrained by the physical design of the source version. Consequently, cosmetic testing of the UI layout is generally limited to the directionality of reading, the alignment of buttons, and the use of color.

Functional testing

Having looked at linguistic and cosmetic testing, let us now turn our attention to functional testing, the final component of localization testing, which is performed in the running application using the localized resource files. The functional testing procedure performed on a localized application ideally mirrors that which was performed on the source-language version. The primary goal of such testing is to ensure that no functionality has been adversely affected by the localization process. However, functional testing also often serves as another round of linguistic and cosmetic testing. As noted above, linguistic and cosmetic testing are performed in static environments, and as such, may fail to reveal problems that involve dynamic aspects of the user interface (such as variable substitutions).

Like linguistic and cosmetic testing, functional testing involves some tasks that can be performed automatically, and others that must be performed manually. Automatic testing can be carried out using functional testing tools such as

Borland SilkTest that verify the proper execution of commands, the proper display of strings without truncation, and so forth. Such tools typically record the results of the testing in a log file. Manual testing ideally involves following the same script used in functional testing of the source-language version, whereby the tester verifies each function and inspects each component of the user interface to ascertain that everything works properly, displays correctly, and makes sense. Thus, whereas an automatic test confirms that there are no problems with the *functionality* of variable substitutions, manual testing verifies that there are no problems with the *meanings* of the composite strings thus created. Although quality can be measured by comparing the number of defects detected in the target version to the number present in the source, the objective of functional testing, as is the case with linguistic and cosmetic testing, is first and foremost to verify that no new defects have been introduced during the localization process.

Localization testing from the vendor's (or practitioner's) perspective: Implications

From the perspective of the vendor or practitioner, localization quality assessment consists of localization testing, and quality is framed in terms of accuracy, equivalence, and consistency between the target version(s) of an application relative to the source. However, as we have seen over the course of the foregoing discussion, there are a number of conceptual and methodological problems inherent in this approach, beginning with the difficulty of simply ascertaining the intended meaning of certain source-language strings. The assessment of accuracy presupposes that the meaning of the text is clear and unambiguous, an invariable benchmark against which any translational deviance can subsequently be identified and quantified. As the foregoing discussion has illustrated, the intended meaning of source-language strings in software cannot always be determined with absolute certainty. In the absence of explicit clarification from the development team or a representative thereof (and an audit trail documenting such clarification), both the localization of ambiguous source materials and the assessment of the accuracy of such localization devolves into a exercise in "creative extrapolation" (Pringle and O'Keefe 2003: 30) whose subjective nature calls into question the reproducibility of the process, the repeatability of the results, and the validity of the testing process itself. After all, if we do not – or cannot – establish a clear, consistent baseline against which to evaluate accuracy, just what "errors" are we counting during localization testing, and what is the validity of any localization quality "measurement" that is subsequently derived?[4]

The assessment of equivalence and consistency in localization poses similar problems. The notion of equivalence tends to posit lexical, syntactic, and semantic symmetry between source and target versions of an application that may not be possible. As we have seen, the grammatical and syntactic assumptions governing the authoring of strings and the use of dynamic variables may make it difficult, or even impossible, to produce translations that are grammatically correct, stylistically appropriate and/or semantically equivalent for a given target locale. Problems of equivalence are amplified by certain physical design attributes of software, such as tabbed dialog boxes and lists whose collation is programmatically driven by the alphabetical order of the source strings, as noted above. Consistency, on the other hand, is not inherently problematic *per se*. However, by taking for granted the quality of the source materials and by assessing consistency of the target version only to the extent that it mirrors that of the source application, linguistic testing fails to address (in)consistency in and across source materials, and the larger issue of the coherence of the source application as a whole, which can – and typically does – impact the perceived quality of the target version as evaluated by client reviewers or target end-users.

The discussion of localization testing thus far is grounded in the context of localization project processes. As such, it reflects the perspective of the *practitioner*, for whom the quality attributes of the target version are implicitly understood, and explicitly evaluated, in terms of their relative equivalence to the corresponding attributes of the source version (see Figure 11). Experienced practitioners understand the challenges inherent in software localization and the extent to which target product quality is constrained by the source version of the application and by the medium of software itself. A client reviewer who does not share this experience, knowledge or perspective on the localization process, and who may not understand the parameters that constrain the very possibility of "quality," will undoubtedly have a very different view.

[4] Indeed, a case study on the use of the LISA QA Model by L&L, a Netherlands-based translation and localization services provider, notes that "[t]he first problem encountered was that the customer wanted quality but then failed to clearly define what quality meant to them. In other words, definitions for critical, major, and minor errors could not be given. In this case, *L&L used its own standards and hoped that this definition suited the customer as well*" (Koo and Kinds 2000: 156; emphasis added).

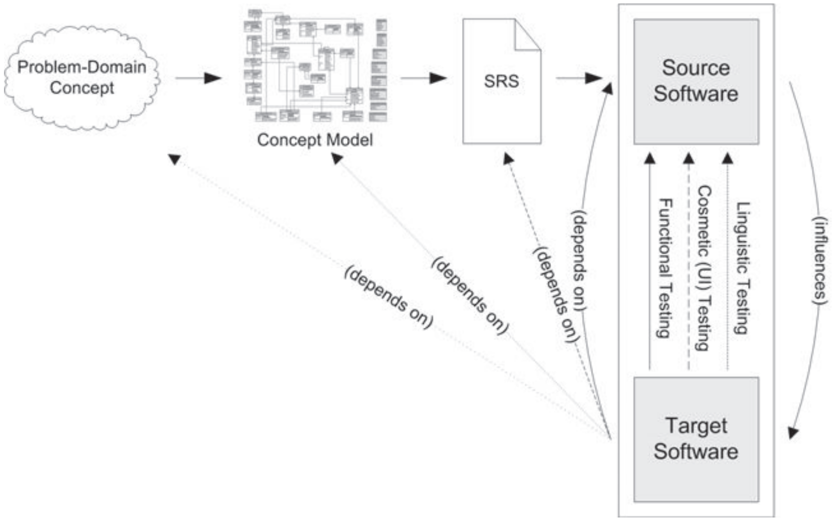


Figure 11. The scope of localized software quality assessment and the factors that shape "quality" from the perspective of the vendor or practitioner during localization testing.

Localized software quality: The client reviewer's (end-user's) perspective

As we have seen, accuracy, equivalence, and consistency – to the extent that they are possible – are not sufficient in and of themselves to ensure quality in localized software, nor do they constitute a formalized specification against which to objectively assess localized software product quality. Because software is pragmatic, functional and informative in nature, the goal of localization from the point of view of translation is – or rather, should be – the production of a "target text" that produces the same effects as the original. By extension, localized software quality is – or rather, should be – understood as the degree to which this objective is achieved.[5] It follows that the goal of localized software quality assessment is – or rather, should be – a determination of the degree of functional equivalence between the source and localized versions of the application. Obviously, neither the effects produced by software nor the degree of functional equivalence between source and localized versions of software can be properly understood or correctly evaluated without reference to users and the context of use of the product in question.

[5] This observation raises fundamental theoretical and methodological questions with regard to quality assessment: What exactly are the effects produced by the original? What effects are we measuring? What should we measure? What *can* we measure? What are the operational variables influencing quality and perceptions thereof, and how can we constrain them?

However, current approaches to localized software quality assessment tend to focus on the relative degree of *formal* equivalence of various characteristics of source and localized software *products*, rather than the degree of *functional* equivalence of the *meanings* and *effects* produced by those characteristics of the source and localized versions. Further complicating things is the fact that in the realm of localization "functional equivalence" is generally understood to mean that the source and localized versions of an application present the same set of *functionalities*. The question of whether the formal characteristics of software are perceived in the same way and hold the same meaning for source and target users usually falls outside the scope of localized software quality assessment. Instead, it is assumed that formal equivalence and replication of functionality translate into functional equivalence of meaning.

The client review process tends to both reveal and amplify the problematic assumptions inherent in the approach to localization quality assessment as understood by the vendor or practitioner. To begin with, client review typically amounts to acceptance testing conducted after localization is complete by a reviewer who is not comparing source and target versions but rather evaluating the target version *on its own merits based on his or her concept of the problem domain and solution*. In addition, the client reviewer rarely possesses the full complement of (theoretically) necessary skills. For the purposes of this discussion, we will presume that the reviewer is a client-side subject-matter expert, employee, or in-country distributor, who is a native speaker of the language into which the software is being localized, and who is familiar with the customer's corporate culture and linguistic preferences, but who has no translation and/or localization expertise. Thus the client reviewer is most likely unaware of the influences and dependencies that implicitly and explicitly shape the localized product (see Figure 11). In addition, since the client is typically neither involved nor consulted during the development of the source-language application, the organization's requirements are generally not captured in the original specification. The absence of a formal specification of client quality requirements undermines the very possibility of objective quality measurement. After all, how can one measure the degree to which a product complies with requirements if the requirements are not specified? The answer, of course, is that in the absence of *identified needs* – requirements – quality is assessed on the basis of *unidentified needs*, namely expectations and preferences (see Figure 12).

Consequently, the typical client reviewer fully expects that the software will meet his or her expectations, and will likely categorize as "errors" any linguistic attributes of the localization that do not reflect his or her terminological and/or stylistic preferences. The widespread identification of such "errors" is highly likely, the best intentions and efforts of linguistic testing notwithstanding, because the focus on accuracy and equivalence during localization testing fails to address the

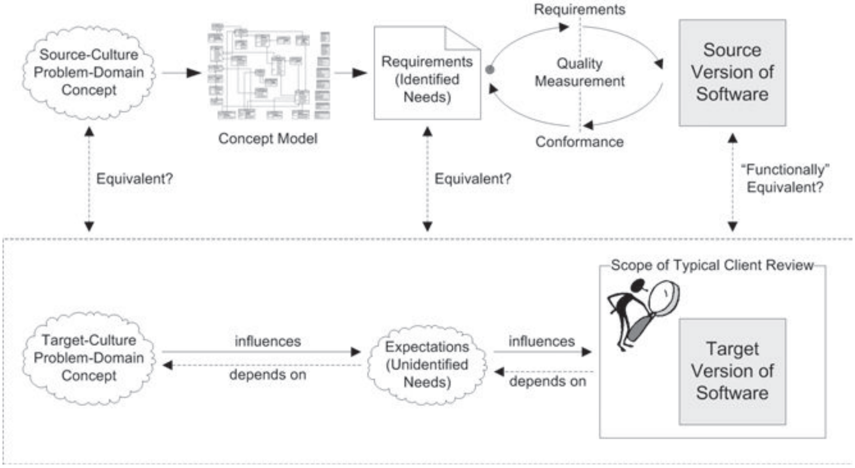

Source-Culture Problem-Domain Concept
Concept Model
Requirements (Identified Needs)
Requirements
Quality Measurement
Conformance
Source Version of Software
Equivalent?
Equivalent?
"Functionally" Equivalent?
Scope of Typical Client Review
Target-Culture Problem-Domain Concept
influences
depends on
Expectations (Unidentified Needs)
influences
depends on
Target Version of Software

Figure 12. The scope of localized software quality assessment and the factors that shape the perception of localization quality during client review.

issue of terminological and stylistic *preferences* when there are multiple ways to translate a given term or phrase. The greater the divergence between what clients (or client reviewers) expect and what they actually receive, the greater the likelihood of customer dissatisfaction. Along similar lines, by focusing on consistency at the level of individual strings, localization testing fails to address the relative consistency of the product as a whole. For example, suppose a source application uses the terms "folder" and "directory" interchangeably. If the terms are translated accurately and consistently across the entire interface, and in all corresponding documentation, the target versions of these materials may still be deemed defective by a client reviewer if the use of synonyms is unacceptable in the target context or if one of the terms is deprecated.

At a more fundamental level, the implicit assumption of functional equivalence underlying the software localization process posits equivalence between source and target users, contexts of use, mental models, and dimensions of culture that may or may not reflect reality (Hall and Hall 1990; Hofstede 1991). Thus, the client reviewer will likely be dissatisfied by any cosmetic and/or functional attributes that diverge from his or her expectations of how the user interface should look, the ways in which the software should perform, and so forth. And why should reviewers not be dissatisfied by anything that does not meet their expectations? In the current marketplace, quality is defined in terms of customer satisfaction, i.e., "customer *perception* as to whether the organization has met customer requirements" – whether those requirements have been identified as such and captured in the form of expectations, or whether they have not been identified and remain tacit expectations and preferences (ISO 2008: 12; emphasis added).

Conclusion

In today's market, quality is whatever the client says it is. In other words, quality of *product* is indissociable from quality of *service*. This reality suggests that localization quality assessment should focus less on end products than on the customer and the customer's requirements with regard to such end products. However, in the current outsourced localization project model, localization typically begins after the development of the source application is complete. When development is divorced from localization, the requirements of the localization customer are outside of the scope of the initial project. In the absence of documented customer requirements, there are no grounds on which to evaluate the quality of the localized product. Therein is the heart of the problem: during outsourced localization projects, clients expect quality products, but their localization quality requirements have been neither captured nor addressed during the development of the source-language application on which the localized version(s) will be based. How can we reconcile this contradiction? How – or to what extent – can we establish consistent definitions of software quality attributes and incorporate them into quantitative evaluation metrics that are valid for any given type of localized application and for any given customer? The answer is that quantitative localization quality assessment is impossible if a cookie-cutter approach is taken:

- Quality is not an absolute, but rather reflects the customer's or user's *perception* of the product (thus the truism "quality is in the eye of the beholder").
- The quality of language, communication, and meaning do not lend themselves well to objective quantification using scientific methods. Results obtained using scientific methods are repeatable, meaning that they can be independently verified by other researchers following the same procedure. The melting point of copper ore at a given atmospheric pressure can be objectively and independently confirmed through experimental observation. The same cannot be said for the clarity of a given software user interface string, for example.
- Quality, as its etymological root indicates, is an inherently *qualitative* – not quantitative – phenomenon.
- From a project management perspective, projects are by definition unique (PMI 2000: 5). It follows that quality assessment is valid only to the extent that it is based on, and measures compliance with, the critical quality requirements of a *specific* customer formulated and captured during the planning stage of a *specific* project.

The market expects scientific precision in localization quality assessment, but scientific approaches are generally not used to measure quality. As Heiman (2001: 49) points out, "an *operational definition* defines a construct or variable in terms of the operations used to measure it." In the current outsourced localization project model, project participants and client reviewers do not share the same operational definition of quality as a variable. Indeed, many language industry stakeholders, vendors and clients alike, do not grasp the fundamental fact that quality is in fact a *construct* shaped by any number of cultural, linguistic, organizational, and commercial factors (among many others), rather than an absolute. What is needed is a complete re-thinking of the operational definition of localization quality as a variable, framed in terms of customer- and project-specific requirements. Localization quality does exist, and it can be measured. However, localization quality is intrinsic to the *assessor*, not the product. The focus of localization assessment needs to reflect that fact.

The problems inherent in current notions of localization quality and the assessment of localized software quality reflect a fundamental terminological confusion. The noun "quality" can mean both *degree of excellence* and *an essential characteristic, property or attribute*. "Quality" as used by localization service providers and practitioners today generally connotes "degree of excellence." However, "quality" as understood by customers (purchasers of localized products and localization services) generally connotes "the degree to which I am satisfied with the product/service," which in turn can be understood as "the degree to which I perceive that the product/service meets my expectations."

Given this divergence between client and vendor perspectives on quality, the proper goal of localization quality management should be to define and control the qualities of the localized product – understood as *characteristics* – that influence the customer's *perception* of product quality and concomitant degree of customer satisfaction. Because the characteristics that influence the customer's perception of product quality are subjective and contextually determined, quality (that is, the characteristics that shape the perception of quality) cannot be *defined,* but rather should be *modeled* on a per-project basis according to the specific project requirements (ASTM International 2006: 5).

Seen from this perspective, the management of localization quality is correctly understood as the management of *expectations*. In order to meet the customer's quality requirements, the vendor must identify the critical quality characteristics or variables for each project/client before starting any localization work. Terminology is arguably the single most critical characteristic in localization (Wright 2001; Dunne 2007). Terminology management provides a means of capturing the customer's terminology preferences in the form of requirements against which compliance can subsequently be measured. Likewise, the use of a style guide provides a framework within which to define critical stylistic characteristics that shape the client's perception of linguistic quality. Documenting preferences in a formal style guide provides a stylistic requirements specification for the project deliverables. (See for example Kohl 2008 and Microsoft 2009.) By circumscribing the critical terminological and stylistic characteristics of a given project, the vendor creates a client- and project-specific localization specification against which to measure the compliance of the localized end product. Localization assessment, in turn, measures the degree to which the localized end product does in fact comply with the specified requirements.

Assessing the quality of a localized product on the basis of the subjective expectations and/or preferences of a reviewer, rather than on the basis of formally specified requirements, is akin not merely to changing the rules in the middle of the game, but rather to allowing the rules of the game to be changed by each new player who enters the playing field. Identifying and documenting client needs, preferences, and expectations in the form of a client quality requirements specification during the project planning phase, before undertaking localization work, and measuring compliance with such requirements after the localization work has been completed, offers a solution to this problem, and a valid basis on which to empirically measure the quality of localized products.

References

ASTM International. 2006. *ASTM F2575-06. Standard Guide for Quality Assurance in Translation.* West Conshohocken, PA: ASTM.

Campbell, Chuck. 2005. "Making It Read Better: Syntactic Strategies." Presentation to the editors of the *Natural Resources Journal* at the University of New Mexico's School of Law. http://infohost.nmt.edu/~cpc/syntax/

Chandler, Heather Maxwell. 2005. *The Game Localization Handbook.* Hingham, MA: Charles River Media.

DePalma, Donald A. and Beninatto, Renato. 2003. *How to Avoid Getting Lost in Translation: Buying and Managing Language Services for Global and Multicultural Business.* Chelmsford, MA: Common Sense Advisory, Inc.

Dunne, Keiran J. 2006. "A Copernican Revolution: Focusing on the Big Picture of Localization." In *Perspectives on Localization*, Keiran J. Dunne (ed), 1–11. Amsterdam/Philadelphia: John Benjamins.

Dunne, Keiran J. 2006. "Putting the Cart before the Horse: Rethinking Localization Quality Management." In *Perspectives on Localization*, Keiran J. Dunne (ed), 95–117. Amsterdam/Philadelphia: John Benjamins.

Dunne, Keiran J. "Terminology: Ignore It at Your Peril." *MultiLingual* 18:3 (April/May 2007): 32–38.

Esselink, Bert. 2000. *A Practical Guide to Localization.* Amsterdam/Philadelphia: John Benjamins.

Guest, John. 2007. “What Are Good Metrics to Measure for Software Localization?” *LinkedIn Answers: Internationalization and Localization.* Feb. 22, 2009. http://www.linkedin.com/answers/international/internationalization-localization/INT_INZ/252322-10744690?browseCategory=

Heiman, Gary W. 2001. *Understanding Research Methods and Statistics: An Integrated Introduction for Psychology.* 2nd ed. Boston and New York: Houghton Mifflin.

Hall, Edward T. and Hall, Mildred Reed. 1990. *Understanding Cultural Differences*. Yarmouth, ME: Intercultural Press.

Hofstede, Geert H. 1991. *Cultures and Organizations: Software of the Mind*. London: McGraw-Hill.

ISO (International Organization for Standardization). 2008. *ISO 9001:2008(E). Quality Management Systems – Requirements*. 4th ed. Geneva: ISO.

ISO/IEC (International Organization for Standardization/International Electrotechnical Commission). 2008. *ISO/IEC 12207:2008(E). Systems and Software Engineering – Software Life Cycle Processes*. 2nd ed. Geneva: ISO/IEC-IEEE.

Kingscott, Geoffrey. 2007. “Translation Quality Assessment.” *Language International*. http://www.language-international.net/articles/Translation Quality Assessment (26-9-07).doc

Kirimoto, Yusuke. 2005. “Quality Programs in Localization Environments.” ENLASO. http://www.translate.com/Language_Tech_Center/Articles/Quality_Programs_in_Localization_Environments.aspx

Kohl, John R. 2008. *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market*. Cary, NC: SAS Institute.

Koo, Siu Ling and Kinds, Harold. “A Quality-Assurance Model for Language Projects.” In *Translating into Success: Cutting-Edge Strategies for Going Multilingual in a Global Age*, Robert C. Sprung (ed), 147–157. Amsterdam/Philadelphia: John Benjamins.

Lionbridge Technologies. 2009. “Localisation Testing.” http://www.lionbridge.com/lionbridge/en-GB/services/localization-translation/localization-testing.htm

LISA (Localization Industry Standards Association). 2004. *LISA QA Model 3.0. Product Documentation.* Féchy, Switzerland: LISA.

LISA (Localization Industry Standards Association). 2008. “About the Globalization Product Assessment.” http://www.lisa.org/Product-Information.932.0.html

Lommel, Arle R. and Ray, Rebecca (ed). 2007. *The Globalization Industry Primer: An Introduction to Preparing Your Business and Products for Success in International Markets*. Domaine en Praël [Switzerland]: LISA.

Luong, Tuok V., Lok, James S.H., Taylor, David J. and Driscoll, Kevin. 1995. *Internationalization: Developing Software for Global Markets.* [New York]: Wiley.

Marcus, Aaron. 2005. “User Interface Design and Culture.” In *Usability and Internationalization of Information Technology*, Nuray Aykin (ed), 51–78. Mahwah, NY: Lawrence Erlbaum.

Melby, Alan K. 2005. “Language Quality and Process Standards: LISA QA Model v3.0.” Language Standards for Global Business Conference, Berlin, Germany, Dec. 12, 2005. http://www.internationalization-conference.org/languagestandards/papers/Panel2_Melby.pdf

Microsoft Corporation. 2009. “Microsoft Language Portal: Style Guide Download.” http://www.microsoft.com/language/en/us/download.mspx

PMI (Project Management Institute). 2000. *A Guide to the Project Management Body of Knowledge (PMBOK® Guide).* 2000 ed. Newton Square, PA: Project Management Institute.

Pringle, Alan S. and O'Keefe, Sarah S. 2003. *Technical Writing 101: A Real-Word Guide to Planning and Writing Technical Documentation.* Research Triangle Park, NC: Scriptorium.

Rätzmann, Manfred and De Young, Clinton. 2003. *Software Testing and Internationalization.* Salt Lake City, UT: Lemoine International.

Rayner, Keith and Sereno, Sara C. 1994. "Eye Movements in Reading: Psycholinguistic Studies." In *Handbook of Psycholinguistics*, Morton A. Gernsbacher (ed), 57–81. San Diego, CA: Academic Press.

SAE International. 2001. *Translation Quality Metric [J2450].* Warrendale, PA: SAE.

Smith-Ferrier, Guy. 2007. *NET Internationalization. The Developer's Guide to Building Global Windows and Web Applications.* Upper Saddle River, NJ: Addison-Wesley.

Symmonds, Nick. 2002. *Internationalization and Localization Using Microsoft .NET.* Berkeley, CA: Apress.

Tek Translation. 2009. "Services: Testing." http://www.tektrans.com/services.asp?subsection=19

Urien, Emmanuel, Howard, Robert and Perinotti, Tiziana. 1993. *Software Internationalization and Localization: An Introduction*. New York: Van Nostrand Reinhold.

W3C (World Wide Web Consortium) Schools. 2009. "OS Platform Statistics." Jan. 14. http://www.w3schools.com/browsers/browsers_os.asp

Wright, Sue Ellen. 2001. "Terminology and Total Quality Management." In *Handbook of Terminology Management,* Vol. II, Gerhard Budin and Sue Ellen Wright (eds), 488–502. Amsterdam/Philadelphia: John Benjamins.

PART III

Professional certification

Lessons from case studies

The predictive validity of admissions tests for conference interpreting courses in Europe

A case study

Šárka Timarová and Harry Ungoed-Thomas
Lessius University College / Université de Genève

Admissions tests are an integral part of conference interpreter education, yet little is known about their effectiveness and efficiency. We discuss general principles of admissions testing, focusing specifically on predictive validity and on measuring aptitude, the main component of interpreter training program admissions tests. We describe the underpinnings of developing admissions tests with high predictive validity of end-of-course exam performance. We then evaluate and report the efficiency of an existing aptitude test by looking at historical records of admissions testing results and end-of-course exam results in one interpreting school. Multiple linear and logistic regression analyses indicate that these tests are poor predictors of students' success rate. Future research should focus on developing tests with better predictive validity assessed on empirical grounds.

Introduction

Spoken language conference interpreting became prominent after World War II, and conference interpreters have since provided their services in a variety of settings, typically involving high-level political meetings and scientific conferences. They have been working for clients, from private companies to governments, to large international organizations, such as the United Nations and the institutions of the European Union. Conference interpreting, especially the simultaneous mode, was once thought to be so complex that there were doubts about whether it could be taught at all, and that one must possess a special innate talent to be able to perform the task. The general belief nowadays is that interpreters are made not born (cf. Mackintosh 1999), and the demand for conference interpreter education has resulted in a larger number of more sophisticated educational programs.

Interpreter education is an interest shared by teachers and academics, large-scale employers (above all international organizations), researchers, and professional associations. AIIC (Association Internationale des interprètes de conférence, International Association of Conference Interpreters), the only worldwide professional association for (spoken language) conference interpreters, has been active in the area of interpreter education for decades. It organized its first symposium on conference interpreter "training" in 1965 (Mackintosh 1999), and offers regular opportunities for conference interpreter educators. AIIC has also developed a standard for conference interpreter education programs.

Until recently, the standards specified that "Admission to CI training shall be on the basis of an entrance test, which verifies language skills, cultural and general knowledge, and aptitude for interpreting" (Mackintosh 1999: 72). This requirement has been recently revised to a *recommendation* to administer an entry-level test, which should ideally demonstrate the candidate's readiness to start interpreter education (AIIC Training Committe 2006). This goes hand in hand with the AIIC Training Committee's acknowledgement that there is currently no reliable test of aptitude for interpreting. The current practice remains to administer an admissions test (AIIC 2006), but despite the highly selective process, the ratio of successful graduates to admitted students remains low (Timarová & Ungoed-Thomas 2008). Very little research has been carried out in the area of predictive validity of admissions tests, and some of the scarce data available show that performance on admissions tests may not be a fully reliable predictor of a candidate's performance in and successful completion of a training program (Gringiani 1990; Tapalova, as cited in Sawyer 2004; Taylor 1997).

The aim of this paper is to start exploring in more detail admissions testing for spoken-language conference interpreter education programs which are assumed to be compatible with the AIIC standards (AIIC Training Committee 2006). In the following section, we will discuss basic concepts of admissions testing, and focus on the link between aptitude and admissions tests as well as on the issue of predictive validity. We will then review previous research in the area of interpreting aptitude test development and present our own analysis of predictive validity of current admissions tests at an interpreting school. Finally, we will discuss the results and their implications, and make suggestions as to future directions of interpreting aptitude research.

Admissions testing

It is assumed that the goal of admissions tests is to select candidates who have the potential to successfully complete a particular educational program. Carroll

(1962) listed five criteria crucial for acquiring a skill. The criteria are: aptitude, general intelligence, time dedicated to training activities, quality of instruction, and time provided for learning. Parry & Stansfield (1990) present a collective volume of papers that considers additional factors, such as personal learning styles, personality of the teacher, classroom environment, level of anxiety, attitude, motivation, brain hemisphericity, and their effects on second language learning. Based on these criteria for skill acquisition, we could arrive at a crude equation:

admissions tests = aptitude + affective variables + curriculum-related factors

Despite the multitude of factors listed as relevant for skills acquisition, the literature in conference interpreter education discusses admissions tests almost exclusively in terms of aptitude, often limited to cognitive and linguistic abilities, such as ability to analyze text, verbal fluency, and memory (Alexieva 1993; Lambert 1991; Longley 1989; Moser-Mercer 1985). In a survey of interpreting schools (Timarová & Ungoed-Thomas 2008), our finding was that the participating interpreting schools test predominantly for skills considered to be directly related to interpreting (language and communication, comprehension, analytical skills, general knowledge), using short consecutive interpreting exercises, summary exercises, written translation, and interviews. Affective variables, such as personality or motivation, are sometimes considered but not formally assessed, and instruction related factors (e.g. correspondence between the school's teaching and the candidate's learning styles) do not seem to be considered at all (Timarová & Ungoed-Thomas 2008).

Since aptitude tests are at the center of admissions testing for conference interpreter education programs, it is pertinent to consider what constitutes aptitude for interpreting. While the term 'aptitude' is used frequently in interpreting studies literature (e.g. Alexieva 1993; Bowen & Bowen 1989; Lambert 1991; Longley 1989), authors hardly ever provide a definition for aptitude. It is usually understood as a pre-requisite for education. However, the pre-requisites are based almost exclusively on intuition and the experience of educators (Gringiani 1990; Lambert 1991; Moser-Mercer 1994; Russo & Pippa 2004).

Bowen & Bowen (1989) refer to Carroll's (1962) definition of aptitude. This definition considers aptitude to be the amount of time necessary for learning a given task or acquiring a given skill. According to this definition, everybody is capable of achieving the goal, but some people require more time than others to do so. This definition was developed for foreign language learning in the 60s. For conference interpreter education purposes, we would like to propose an adjusted version of Carroll's definition of aptitude. Aptitude in this chapter is defined as the capacity to acquire consecutive and simultaneous interpreting skills to a criterion level within a period of time corresponding to the length of a particular interpreting program, where the criterion is the ability to provide interpreting of a quality acceptable for entry into the profession. Conference interpreter education is typically offered as a one- or two-year master-level university course, so skills acquisition is assumed to be possible within this limited period of time, plus a possible extension beyond the period in which classes are offered. For example, students may be allowed to take their exams up to one year after the study period has officially finished. As Bowen and Bowen (1989) point out, the time factor in the definition effectively means that different schools will have to develop their own admissions tests depending on the parameters of their program. However, there are other more serious problems with admissions tests for interpreting programs based on aptitude: there is no reliable evidence of what aptitude for interpreting is, above and beyond practitioners' intuitive views, as aptitude tests have not been consistently researched and validated (Moser-Mercer 1994; Sawyer 2004).

Basic concepts of aptitude testing

The basic aim of an aptitude test is to *predict* whether an individual will be able to acquire a skill (Boyle & Fisher 2007). Predictive validity thus becomes the most important feature of aptitude tests. The predicament is how to measure a skill that the individual has not acquired yet. This is the crucial difference between aptitude tests and other forms of assessment, such as achievement tests and ability tests (Boyle & Fisher 2007). To use a simple analogy, we can find out easily whether or not a child can ride a bike by simply asking her to ride it (achievement). However, how do we find out whether a newborn baby will be able to ride a bicycle in the future (aptitude)? In this case, seating her on a bicycle will not help. And this is the basic conundrum of aptitude testing. How do we assess potential for skilled behavior without having to perform the skilled behavior? This is where we see a serious gap in current research, which needs to be addressed.

Determining aptitude for future skills requires identifying key elements of that skill and testing for the *elementary underlying* abilities. If the skilled behavior builds on these elementary abilities, they should be present even in the absence of skilled behavior, and should be measurable. To continue with the baby and bike analogy, we may be looking for basic motor coordination and balance. These abilities can be measured even in small children, who are too young to be able to ride a bicycle. This approach is rare in interpreting. The interpreting literature often advocates or rejects tests based on their face validity, that is, on their resemblance to interpreting. For example, Bowen and Bowen (1989) rejected vocabulary tests in part on the grounds that, the tests "do not give any indication of the candidate's comprehension of text or of writing ability" (ibid: 112). Yet there is some evidence that word knowledge is a variable that may be an important component of interpreting (Padilla, Bajo & Macizo 2005). Similarly, Setton (1994) argued for cloze tests which require completing a sentence, rather than filling in a missing word, while Gerver, Longley, Long & Lambert (1989, see more details below) showed that one-word cloze tests could predict final exam interpreting performance. Based on research in language testing (e.g. Bachman & Palmer 1996), Angelelli (2007) justified the principle of authenticity, which requires that the test correspond to real-life situations. This leads to candidates being tested on their ability to interpret as a prerequisite for their admission into an interpreting program. Such a requirement is very common as evidenced by the high incidence of short consecutive interpreting tests being part of admissions tests (Timarová & Ungoed-Thomas 2008). However, this places a special demand on candidates with no previous interpreting experience by asking them to perform a task they have very likely never performed or practiced before. We believe such an approach constitutes a test of current ability (can a baby ride a bike?), but runs contrary to the idea of aptitude testing (will a baby be able to ride a bike in the future?).

To borrow an example from a related discipline of what we believe is a better approach to determining aptitude, foreign language learning boasts several tests of language learning aptitude. As mentioned previously, a classic test was developed by Carroll (1962), who found four predictors of language learning aptitude: phonetic coding ability, grammatical sensitivity, rote memory, and inductive learning ability. A similar test was developed by Pimsleur (1963). For example, a task in these tests may present the testee with a very simple sample sentence (subject – verb – object) in an imaginary language. The sentence is explained and testees are then asked to indicate, in a multiple choice format, correct translations of other similar sentences. Note that the task does not ask for performance in authentic language use circumstances and that the test would not satisfy the authenticity/face validity principle advocated in interpreting studies. The test does, however, simulate the instruction environment, and it is strongly related to the learning experience, the skills acquisition stage, rather than the target performance. These tests have been shown to have a predictive validity of people's ability to learn a foreign language. Since the 1960s, much research has been carried out in the area of foreign language learning. The concept of aptitude for foreign language learning (in the narrow psycholinguistic sense) has been expanded to include affective variables (personality, motivation, anxiety, learning styles, etc.), but none of these variables has substantially improved the predictive validity of Carroll's test (Sparks and Ganschow 2001). More research along these principles needs to be carried out in order to determine aptitude for interpreting, both in terms of the interpreting skill itself and the ability to acquire the skill (cf. Shaw, Timarová & Salaets 2008). We will now discuss two lines of research reported in interpreting studies literature: research focusing on developing new aptitude tests, and validation of current admissions tests and their usefulness as predictors of final exam performance.

Development of new aptitude tests

This line of research focuses on development of new tests of various skills and abilities, which are assumed to be closely related to interpreting skills. Performance on the tests is then compared to scores achieved on aptitude tests employed in educational practice and/or end-of-course exams. Much interesting research of this type has been carried out at SSLMIT (Scuola Superiore di Lingue Moderne per Interpreti e Traduttori – Faculty of Interpreting and Translation Studies) in Trieste, Italy (Russo 1989, 1993; Pippa & Russo 2002; Russo & Pippa 2004) and research has also been carried out at the Institute of Translation Studies in Prague, Czech Republic (Rejšková 1999).

The Institute of Translation Studies, Prague, Czech Republic, offers a five-year masters program. In the first three years, students take courses on general subjects (languages, linguistics, translation theory), and in the last two years they specialize in translation or interpreting. All students, regardless of their wish to specialize in translation or interpreting in the fourth and fifth year, must take an introductory consecutive course in the third year. Performance in this course serves as a basis for recommendation to pursue (or not to pursue) the interpreting specialization. To find out if consecutive interpreting skills can predict simultaneous interpreting skills, Rejšková compared student performance on the end-of-course examination in the introductory consecutive course with performance on a battery of six tests, which she designed in order to assess aptitude for simultaneous interpreting. These tests included:

1. Shadowing (Lambert 1992): Students listened to a short speech in a foreign language and simultaneously repeated verbatim what they heard.
2. "Personalized" cloze test: Students listened to a short piece of text in which the speaker provided some basic details about himself. Students repeated the text verbatim in the foreign language and replaced all personal information – name, age, nationality, etc. – with information about themselves.
3. Interpreting from a foreign language into the mother tongue of a simple text designed specifically for the aptitude test.
4. Interpreting from a foreign language into the mother tongue of a specifically designed procedural text.

5. Interpreting from a foreign language into the mother tongue of a fairy tale with a twist.
6. Interpreting from a foreign language into the mother tongue of an authentic conference speech with high redundancy.

Comparison of exam grades and these simultaneous interpreting exercises correlated only weakly (r = .498). Rejšková concluded that performance in consecutive interpreting is not a reliable predictor of future performance in simultaneous interpreting and that aptitude for simultaneous performance must be tested separately (cf. Alexieva 1993). Regretfully, Rejšková did not collect further data to compare the scores with actual performance in the following (simultaneous) interpreting courses, which may have allowed her to evaluate the predictive validity of her battery of tests in terms of aptitude for simultaneous interpreting.

The creation of a standardized, valid and reliable test of interpreting aptitude has been pursued by Russo (1989, 1990, 1993; Pippa & Russo 2002; Russo & Pippa 2004) at SSLMIT in Trieste. Based on theoretical assumptions of common underlying cognitive and linguistic components and processes, Russo has been using paraphrasing as her basic method for assessing students' aptitude for simultaneous interpreting. Her most important aim is to establish whether there is a reliable association between paraphrasing scores and later performance in interpreting courses. The latest available results (Russo & Pippa 2004) report on a sample of 46 students who took the test at the beginning of their studies. Their scores were analyzed in relation to their final exam scores and the number of semesters it took them to get through the school, two criteria which have been previously shown to be related (Russo 1993). The authors found that there was a significant negative correlation of the results with study time ($r(44) = -.32$, $p < .05$), meaning that students who scored better on the test graduated earlier. There was also a positive correlation with exam results ($r(44) = .38$, $p < .05$), which indicates that students who scored better on the test achieved better exam results.

A psychological approach to measuring aptitude was adopted by Gerver, Longley, Long & Lambert (1989). In 1977, the authors set out to "evaluate objective tests which were intended to assess interpreting candidates' ability to grasp rapidly and convey the meaning of spoken discourse" (1989: 724). The tests were not part of the admissions procedure at the interpreting school (Polytechnic of Central London), but were given to all of the students who were admitted to the school in 1977. The students scoring low marks for the tests therefore still went on to take their final exams.

The authors designed two groups of tests and tested their validity by administering them alongside existing admissions tests and by correlating them with final exam performance. A first group consisted of these text-based tests:

1. Logical memory test (Wechsler 1945): Tests were chosen from the Wechsler Memory Scale, each test consisting of a short text of 65 words divided into 24 memory units. Students listened to the text and then were evaluated on the number of memory units they successfully recalled.
2. Text memory test: A speech of 1000 words was read aloud, students were told to listen and, once it was over, instructed to write an information summary.
3. Cloze test: The task was to restore words that were missing from taped speeches.
4. Error detection test: Students were asked to correct an auditory text with around 50 intentional lexical, syntactic and pronunciation errors.

The second group consisted of sub-skills based tests:

1. Synonyms production test: Students were asked to write as many synonyms as possible for four words.
2. Expressional fluency test: Students had to rewrite sentences according to specified criteria.
3. Vocabulary test: This was a multiple-choice synonym test.
4. Speed stress test: Students had to complete mental acuity tests under time pressure.

The authors found that the scores of some, but not all, of these tests predicted final exam scores. Overall, the text-based tests were the best predictors, with the two cloze tests having the highest correlation with the final simultaneous exam (r = .44 and .56) and the two logical memory tests correlating significantly with the final consecutive exam (r = .48 and .63). Of the sub-skill based tests, only the synonyms test correlated significantly (r = .50) with a final exam result for a consecutive interpreting exam.

Gerver et al. went on to state that if these results were accurate, the prediction accuracy of admissions tests could be raised from 59% to 75% with the introduction of text memory tests, cloze tests and error detection tests, and to 94% with the introduction of a logical memory test. The authors concluded that the processing of connected discourse constitutes a crucial feature of the interpreter's task which needs to be embodied in selection tests. It is very surprising to see that this research has received very little follow-up in interpreting research. Even if the estimate of prediction accuracy of Gerver et al. was inflated, the tests are still worthy of closer scrutiny. They also demonstrate how fruitful it is to look for simpler tests of underlying abilities, which do not necessarily bear superficial resemblance to interpreting, i.e. lack face validity.

Validation of existing admissions tests

Another way to approach the issue of effectiveness of admissions tests is to look at the current admissions testing practices and determine whether they predict student success. One such study was reported by Gringiani (1990) from SSLMIT in Trieste, Italy. Similarly to the Institute for Translation Studies in Prague, SSLMIT offers an undergraduate program with initial education in translation and subsequent education in interpreting. It is constrained by the Italian educational settings in that admission into the interpreting specialization late in one's studies is a student's decision. While the school conducts interim aptitude tests, these have very limited power, and serve only as a guide. In practice, students' performance on the tests cannot prevent them from pursuing interpreter training. This creates a unique research opportunity: by effectively admitting candidates who are not deemed to have the necessary skills, the researchers have a valuable opportunity to put the judgment of the juries to the test. As Gringiani (1990) reports, this judgment appears to be fallible. Gringiani compared aptitude test results with actual performance on an end-of-course examination, and found that out of 25 students who failed the test, 7 successfully completed the course, achieving more or less the same grades on their examinations as those students who were considered to be better equipped to become interpreters, according to aptitude test results. In three cases, students who failed the admissions tests completed the course with fewer re-sits than those students. On the other hand, out of 17 students who successfully passed the aptitude tests, 8 withdrew without completing the course of study. There is, of course, a host of possible explanations (confounding variables) for these outcomes. The point here is that these tests failed to have a high predictive validity.

The study reported by Gringiani exemplifies research based on actual aptitude test scores and actual examination performance. The value of such an approach lies in the fact that a comparison of the two results provides some indication of how useful the administered admissions tests are to this particular program. The disadvantage is that these comparisons will tell the researcher nothing about the nature of the tests. For example, it will not be possible to determine which part of the aptitude test is or is not valid (whether it measures what it is supposed to measure). Another disadvantage is that such research is likely to be closely associated with a particular interpreting program (curricula, instructors, etc.), and the results gleaned cannot be directly transferable to other schools and other programs. This is an inherent limitation of conducting a correlation analysis in isolation: The simple fact of establishing a relationship between two variables does not establish a causal relationship. The causes of the correlation in this cited study are unknown. However, determining the relationship between the admissions tests and the final exam results can be used as one tool towards collecting data to determine the effectiveness of the admissions tests in their current form – that is, their predictive validity.

It is worth noting, however, that the correlational analysis did serve to point to the need for a program evaluation that considers all facets of the interpreter education program. Aptitude testing is one of those facets. As Sawyer (2004: 111) points out, schools may administer admissions tests in order to merely determine readiness to start a program, without any ambition to predict completion. However, as stated above, the interpreting literature discusses aptitude as the main component of admissions tests, which implies interest in selecting candidates with a high potential to complete the program successfully. If that is the case, such schools will need to validate their admissions procedures. The principles outlined in the introductory part of this paper provide a framework for the development of an admissions test which measures aptitude, personality, motivation and other traits that may contribute to success in an interpreting program. As a result, such a test should be able to help discriminate above and beyond the chance level between those candidates who will complete the program, and those who will not, assuming success is also measured in a valid and reliable way. To our knowledge, the above mentioned study by Gringiani (1990) and a similar one by Tapalova (as cited in Sawyer 2004) are the only available studies on validation of conference interpreting admissions tests. Both concluded that the admissions tests investigated had low predictive validity. Clearly, the issue of effectiveness of current admissions practices needs more empirical attention.

In what follows, we will present our own investigation into the predictive validity of admissions tests based on available historical records provided by an interpreting school. Our concern lies with determining the predictive validity of the admissions tests; that is, we wish to find out whether or not the tests help reliably select candidates with a high potential to successfully complete the program.

Admissions tests and final exams at Wilhelm University

Methods

Wilhelm University, a pseudonym for a European conference interpreter education school, kindly made its records available for this research project. For methodological purposes, we can only present an analysis of one school's data, but it should be said that Wilhelm University satisfies the criteria set by AIIC for interpreting programs.[1] Also the type of admissions procedures and the sorts of tests employed are deemed by this research team to be representative of other schools' admissions tests (Timarová & Ungoed-Thomas 2008).

Since 1998, Wilhelm University has been operating a two-stage entrance exam for its one-year graduate conference interpreting course. As in other schools (Gringiani 1990; Lambert 1991; Moser-Mercer 1994; Russo & Pippa 2004), the tests are intuitive, based on extensive training experience of the faculty. The first is a written exam where the candidate student is required to complete translations on a non-technical subject into her active language(s) (350 words into an A[2] language, 250 words into a B language). If candidates pass this test, they are invited to take the oral admissions test in the presence of a panel of at least three jurors. All jurors are experienced professional interpreters and typically interpreter trainers at Wilhelm University. The jurors discuss the candidate's performance and make decisions based on group consensus. The oral test consists of two parts and lasts between 30 and 60 minutes, depending on the student's language combination (that is, a student with three working languages will necessarily require more testing time than a student with two languages). In the first part, the student is required to summarize, in her active language(s), short oral presentations of three minutes given in her passive languages. In the second part, the student is given a choice of three subjects and is required to prepare her own presentation of three minutes in her mother tongue, for which she is given five minutes of preparation time.

Until 2005, once students were admitted to Wilhelm University, they had two semesters in which to prepare for their final exams. If they did not pass all of their exams on the first attempt, they were allowed two further attempts. If they were unsuccessful on the third attempt, they were dismissed from the program. Students graduate from Wilhelm University when they pass all of the exams. To do so, they have to score 40 or above (out of 60; this is the assessment scale applied in the country where the university is located) in three exam components for each

[1] The most important standards are: the program is taught preferrably at a graduate level; admission into the program is ideally on the basis of an admissions test; an undergraduate university degree is required as a prerequisite for admission; the curriculum includes training in consecutive and simultaneous interpreting, and has a theoretical component; faculty members are practicing conference interpreters, who ideally have some teacher training; final examination juries are composed of course instructors and external examiners (practicing conference interpreters) (AIIC Training Committee 2006).

[2] A language – mother tongue, B language – non-native language, of which the interpreter has near-native command, the interpreter works from B to A language, and also from A to B language, C language – the interpreter works from C to A language. A and B are also known as active languages, Cs as passive languages. For more details see the AIIC classification at http://www.aiic.net/glossary/default.cfm.

of their working languages. The exam components are consecutive interpreting, simultaneous interpreting of a spontaneously delivered speech, and simultaneous interpreting of a read speech for which the text was made available to the interpreters at the time of interpreting.

We wanted to assess how well the admissions tests selected successful graduates by determining whether there was a direct relationship between a candidate's result on the admissions tests and the final exam. For the purposes of this analysis, the data included were restricted to results for the first attempt at final exams, based on the assumption that students selected are considered to be trainable within the standard length of the program. Also, there are periods of up to 16 months between retakes, and we did not consider data from exams taken more than a year apart to be equivalent.

Data collection

The data used for this analysis consist of examination and test results from written and oral admissions tests and first attempts at final exams at Wilhelm University. The results cover a period from 1998, when the current entrance exams were introduced, to 2005. The records are extensive, but not complete, as it proved difficult to find results for all evaluations for each year. We were able to collect records for the admissions tests and/or final exams of a total of 184 participants. (It is estimated that around 1,000 candidates applied during the time period, of which some 170 were admitted). There were several different kinds of records depending on how far a student advanced. Table 1 is an example of the four different kinds of records included in our analysis. The data from the aforementioned 184 students were classified according to these four types. Record Type 1 represents students who progressed all the way to the final exams (40 being the minimum grade required for a pass), Record Type 2 represents students who failed the orals, Record Type 3 represents students who failed the written admissions test, and Record Type 4 represents students who graduated from the translation program at Wilhelm University, and so were exempt from the written test and proceeded directly to the oral admissions test.

Students Type 1 passed both the admissions tests and took the final exams. Although in the example given in the table, the student actually failed the C2-A simultaneous (i.e. second foreign language into mother tongue) final exam (achieving a non-passing grade of 38), which later will be retaken, this is as far as we decided to follow her. Other students of Type 1 were successful on all exams, others may have failed. Importantly, we had 35 such records with data from all of the tests (admissions and final) and used them to analyze the correlation

Table 1. Example of student records

Record type (language combination)	**Written admissions test**	**Result**	**Oral admissions test**	**Result**	**Final exam (1st attempt)**	**Result**
Type 1	Translation		Oral summary		C1-A sim	40
ACC	C1-A	45	C1-A	40	C2-A sim	38
(n = 35)	C2-A	40	C2-A	45	C1-A con	42
			Presentation A	49	C2-A con	45
Type 2	Translation		Oral summary			
AA	A1-A2	42	A1-A2	35		
(n = 66)	A2-A1	48	A2-A1	42		
			Presentation A	45		
Type 3	Translation					
ABC	A-B	40				
(n = 45)	B-A	35				
	C-A	32				
Type 4			Oral summary			
ACCC			C1-A	40	C1-A sim	50
(n = 23)			C2-A	43	C2-A sim	51
			C3-A	34	C1-A con	50
			Presentation A	50	C2-A con	44

Languages: A – mother tongue, B – active working language, C – passive working language
sim = simultaneous, con = consecutive interpreting

between written, oral, and final exams. Students Type 2 passed the written exams and failed the oral exams, and so were not admitted to the school. We had 66 such records which served to analyze the relationship between written and oral tests scores. Students Type 3 failed the written test. Those records are of no use to our purposes. There were 45 such records. Students Type 4 proceeded directly to the oral tests, were admitted, and passed the final exams. We have included their data to analyze the relationship between the oral and final exams. We have 23 such records, some of which come from students who proceeded directly to the final exams, some of which come from records where the scores for the written entrance exams were missing. There were also 15 records with data for written tests and final exams with no data for oral tests.

Results

A number of regression models were estimated, using simple and multiple linear and logistic regression. Selected examples are shown in Table 2.

Each row in Table 2 represents a regression model. The first column specifies the dependent variable; that is, the variable we want to predict. The second column lists the regressors; that is, variables on the basis of which we are making the prediction. The third column lists the standardized regression coefficients which tell us about the relative importance of each predictor: in other words, how much it contributes to our prediction, and whether the contribution is significant (marked by an asterisk). The fourth column contains the value of adjusted R^2, a measure of how well the model predicts the dependent variable. Finally, the fifth column shows the size of the sample included in the regression model.

As an example and to clarify further, in the first row we used the average grade for the written admissions test (across the candidates' languages), the average grade for oral summaries, and the average grade for oral presentations to predict final average grade. Beta coefficients for the three predictors show that only the oral summaries made a significant prediction of the dependent variable. However, the three predictors taken together explained a mere 10.8% variation in the final grade, which is considered to be a poor fit.

The most important information can be found in the fourth column, which indicates the strength of the regression model. The adjusted R^2 value can be multiplied by 100 for convenience to indicate the percentage in the dependent variable variation explained by the predictors. For all the models we estimated, the fit is very poor. The predictor variables explain typically less than 10% of the variation in the dependent variable, and with the exception of prediction of final simultaneous with text in the first foreign language, do not reach significance. This interpretation is supported by the bivariate correlation coefficients listed in Table 3. The table shows a matrix of the individual components of admissions tests and final exams, and also composite scores (so that "average oral" gives the average grade for all oral components of the admissions tests). It is apparent that the individual components of admissions tests are related to each other, and so are the individual components of final exams. However, the relationships between the admissions tests and the final exams are not strong.

The predictions do not become stronger nor are they significant with the use of cruder measures, such as simple pass-fail predictions, as opposed to more fine-grained predictions of specific grades (the regression models are basically the same as those listed in Table 2). Based on these analyses, we have to conclude that the present composition of the admissions tests does not predict well the outcome of the final interpreting exams.

Table 2. Selected linear regression models predicting final exam results

Dependent variable	Predictors	Beta	Fit (Adjusted R^2)	Sample size
Average final grade	average written average oral summary average oral presentation	.084 .424* –.066	.108	35
Average final consecutive grade	average written average oral summary average oral presentation	.077 .335 –.002	.047	35
Average final simultaneous grade	average written average oral summary average oral presentation	.490 .405* –.126	.069	34
Final consecutive grade, 1st language	written 1st lang. oral summary 1st lang. oral presentation 1st lang.	–.076 .095 .011	–.082[†]	34
Final simultaneous grade, 1st language	written 1st lang. oral summary 1st lang. oral presentation 1st lang.	.235 .339 –.155	.068	33
Final simultaneous with text grade, 1st language	written 1st lang. oral summary 1st lang. oral presentation 1st lang.	.330 .391* –.123	.172*	33

*p < .05; [†]Negative value generated by SPSS

Table 3. Correlation matrix of main predictors and dependent variables

	Average written	Oral summaries	Oral presentation	Average oral	Average admission	Consecutive	Simultaneous
Oral summaries	.287**						
Oral presentation	.281**	.695**					
Average oral	.301**	.953**	.873**				
Average admission	.779**	.871**	.795**	.910**			
Consecutive	.085	.182	.205	.231	.147		
Simultaneous	.186	.266*	.113	.255	.236*	.580**	
Average final	.158	.251	.188	.276*	.216	.906**	.870**

* p < .05, ** p < .01

Discussion: Issues raised

There are at least two ways that we could judge the effectiveness of these admissions tests. On the one hand, we could measure predictive strength of the candidates' scores on admissions tests towards final exams, as we have done above. This analysis would lead us to believe that the admissions tests only weakly predict final exam performance. On the other hand, we could ask how many of the students selected by the admissions tests actually pass the final examination? In the case of Wilhelm University, this figure is higher than 70%, which might suggest these particular admissions tests are very effective at predicting which students will pass at this particular school (Timarová & Ungoed-Thomas 2008). How can we explain this paradox?

Unfortunately, our research design does not allow us to separate these issues. Other methods and other types of data would have to be used to that end, an obvious area for future research. Having said that, we consider some possible explanations for the lack of predictive validity of this particular admissions test, each of which warrant further research.

First of all, the data above show the overall pass rate – including students who retake the exam. However, our analysis focused on the pass rate only for first attempt at the final exams. In this respect, an analysis of the student records shows that only 17% of students pass the final exams on the first attempt in all exam components and in all their languages, with a further 30% of students who pass on the first attempt for all exam components in at least one of their languages. Altogether, the partial pass rate for the first attempt at final exams is thus around 47%. Another plausible explanation is that Wilhelm University is turning down a number of students who have the potential to graduate from its course. By sorting candidates on the basis of test scores with weak predictive validity, and then selecting only the strongest performers, the school achieves a high pass rate, but other potential graduates are possibly being rejected.

The results of the present analyses raise several important theoretical and methodological issues. First, if the admissions tests do not predict the final outcome, what do they measure? Despite the low predictive validity, the tests are not necessarily without a purpose. They apparently do not discriminate between successful and unsuccessful graduates, but they may still measure very important skills. For example, the oral presentation test may not distinguish between successful or unsuccessful students, but since all interpreters need to have good presentation skills (which can be considered a component of aptitude), it weeds out candidates who would not satisfy this basic pre-condition to interpreting. The level of presentation skills in all admitted students may therefore be so similar that the variable cannot make a unique contribution to the differences between those who pass and those who fail.

Conclusion

Directions for future research

Program evaluation for interpreter education programs is essential, and research needs to continue to be conducted in this area. To illustrate the urgency of the issue, among the 16 programs in Western, Central, and Eastern Europe, and 2 outside of Europe (Asia and North America) surveyed by Timarová & Ungoed-Thomas (2008), only two schools have an overall pass rate higher than 80%, with another three at 70% or above. For some schools, the pass rate was found to be as low as 20%. In the present analysis, we have seen that a program with an overall pass rate of 70% can have a pass rate on the first attempt at exams, that is, at the end of the standard length of the program, as low as 17%. Empirical research is needed to interpret these statistics to inform test and curricula design.

Future research will need to take into consideration several issues. First among them is the issue of validity. More research needs to focus on finding good predictors of final exam results, perhaps using methods and tests as described in the introductory part of this paper. The tests should be ideally based on empirical evidence rather than only on face validity, i.e. superficial resemblance to interpreting and intuitive appeal. The tests used by Wilhelm University are very commonly used by other schools, which have similar overall pass rates (Timarová & Ungoed-Thomas 2007), and while we examined data from a single school, it might not be unreasonable to postulate that other schools' records would produce similar results, as they are typically based on intuitive practice and similar rationale, and lack validation.

Any new admissions tests should aim at assessing skills related directly to interpreting and/or the ability to acquire the interpreting skill. However, the tests should not make unreasonable expectations of candidates who have not been trained in interpreting yet. Administering consecutive interpreting exams to candidates may lead to circular argumentation: only candidates capable of consecutive interpreting will be admitted to study consecutive. As Moser-Mercer (1994) says, such an approach only makes sense if we believe interpreters are born, not made. It would be preferable for researchers to come up with tests that target the appropriate latent constructs (i.e. something inherent in developing a particular interpreting skill) while administering a task that the candidate can be reasonably expected to complete without prior training. Such potential is, in our opinion, in the research reported by Gerver et al. (1989).

Secondly, for reasons of economy and ease of administration, new tests should make a unique contribution to prediction. Interpreting is a very complex task, and it is not reasonable to expect that one supertask will be found which will serve as the sole predictor. Rather, a good mix of simple tasks (both in terms of completion for the candidate, and in terms of administration and scoring for the testing institution) should be found, where each task makes a valid contribution to the overall prediction. This may result in better fitting regression models than those reported in the present study.

Limitations

Methodologically, it is also important to keep in mind that significant correlations are not always the ultimate and only goal, as mentioned previously. It is the size of the relationship (measured by squared correlation coefficient) that reveals the true contribution of the variable to the variation in the dependent variable. For example, a relationship with a correlation coefficient r = .3 will account for $r^2 = .09$, or 9%, of the variation in the predicted variable, independently of the correlation being statistically significant. Also, it must be remembered that a correlation does not indicate a causal relationship, regardless of size: causes remain unknown. Studies in real educational settings, working with real students, and analyzing their real test scores are necessary.

We have attempted here to look at general trends, as an exploratory approach to understanding aptitude testing in interpreting programs. As we stand, this is the third study to produce empirical evidence that current admissions tests are nor strongly predictive of successful completion (as measured by course grades or scores on exit exams).

Research in aptitude for foreign language learning shows that it is possible to develop robust tests of aptitude. Carroll's test referred to above continues to be a test of aptitude independently of type of instruction, language, or age of the students. We believe it should be possible to find aptitude tests for interpreting which will demonstrate similar robustness, and will be applicable in different educational settings.

Most importantly, this line of research depends heavily on good quality data. In the preparation stage for the present study, we were granted permission to use historical records from three schools, but only the data of Wilhelm University was in a form that allowed analysis. Even then, there were a number of records with missing data. Schools need to keep more detailed and better managed records of both the admissions tests and final results if the relationships between the two are to be analyzed. In addition, to be able to interpret correlational analyses, researchers should be provided with testing materials that they can subject to analysis.

In conclusion, admissions tests for interpreting courses present a challenging line of research, but also one with very immediate effects for applied practice. They deserve the attention of the training, education, and research community, and much more focus in the future. We strongly support exploring alternative tests of interpreting aptitude that could complement and/or replace existing admissions tests. Of course, it is still possible that schools are not interested in admitting students with the highest potential to successfully graduate, but those who have skills enabling them to start an interpreting program (Sawyer 2004: 111). Naturally, it is within each school's discretion to specify its own admission criteria, as well as criteria for the award of a diploma. In such cases, however, schools need to be aware of the limitations of their admissions procedures, and should refrain from generalized claims about measuring candidates' interpreting aptitude.

References

AIIC. 2006. "*AIIC Directory of Interpreting Schools*." Available at http://www.aiic.net/schools/ (last accessed June 27, 2007).

AIIC Training Committee. 2006. "*Conference Interpreting Training Programmes: Best Practice*." Available at http://www.aiic.net/ViewPage.cfm/page60 (last accessed November 17, 2008).

Alexieva, Bistra. 1993. "Aptitude Tests and Intertextuality in Simultaneous Interpreting." *The Interpreters' Newsletter* 5: 8–12.

Angelelli, Claudia. 2007. "Assessing Medical Interpreters. The Language and Interpreting Testing Project." *The Translator* 13 (1): 63–82.

Bachman, Lyle F. and Palmer, Adrian S. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bowen, David and Bowen, Margareta. 1989. "Aptitude for Interpreting." In *The Theoretical and Practical Aspects of Teaching Conference Interpretation*, Larua Gran and John Dodds (eds), 109–125. Campanotto: Udine.

Boyle, James and Fisher, Stephen. 2007. *Educational Testing: A Competence-Based Approach*. Oxford: BPS Blackwell.

Carroll, John B. 1962. "The Prediction of Success in Intensive Foreign Language Training." In *Training Research and Education*, Robert Glaser (ed), 87–136. University of Pittsburgh Press.

Clifford, Andrew. 2005. "Putting the Exam to the Test: Psychometric Validation and Interpreter Certification." *Interpreting* 7 (1): 97–131.

Gerver, David, Longley, Patricia E., Long, John and Lambert, Sylvie. 1989. "Selection Tests for Trainee Conference Interpreters." *Meta* 34 (4): 724–735.

Gringiani, Angela. 1990. "Reliability of Aptitude Testing: A Preliminary Study." In *Aspects of Applied and Experimental Research on Conference Interpretation*, Laura Gran and Christopher Taylor (eds), 42–53. Campanotto: Udine.

Lambert, Sylvie. 1991. "Aptitude Testing for Simultaneous Interpretation at the University of Ottawa." *Meta* 36 (4): 586–594.

Lambert, Sylvie. 1992. "Shadowing." *The Interpreters' Newsletter* 4: 15–24.

Longley, Patricia E. 1989. "The Use of Aptitude Testing in the Selection of Students for Conference Interpretation Training." In *The Theoretical and Practical Aspects of Teaching Conference Interpretation*, Laura Gran and John Dodds (eds), 105–108. Campanotto: Udine.

Mackintosh, Jennifer. 1999. "Interpreters Are Made Not Born." *Interpreting* 4 (1): 67–80.

Moser-Mercer, Barbara. 1985. "Screening Potential Interpreters." *Meta* 30 (1): 97–100.

Moser-Mercer, Barbara. 1994. "Aptitude Testing for Conference Interpreting: Why, When and How." In *Bridging the Gap: Empirical Research in Simultaneous Interpretation*, Sylvie Lambert and Barbara Moser-Mercer (eds), 57–68. Amsterdam/Philadelphia: John Benjamins.

Padilla, Francisca, Bajo, María T. and Macizo, Pedro. 2005. "Articulatory Suppression in Language Interpretation: Working Memory Capacity, Dual Tasking and Word Knowledge." Bilingualism: Language and Cognition 8 (3): 207–219.

Parry, Thomas S. and Stansfield, Charles W. (eds). 1990. *Language Aptitude Reconsidered.* Englewood Cliffs, New Jersey: Prentice Hall Regents.

Pimsleur, Paul, Sundland, Donald M. and McIntyre, Ruth D. (1963). *Under-Achievement in Foreign Language Learning*. Final Report, Ohio State University, Research Foundation.

Pippa, Salvador and Russo, Mariachiara. 2002. "Aptitude for Conference Interpreting: A Proposal for a Testing Methodology Based On Paraphrase." In *Interpreting in the 21st Century: Challenges and Opportunities*, Giuliana Garzone and Maurizio Viezzi (eds), 245–256. Amsterdam/Philadelphia: John Benjamins.

Rejšková, Jana. 1999. "Establishing A Correlation Between Performance in Consecutive Interpreting and Potentially Good Performance in Simultaneous Interpreting." *Folia Translatologica* 6: 41–60.

Russo, Chiara. 1989. "Text Processing Strategies: A Hypothesis To Assess Students' Aptitude for Simultaneous Interpreting." *The Interpreters' Newsletter* 2: 57–64.

Russo, Mariachiara. 1993. "Testing Aptitude for Simultaneous Interpretation: Evaluation of the First Trial and Preliminary Results." *The Interpreters' Newsletter* 5: 68–71.

Russo, Mariachiara and Pippa, Salvador. 2004. "Aptitude to Interpreting: Preliminary Results of a Testing Methodology Based on Paraphrase." *Meta* 49 (2): 409–432.

Sawyer, David B. 2004. *Fundamental Aspects of Interpreter Education*. Amsterdam/Philadelphia: John Benjamins.

Setton, Robin. 1994. "Experiments in the Application of Discourse Studies to Interpreter Training." In *Teaching Translation and Interpreting 2*, Cay Dollerup and Annette Lindegaard (eds), 183–198. Amsterdam/Philadelphia: John Benjamins.

Shaw, Sherry, Timarová, Šárka and Salaets, Heidi. 2008. "Measurement of Cognitive and Personality Traits in Determining Aptitude of Spoken and Signed Language Interpreting Students." In *Proceedings of the 17th National Convention of the Conference of Interpreter Trainers: Putting the Pieces Together: A Collaborative Approach to Educational Excellence,* L. Roberson and S. Shaw (eds), 91–109. CIT.

Sparks, Richard and Ganshow, Leonore. 2001. "Aptitude for Learning a Foreign Language." *Annual Review of Applied Linguistics* 21: 90–111.

Taylor, Christopher. 1997. "Course Profile: Degree in Conference Interpreting / Translation." *The Translator* 3 (2): 247–260.

Timarová, Šárka and Ungoed-Thomas, Harry. 2008. "Admissions Testing for Interpreting Courses." *The Interpreter and Translator Trainer* 2 (1): 29–46.

Wechsler, David. 1945. "A Standardized Memory Scale for Clinical Use." *Journal of Psychology* 19: 87–95.

Getting it right from the start

Program admission testing of signed language interpreters

Karen Bontempo and Jemina Napier
Macquarie University

This chapter presents data from two related studies concerning signed language interpreter education in Australia. In the first study, 110 signed language interpreters were surveyed on their perceptions of the efficacy of interpreter education programs in Australia in preparing graduates for work as an interpreter. The second study was designed by drawing on the qualitative survey findings of the first study, coupled with previously published results from the survey (Bontempo & Napier 2007), which identified the skills gaps of interpreters that need to be addressed in interpreter education programs. To this end, a program admission test was designed to include six elements considered potentially predictive of performance, and was piloted with a cohort of applicants to a signed language interpreter education program in Australia. Eleven out of 18 screened students were accepted into the program. The exit outcomes showed however that only 55% of the students successfully completed the program; thus the screening test results were not predictive of student performance. We present discussion of the relationship between admission testing and achievement in signed language interpreter education, and make recommendations for researchers and interpreter educators.

Getting it right from the start: Program admission testing of signed language interpreters in Australia

Two research areas of applied linguistics that heavily overlap in terms of common issues, approaches and research questions are language testing and second language acquisition (Bachman & Cohen 1998). Both of these research areas are also relevant to translators and interpreters, as bilingual professionals are required to undertake various language tests to provide evidence of proficiency in their working languages. Many translators and interpreters are also tested on their practical translation or interpreting skills, but there is a dearth of research in this area. Existing research focuses on testing spoken language translators and interpreters, although this is still an under-researched topic (see Angelelli 2007; Clifford 2005; Colina 2008; Hale & Campbell 2002; Kozaki 2004; Lauscher 2000; Mortensen 2001; Niska 2005; Slatyer & Carmichael 2005; Slatyer, Elder, Hargreaves, & Luo 2006; Stansfield & Hewitt 2005; Stansfield, Scott, & Kenyon 1992).

The concept of testing can also be applied to screening applicants for admission into interpreter education programs. Such admission tests may be based on an individual's aptitude; that is, assessing a person's capacity for interpreting, or learning the art of interpreting, with a view to predicting their general suitability for the occupation. More commonly, however, program admission tests evaluate what candidates can currently demonstrate, by testing existing sub-sets of skills, knowledge and abilities required for the task of interpreting. Testing for aptitude is different from testing for existing ability, and some program admission procedures may incorporate a mix of both aptitude and ability tests.

Program admission testing

The practice of program admission testing is pervasive; with the selection of suitable candidates for interpreter education courses naturally a major concern of interpreter educators. The literature available on this type of testing mostly relates to spoken language interpreter education programs, and in particular, conference interpreting program admission testing (Bernstein & Barbier 2000; Gerver, Longley, Long, & Lambert 1984, 1989; Goff-Kfouri 2004; Lambert 1991; Moser-Mercer 1985; Sawyer 2004).

Defining the knowledge, skills and abilities relevant to the complex task of interpreting and distilling some of these down into discrete measurable components that can be reliably assessed at program entry appears to have been a nebulous process to date. Campbell and Hale (2003) note that in spite of considerable developments in language testing in general (with regard to spoken languages) and increased understanding of second language acquisition, little of this knowledge appears to have been used by interpreter educators for the purposes of test design for interpreter program entry. Additionally, many interpreter educators do not have a background in educational measurement.

As a result, it seems that the process of admission testing has been very "hit and miss" thus far. Indeed, questions have been raised about the effectiveness of admission testing for interpreter education programs due to the subjective nature of many admission tests, and the lack of predictive power of such tests, despite their common use in the field (Gerver et al. 1989; Dodds 1990). For example, Timarova and Ungoed-Thomas (2008), surveyed the admission tests of 18 different spoken language interpreter education programs, mostly based in Europe, and found that admission testing was a poor predictor of performance, with 44% of admitted students across the 18 institutions failing to successfully complete their program.

Sawyer (2004) also expresses concern about the weak predictive validity of program admission testing and the lack of scientific evidence supporting their use. He argues that educators should not be describing entry level tests as "aptitude tests" when predictive validity has not been demonstrated, and that most program admission tests are in fact of a diagnostic nature, testing existing abilities rather than assessments of aptitude as such. This diagnostic testing can determine "readiness" for interpreter training by diagnosing current skill level, and in particular any skills deficits (for example, identifying whether greater proficiency in working languages is needed before course commencement) but cannot determine probability of success in an interpreter education program. Sawyer also makes a cautionary comment about the impact of program duration on admission standards – the shorter the program, the higher the entry level standard required, therefore the more rigorous the diagnostic testing admission process needed. In addition, Sawyer notes that, by necessity, program admission testing will vary from institution to institution depending on what entrance level skills are needed by that program, in light of duration, content, emphasis, resources etc.

There is acknowledgement that there is no absolute guarantee or accurate predictor of interpreting performance (Lambert 1991); however, in spoken language interpreter studies, the links between cognitive/affective factors and interpreting skills are considered to be extremely strong (Brisau et al. 1994) and cognitive and affective factors are known to impact on second language learning achievement (Onwuegbuzie et al. 2000). Some studies suggest program admission selection instruments appear to be effective in discouraging or rejecting candidates with little or no chance of succeeding as practitioners (Lambert 1991), and student results on selection tests correlated significantly with performance on final interpreting examinations in the spoken language interpreting field (Moser-Mercer 1985; Gerver et al. 1989).

In regard to selection instruments, there does appear to be commonalities across programs in regard to admission test content (Campbell & Hale 2003). Most institutions seem to agree that an interview is a vital component (conducted in the "B" language, addressing language proficiency, general knowledge, etc.). The admission test also often consists of a selection of the following exercises: shadowing, paraphrasing/summarizing, memory/recall, a translation exercise of some kind (written or sight translation), cloze tests, an essay and a dual processing task of some type (Moser-Mercer 1985; Gerver et al. 1989; Lambert 1991; Russo & Pippa 2004; Pippa & Russo 2002; Sawyer 2004). In addition, many programs include a consecutive interpreting task in their admission testing procedure (Timarova & Ungoed-Thomas 2008).

It seems most admission tests for spoken language interpreters are not developed based on evidence-based research, are not standardized, are subjectively graded, and are typically designed based on the intuition and experience of interpreter educators in individual programs (Campbell & Hale 2003). The reliability and validity of such tests are questionable, and nearly 20 years ago Gerver et al (1989) and Dodds (1990) strongly called for further research on the issues of interpreter aptitude and objective interpreter testing.

There has never been any empirical research conducted on the efficacy of tests for signed language interpreters in Australia: this is true both for tests for certification/qualification and tests for admission into signed language interpreter education programs. Without reliable data available, an understanding of what factors might be predictive of performance remains unknown at present. Therefore the results of current measures used in program admission tests should perhaps be interpreted with caution, despite the strong inclination of the field to apply entrance testing to program applicants and to accept or reject students exclusively on the basis of these results.

Despite a lack of empirical evidence demonstrating correlation between program admission testing and performance during training and in the profession, the interpreting education field remains convinced of the merits of screening. Admittedly, it is a logical position to take, and occupational screening for suitability occurs in other professions, particularly those where the psychological demands of the position are quite high. This is true of interpreting, where the management of stressful conditions and cognitive load are paramount to effective interpreter performance (Kurz 2001).

Borum, Super, and Rund (2003) note pre-employment screening and "fitness for duty" evaluations are commonplace for workers dealing in high risk jobs, and report that psychological profiling of applicants for courses of study or jobs in stressful occupations (such as law enforcement, airline pilot, air traffic controller) is widespread. Of particular interest is the specific reference by Borum et al. to occupations such as a pilot or air traffic controller – a study by Moser (1985, as cited in Kurz 2003) found 18% of interpreter respondents likened their job to a pilot or air traffic controller due to the constant stress and level of concentration required in performing their duties.

In line with this thinking, Humphrey (1994) asserts effective screening strategies at program entry can assist in predicting the successful performance of signed language interpreting students. She provides significant detail regarding the nature and format of screening tools used and the duration of the testing period for entry to one signed language interpreter program in Canada. Unfortunately however, no data at all is provided regarding number of participants, over how many years entrance screening had been conducted by the institution, program duration, content, qualifications and experience of educators in the program, etc. and just a brief reference is made to an overall 98.5% graduation rate at the end of the program.

Roberts (1994) and Monikowski (1994) also support more appropriate admission testing and selection of signed language interpreting students. They suggest that by establishing and implementing a standard of skills, knowledge and abilities required at program entry and testing for these at the time of intake, that these standards can be correlated with end-of-program competencies. They further argue such an approach may result in better outcomes, both during the program of study and in future practice, and may reduce attrition rates in programs, and later in the profession.

In signed language interpreter education programs in several countries there are pre-requisites for entry, suggestions for program screening, and program content sequencing initiatives, albeit not based on empirical evidence (For example in Canada – Humphrey 1994; USA – Monikowski 1994; Finton 1998; Solow 1998; Patrie 2000; Shaw, Collins & Metzger 2006; and Finland – Nisula & Manunen 2009). Much of this literature claims more stringent admission criteria, screening processes and appropriate program sequencing will result in better student outcomes.

Screening for aptitude or ability?

Key issues remain, however, particularly in regard to screening for the seemingly obscure concept of "interpreter aptitude" at program entry. While there appears to be general agreement about some of the skills needed in a candidate that may be assessable by an ability test at program admission (such as knowledge of working languages), less agreement and substantially less research supports factors of aptitude that may be predictive of interpreter performance. Which personality/affective factors (such as anxiety, motivation, stress-resistance, emotional sensitivity, and confidence, among others) and cognitive abilities (for example, intelligence, memory capacity, processing speed, attention span etc.) are predictive of individual performance on an interpreter's course? Of these, which are inherent and cannot be taught, and which can be acquired (or learned to be controlled/enhanced) during a program of study and on the job? How exactly can aptitude for learning the complex skills required in interpreting be assessed in an efficient and effective manner at the time of program entry screening for signed language interpreters? What screening tools can be developed to measure the personal traits, social capital, and cognitive abilities that may suggest candidates possess aptitude for successful completion of an interpreter education program? Is testing for aptitude relevant, or is the more prevalent current approach using ability tests meeting our needs sufficiently well?

With available data and research mostly concentrating on tests of ability, and demonstrating less than convincing links between program admission test outcomes and end of program examinations, it would appear the current ability tests are not meeting our needs sufficiently well. An increasing body of research points to the importance of aptitude, in addition to ability, for potential interpreters.

Organizational psychology literature confirms that occupational performance can not only be improved through the development of competencies via training, practice and experience, but it is also significantly influenced by talent, temperament, "person-vocation fit" and motivation (Maurer, Wrenn, Pierce, Tross & Collins 2003; Losier & Vallerand 1994). Personal interests, as well as cognitive ability, have considerable influence on career choice and successful performance in one's chosen career (Ree, Earles & Teachout 1994; Reeve & Heggestad 2004). Some personality factors are predictive of job performance (Bozionelos 2004; Button, Mathieu & Zajac 1996; Choi, Fuqua & Griffin 2001; Oakes, Ferris, Martocchio, Buckley & Broach 2001), but overall, general mental ability is the single best predictor of occupational performance (Schmidt & Hunter 1998). For these reasons, suitable tests of aptitude, and not just ability, have a place in screening interpreters for program admission.

Given the apparent relationship between cognitive and affective factors and their impact on occupational performance, developing a profile of the skills, knowledge, and abilities needed by a competent interpreter; as well as the personal characteristics and traits needed by a prospective signed language interpreting student would prove very useful (Bontempo 2008; Lopez Gomez et al. 2007; Shaw & Hughes 2006; Stauffer & Shaw 2006). An increased understanding of what kind of foundation is needed in an interpreting student right from the start in terms of both aptitude and ability (and what can be built into training courses to account for any gaps in skills, knowledge, abilities and traits within a student cohort) may help increase the depth and speed of skill acquisition and improvement in performance required of students in interpreter education programs. It may also assist in mitigating the "readiness to work" gap identified in American signed language interpreting graduates by Anderson and Stauffer (1990) and Patrie (1994), and similarly found in Australia (Bontempo & Napier 2007).

Some pioneering studies which attempted to profile the psychological make up of the signed language interpreter point to the potential role of personality in successful occupational performance as an interpreter but the breadth of the research conducted was limited, and based on small samples of interpreting practitioners (see Rudser & Strong 1986; Doerfert & Wilcox 1986; Schein 1974: cited in Frishberg 1986; and Frishberg & Enders 1974, as cited in Frishberg 1986).

However, more recent studies including slightly larger samples of participants and a wider range of cognitive and personality measurements have begun to identify some common themes of interest. Bontempo (2008), Lopez Gomez et al. (2007), Seal (2004), and Shaw and Hughes (2006) found that having certain cognitive abilities, aptitudes and personality traits are significant predictors of performance in the signed language interpreting profession. While it is recognised that temperament and other psychological characteristics influence performance across a range of occupations, these considerations are often not properly considered when screening and selecting trainee interpreters. Understanding the role that disposition and traits may have in influencing competence, and therefore performance, as an interpreter may be relevant, in addition to the capacity of an individual to handle the inherent technical skills required, i.e. the linguistic and cognitive processing aspects of the job.

Using existing psychometric tests to measure the dispositional traits of 110 signed language interpreters in Australia, Bontempo (2008) found a strong correlation between high levels of negative affect (neurotic, anxious, emotionally reactive / easily distressed) and lower reported levels of competence in interpreters. In addition, a positive correlation between self-efficacy (belief in personal ability to succeed and accomplish tasks / self-confidence) and higher levels of interpreter competence was reported.

Lopez Gomez et al. (2007) administered a battery of tests to 28 signed language interpreting students in Spain, examining perceptual-motor coordination; cognitive skills; personality factors; and academic background, and comparing results with an expert trainers' evaluation of the students' proficiency in sign language and interpreting. Perceptual-motor coordination was found to be the most significant predictor of proficiency in a signed language, while cognitive and personality factors were considered influential in developing signed language interpreting abilities. In particular, the personality factor of dominance was found to be of interest – high scores on this factor indicated a person was assertive, resourceful, confident, task-oriented, responsible, stress-resistant etc. Low scores on this factor point to low self-confidence, rigidity in problem solving, unreliability and so on. Lopez Gomez et al. found this trait of dominance of relevance to success in the achievement of signed language interpreting abilities. This latter finding is supportive of Rudser and Strong's (1986) earlier work. In addition, cognitive abilities such as abstract reasoning and memory skills were found to be significant, supporting the findings of Seal (2004).

Shaw and Hughes (2006) identified characteristics such as self-motivation and confidence as defining features, as well as the ability to multi-task and to process information rapidly, in their study of over 1000 signed language interpreter education program participants in North America and Europe. They note that although personal qualities and traits may seem to influence success in the profession, given that the predictive abilities of these characteristics are as yet not fully defined or validated, it might be premature to apply them in a testing process as admission screening tools. Shaw and Hughes argue it would be more effective to incorporate trait awareness and skill building on elements such as assertiveness and so on into interpreter education course curricula, given "student's ability to learn, develop and enhance critical personal and cognitive characteristics" (2006: 218).

Shaw, Grbic and Franklin (2004) explored and compared the perceptions of spoken and signed language interpreting students about factors that contribute to, or inhibit, readiness to apply language skills to interpreting performance. They found that students experience a period of transition after entering an interpreter education program as they realize the task of interpreting is more complex than simply being able to use two languages fluently. They specifically identify confidence and risk-taking as primary personality assets that contribute to successful adaptation through this period of transition, and resulting performance by students on interpreter training courses. Similar findings regarding the personality of learners were found by Onwuegbuzie, Bailey and Daley (2000) in their study of foreign language learning students. Additionally, Riccardi, Marinuzzi and Zecchin (1998: 98) note that accomplished performance as an interpreter may not be possible unless skill sets "are matched by specific personality elements."

Brisau, Godijns and Meuleman (1994) attempted to develop a psycholinguistic profile of the interpreter by forcing a distinction between the linguistic and non-linguistic parameters that determine interpreting performance. The linguistic parameters included vocabulary, syntax, listening comprehension and delivery. The non-linguistic parameters were psycho-affective factors including self-concept, cognitive style, real-world knowledge, anxiety, attitude, stress resistance and meta-cognition. Additionally, neurolinguistic factors rate a mention, with attention and memory stressed as indispensable factors for interpreters.

Schweda Nicholson (2005: 28) in her application of the Myer-Briggs Type Inventory (MBTI) to 68 spoken language interpreter students ultimately found "the profession may offer opportunities for all personality types to exercise their preferred ways of interacting, deciding and being." Although participants were represented across all personality types in the MBTI, the largest group represented was the "Introverted, Sensing, Thinking, Judging" Type (or ISTJ), accounting for 18% of her interpreting sample (which was 75% female), but is a type normally represented by only 7% of the wider population (and further, only one third of this 7% are women). In particular, the thinking/feeling dimension of the MBTI results showed an extremely significant number of "thinking" types – a dimension associated with impersonal and logical analysis of ideas and information, being thorough, attending to detail, organizing and synthesizing information, setting high standards of achievement, and coping with stress.

Other factors influencing student achievement

If we draw together the above mentioned research, it would appear that the successful interpreting student (and practitioner) is a "package" of skills, knowledge, abilities, experiences (academic and vocational experience, as well as life experience), personal characteristics, traits and attributes. However, the outcome for students in an education program is not just dependent on individual difference, but is also heavily influenced by the interaction between an individual students' aptitude, learning style, motivation, willingness and opportunity to practice and perform, and the pedagogical approach and competence of their teacher/s (Robinson 2002; Moser-Mercer 2008a).

In a very large scale meta-analysis of research studies across all age ranges and education settings to determine what most influences student achievement in an academic setting, Hattie (2003) found 50% of the variance in academic achievement was accounted for by the student, and notes, "it is what students bring to the table that predicts achievement more than any other variable. The correlation between ability and achievement is high, so it is no surprise that bright students have steeper trajectories of learning than less bright students" (Hattie 2003: 1). Teachers accounted for the next greatest single amount of variance in achievement, directly responsible for approximately 30% of the learning that takes place, with Hattie noting what teachers "know, do, and care about" has a powerful impact on students. The combined effects of resources and facilities, family, influence of peers, institutional culture etc, altogether accounted for the remaining 20% of variance in achievement.

Thus, dynamics such as the interaction between peers in the classroom, access to resources and necessary materials, family support, the institution, and so on, will likely all also have an impact on student outcomes. The impact will be to a significantly lesser extent than the previously mentioned critical student aptitude/teacher effectiveness factors, however.

Program admission testing in Australia

In terms of getting it right from the start, formal admission testing of signed language interpreters accessing entry level courses in Australia should logically form an integral part of the process of attempting to effectively select students who are program-ready. However, there is currently no uniform approach to the screening of potential Auslan (Australian Sign Language) / English interpreters in Australia.[1] Individual programs and educators implement such measures on an ad hoc and informal basis.

With little other than subjective opinion and years of experience shaping decisions, interpreter educators in Australia are presently conducting screening interviews and accepting or rejecting students for interpreter education program entry on the basis of variable admission testing procedures that lack clarity about testing both language proficiency in Auslan, and interpreter course readiness. This is not to suggest that the capacity of a grader to make an intuitive assessment based on years of experience is no longer of any merit whatsoever, more so that the introduction of systematic formal screening procedures would enhance processes by adding standardization to current intuitive approaches. Presently different colleges throughout the nation use varying admission procedures, improvised on the experience of the incumbent coordinator, and based on the sketchy requirements outlined in the national interpreter education curriculum (personal communications, D. Goswell, P. Bonser and M. Bartlett 2006, 2008).

The existing admission procedures in the four colleges that regularly conduct annual interpreter education programs for Auslan interpreters in Australia (based in Melbourne, Victoria; Sydney, New South Wales; Brisbane, Queensland; and Perth, Western Australia) included a mix of the following: an informal interview, an English essay, written responses to English questions, sight translation of an English text into Auslan, consecutive interpreting task, English grammar test, English terms/concepts definitions, comprehension tasks in Auslan, questions regarding motivation for applying for the course, and questions about Deaf culture, the Deaf community and/or signed language interpreting. Screening is often conducted by only one or two people (sometimes separately, to screen applicants more quickly) but always includes a native signer (although not always a deaf person – the native signer may be a hearing person with Auslan as their first language due to having deaf parents), and time constraints and funding limitations prevent an in-depth admission procedure. The challenge of the process is exacerbated by the absence of a reliable and valid language assessment tool that can effectively measure proficiency in Auslan. This lack of standardized entry testing and informal approach to admitting students into signed language interpreter education programs has been acknowledged in other countries (Lopez Gomez et al. 2007).

Admission testing remains a necessary component of interpreter education to maximize the recruitment and retention of suitable applicants, however. Program

[1] Hereafter referred to as Auslan interpreters.

coordinators and educators need to be able to distinguish the qualities and skills of prospective students, and to predict 'interpreter-training-potential', in order to select a suitable cohort of students for program commencement. Without some form of screening or assessment pre-entry, class sizes may become unwieldy for interpreter education purposes.

Given that interpreter education programs are practice-oriented, students with less than adequate existing skills (for example, in language proficiency) will impact on class dynamics and group progress. As a result, interpreter educators may not be able to focus their energies appropriately to maximize student outcomes and quality standards in an interpreter education program. At present, time and money is expended in training many people in interpreter education programs who do not pass the final examination, or never work in the field, or only practice for a short time. Poor performance on a program or in the profession is disheartening to the individual, and is difficult for educators and employers to manage, not to mention the potentially grievous impact on service users. Admission testing is therefore expected, if not demanded, by many stakeholders, such as teachers, employers, and the Deaf community, as well as students themselves.

Anecdotal evidence from interpreter educators, employers, and the Deaf community in Australia suggests that some candidates currently enter interpreter education programs without the aptitude or the pre-requisite skills, knowledge, and abilities for effective program participation, and correspondingly exit without the level of competence required to function adequately in the profession. At present the system does not seem to be meeting the performance expectations or the needs of the Deaf community (Orima 2004; Napier & Rohan 2007), with only a relatively small number of practitioners nation-wide meeting the growing demand for competent practice in Australia (Bontempo & Napier 2007).

The context for interpreting and interpreter education in Australia

In Australia, signed language interpreter education programs have existed in educational institutions for more than 20 years (Bontempo & Levitzke-Gray 2009). Interpreters can become certified by either undertaking an approved education program, or sitting for an interpreting examination without attending any course of study. Signed and spoken language interpreters are both accredited according to the same standards as determined by NAATI (the National Accreditation Authority for Translators and Interpreters).

NAATI accreditation is the only officially recognized certification for interpreting throughout the nation. Accreditation for signed language interpreters is currently available at "Paraprofessional" or "Professional Interpreter" level.[2] Accreditation suggests practitioners have met the minimum standards required to competently perform interpreting related tasks associated with the level of accreditation at the time of being examined.[3]

The minimum standard for professional practice in Australia is NAATI accreditation at the Professional level. The Paraprofessional level of accreditation is supposed to be limited to practice by interpreters of new and emerging languages in Australia; is for conversational interpreting purposes only; and is considered a stepping stone towards Professional Interpreter level of practice (NAATI 2007).

Regardless of this stepping stone structure, the vast majority of the Auslan interpreter population is accredited as Paraprofessionals, partly because of the lack of interpreter education programs which lead to Professional Interpreter level accreditation in Australia. Although NAATI provides descriptors of the nature of work expected at each level of accreditation, many Paraprofessionals are known to practice in settings that would normally be considered the domain of (Auslan) Professional Interpreter level practitioners, such as higher education, court, and conferences (Ozolins and Bridge 1999; Bontempo and Napier 2007).

As of September 30, 2008, NAATI had accredited a total of 888 Auslan interpreters since testing commenced in November 1982. These include 768 interpreters accredited at the Paraprofessional level; and only 120 practitioners accredited at the Professional Interpreter level.[4] Australia, not unlike many other countries, faces a challenge whereby the demand for competent interpreters greatly outstrips

[2] NAATI defines Paraprofessional Interpreter level as "a level of competence in interpreting for the purpose of general conversations. Paraprofessional Interpreters generally undertake the interpretation of non-specialist dialogues. Practitioners at this level are encouraged to proceed to the professional levels of accreditation". NAATI Professional Interpreter level is defined as "the first professional level and represents the minimum level of competence for professional interpreting. Interpreters convey the full meaning of the information from the source language into the target language in the appropriate style and register. Interpreters at this level are capable of interpreting across a wide range of subjects involving dialogues at specialist consultations. They are also capable of interpreting presentations by the consecutive mode. Their specializations may include banking, law, health, and social and community services" (NAATI 2007). Work is currently underway to develop another level of accreditation for signed language interpreters – "Conference Interpreter". This level presently exists for spoken language interpreters but has not been available to Auslan interpreters to date.

[3] For an overview of accreditation standards of signed language interpreters in Australia as compared to the USA and UK, see Napier, J. (2004). Sign Language Interpreter Training, Testing & Accreditation: An International Comparison. American Annals of the Deaf 149(4): 350–359.

[4] G. Lees, personal communication, September 30, 2008.

available supply (Orima 2004). Paraprofessionals therefore have little difficulty obtaining employment in the current tight labor market, meaning there is little incentive to upgrade to the higher level of accreditation.

This raises questions about Paraprofessional interpreters' capacity to perform the work often allocated to them due to market demand. The disparity between the level of accreditation and skill of Paraprofessionals and Professional Interpreters was examined in a research study by Bontempo and Napier (2007). Survey respondents were asked to rank the level of importance of a range of skills, knowledge, and abilities for signed language interpreters; then to rate their own level of competence for each of these skills, knowledge and abilities. The results provided evidence of a clear skills gap where an interpreter rated a particular skill as being very important for an interpreter, but rated their own level of competence lower on the same skill set. Given the expectation that Professional level interpreters should have more sophisticated linguistic and interpreting skills, it is not surprising that these respondents demonstrated fewer skills gaps, and higher levels of competence overall, in comparison to Paraprofessionals. Paraprofessionals, in contrast, self-identified that they lacked a number of skills that they had ranked as vitally important for interpreting. This finding is of concern.

Technical and Further Education (TAFE) colleges conduct language acquisition programs for Auslan students and Paraprofessional level interpreter education programs on an annual basis in the larger capital cities in Australia. TAFE community colleges are vocational education and training institutions, delivering courses with a trade and skills based focus, typically with an emphasis on practical skill development suited to the relevant industry. This includes courses that are apprenticeship or traineeship based (hairdressing, plumbing and carpentry for example), as well as Certificate and Diploma qualifications in a diverse range of careers, stretching across fashion, photography, child care, music, real estate, languages and tourism to name just a few fields of study. Academic qualifications issued by TAFE are "pre-degree" level qualifications, although some TAFE awards can result in exemption of selected first year university units via recognition of prior learning. TAFE courses are shorter than university degrees, tend to blend theory and practice in a vocational context, and are a fraction of the cost of university studies.

The Paraprofessional interpreter programs are provided as a vocational education program at TAFE and successful completion results in a "Diploma of Interpreting" (which leads to NAATI Paraprofessional level accreditation). The interpreter education programs are conducted over one year part-time (approx. 8 hours per week), after completing the requisite language acquisition programs at a training institution like TAFE,[5] or by obtaining linguistic fluency via another avenue. This entry pathway is available to spoken language interpreters also, and the same Diploma of Interpreting curriculum applies to both spoken and signed language interpreting students at TAFE. Spoken language interpreters in Australia however are able to access interpreter education programs at *either* TAFE or at university in most states, where undergraduate and postgraduate interpreting degrees are available in several spoken languages.

Macquarie University in Sydney is currently the only university in Australia offering a degree in Auslan interpreting; however the program is at postgraduate level for experienced NAATI accredited Paraprofessional interpreters to advance their skills and to gain Professional Interpreter accreditation upon successful course completion. At the time of writing there is no university program at undergraduate level in Australia geared towards entry level Auslan interpreters. The current postgraduate program in Auslan/English interpreting at Macquarie University is unique; therefore reference in this chapter to interpreter education programs in Australia will for the most part be in regard to TAFE colleges nationwide, with a specific emphasis on vocational Diploma level training for Paraprofessional level interpreters.

Programs of study at TAFE colleges in Australia are based on a national competency based curriculum, and therefore contain a degree of consistency in regard to learning outcomes on paper. Nonetheless, there is variation "on the ground' in terms of operational factors and logistics, such as admission testing; course delivery; sequencing of skills development stages; qualifications and quality of teaching personnel; availability of suitable resources and equipment, etc.

The time limitations of current TAFE interpreter education programs prevent educators from being able to allocate time and resources to those students who do not meet a certain level of competence in various domains at the time of program entry. Those ill-equipped to meet program demands are less likely to reach exit level competence; and if they do scrape through an end of year examination, they may struggle to perform adequately in the profession. The relatively high attrition rates observed in programs and in the field may be partly a result of poor admission screening to begin with. People who lack the confidence and skills to remain in the profession appear to either withdraw from the profession, or are actively excluded from practice by either the Deaf community or by service providers, based on feedback on performance.

[5] Completion of the Diploma of Auslan (language acquisition studies) does not guarantee entrance to the Diploma of Interpreting. Language proficiency may not be at the standard required for interpreter education program entry.

These issues have a bearing on the ability to provide quality interpreting services to the Australian Deaf community and other service users. Potential solutions to this pressing concern in Australia are threefold – to examine admission testing and the outcomes of entry level interpreter education programs; to increase the skills level and capacity of qualified Paraprofessionals to that expected at the Professional Interpreter level of accreditation; and to appropriately target ongoing training opportunities for Professional Interpreter level practitioners to minimize the risk of their advanced skills becoming fossilized. The focus of this chapter is the first concern – admission testing and the outcomes of entry level interpreter education programs.

As already mentioned, there are no research studies available that describe the methods by which signed language interpreters are currently screened, and selected into interpreter education programs in Australia. There are no clearly articulated national protocols regarding program entry, no coordinated databanks providing clear directions and information on assessment during interpreter education programs or on the final test outcomes of such programs. In addition there is no transparent and easily accessible information on the full range of accreditation outcomes of direct NAATI testing on Auslan interpreters, either by testing alone, or accreditation via approved programs of study. Given the importance of effective pedagogical assessment for evaluating student progress, and accreditation assessment for determining standards of performance upon completion of an education program, the paucity of research and scholarly contributions on interpreting assessment, particularly in Australia, is surprising. Nothing at all appears to exist regarding measuring interpreting aptitude in Australia, and very little exists on this subject in the signed language interpreting field internationally. This exploratory study surveying practicing interpreters regarding their perceptions of interpreter education; the development and administration of an assessment tool in the form of an admissions test; and comparison with final examination results, is therefore timely and much needed in Australia.

Study procedure

In order to investigate factors that may be predictive of Auslan interpreter competence, as well as perceptions of the efficacy of interpreter education programs, a detailed survey was administered to accredited Auslan interpreters throughout Australia. The lengthy survey was designed to determine the incidence, distribution, and interrelations among sociological and psychological variables: that is, to examine demographic details and personality test results in conjunction with individual ratings of perceived linguistic skill; other interpreting-related knowledge, skills and abilities; self-ratings of overall competence and perceptions of the efficacy of interpreter education programs (Bontempo 2005). Feedback from the survey provided the impetus to more closely examine interpreter education programs, and in particular the process of screening applicants to interpreter education programs in Australia, with a view to improving standards in programs, and upon exit from programs and entry into the profession.

Research questions

The research questions for this study were as follows:

1. Are signed language interpreter education programs in Australia perceived by practitioners to be preparing interpreters for effective performance in the profession?
2. Can the interpreter education program admission tests commonly referenced in the literature for spoken language interpreters be adapted and applied to signed language interpreters for entry level screening purposes?
3. Are the results of program admission tests developed and administered in this study predictive of final examination performance?

In 2007 Bontempo and Napier conducted a survey that identified significant skills gaps in accredited Auslan interpreters. The present study extends that research by exploring previously unpublished data from the original survey concerning practitioners' perceptions of interpreter education programs in Australia. This chapter will provide an overview of the relevant survey results; and will discuss the subsequent interpreter education program admission test developed. In order to determine whether we are "getting it right from the start", we will provide quantitative and qualitative analyses of the effectiveness of the test by contrasting the program admission scores with exit outcomes, and comments gathered from program participants regarding their perceptions of the admission test.

We acknowledge that there are several variables that may have an impact on the correlation between admission scores and exit outcomes, such as student motivation, learning style, program content and delivery, and the pedagogical approach and competence of the interpreter educators involved in the course. For the purpose of this paper however, we are exploring only the explicit relationship between entrance and exit scores rather than hypothesizing in any detail on other confounding variables that may have impacted on exit scores, and we have not controlled for these variables. We also acknowledge that this discussion can only be considered as a small pilot study, as the admission test was administered with one cohort of signed language interpreting students at one college. Nonetheless the results generate interesting food for thought in relation to interpreter education, testing and accreditation, and lead us to make recommendations regarding further research.

This research has two components: Firstly, a study was undertaken to explore characteristics and parameters of the signed language interpreter population through a survey instrument. The purpose of this initial study was to identify factors that may be potential predictors of successful performance in Auslan interpreters, and to obtain participant views of interpreter education programs in Australia. These results then fed directly into the second part of the study – the development of a screening test to be piloted with applicants to an Auslan interpreter education program. Here we discuss first the methodology and results for study 1, the survey; and then give an overview of the methodology and results for study 2, the development and administration of the program admission test.

Study 1: Survey instrument

Methodology

A mail questionnaire instrument was designed drawing on literature from organizational psychology, interpreting and translation, and applied linguistics. The survey was designed to obtain data to determine the incidence and distribution of, and interrelations among, sociological and psychological variables. We compared respondents' personal facts (such as route to qualification, level of qualification, years of interpreting experience, etc.) with their opinions and attitudes about general linguistic skill, other knowledge and abilities, overall competence and some personality measures. More details regarding the instrument are provided in the "materials' section. Demographic information was obtained from participants in order to develop a profile of the participants and the profession, and also to allow for examination of the relations among these variables as well as the overall interpreting competence rating reported by respondents.[6] In addition, participants were asked their perception of the effectiveness of interpreter education programs for Auslan interpreters.

[6] Findings with regard to interpreter perceptions of competence and reported skills gaps are discussed in detail in Bontempo & Napier (2007).

Participants

NAATI accreditation as an Auslan/English interpreter was an essential criterion for participation in the study to ensure only practitioners who had met benchmarks for work in the field could respond to the study.[7] Survey respondents had passed an interpreting examination in Auslan/English at a prior point in time (either via a NAATI approved course of study or direct NAATI testing) that had deemed them competent to practice at either Paraprofessional or Professional Interpreter level. NAATI accreditation is the only recognized licence to practice as an interpreter in Australia (in both spoken and signed languages), and federal legislation such as the Disability Discrimination Act and state government language policies protect the rights of deaf people in requesting a NAATI accredited Auslan interpreter when accessing services in the wider community.

As only NAATI accredited Auslan interpreters could participate, potential subjects were able to be identified and sourced via a number of avenues which would potentially allow for multiple hits on individuals. Information regarding the study was distributed nationally using direct mailing lists and through snowball sampling. Practitioners were asked to pass on the information about the research study to other practitioners they knew and who may not have received the information via a direct mailing list.

A flyer regarding the study was posted or emailed out to 500 accredited Auslan interpreters on the NAATI mailing list at that time. All accredited Auslan interpreters were eligible and could self-select to participate in the study – no sampling was conducted. Information was also distributed by the main employers of signed language interpreters in Australia at the time, and by the Australian Sign Language Interpreters' Association (ASLIA). Employers and ASLIA would have had access to most of the same population contacted directly by NAATI, with information estimated to have reached approximately 200 working interpreters via employers and approximately 300 members of ASLIA at that time.[8] It is estimated

[7] Unaccredited Auslan interpreters are used infrequently in Australia in community based interpreting settings such as medical, legal or government appointments; however it is not uncommon for unaccredited practitioners in some Australian states to gain employment in education settings, working with deaf students accessing a mainstream education at school, TAFE or university. Unaccredited practitioners were not able to participate in the study, so data collected by the survey would therefore be from participants who had met a certain tested standard of practice in interpreting already.

[8] ASLIA allows student membership, and "inactive" interpreters can also retain membership. Membership of ASLIA is not compulsory for practice in Australia, so ASLIA membership numbers are not directly reflective of the number of accredited and active Auslan interpreters in Australia.

that approximately 500 accredited Auslan interpreters received information about the study via one or more sources.

Interested parties contacted the researchers and were then sent the questionnaire[9] either via email or regular mail according to their preference, along with introductory information, and a stamped addressed envelope if requested. A total of 82 Auslan interpreters requested a copy of the questionnaire. Surveys were not coded in any way, so it is not possible to determine if all of the requesting parties returned a completed questionnaire.

A total of 110 completed questionnaires were received from interpreter respondents via email or mail, which is more than the number of requests received for the survey. It is assumed that some respondents may have passed a copy of the questionnaire onto colleagues, or that some employers forwarded copies of the survey directly to interpreter employees, circumventing the need for potential respondents to contact the researchers directly to obtain a copy of the survey.

Whilst 722 interpreters had been accredited by NAATI between 1982 and the release of the survey in early 2005, a report commissioned by the Federal Government Department of Family and Community Services noted that only 257 accredited interpreters were working in the field at that time (Orima 2004). This was a little more than the figure reported by employers (approximately 200 active interpreters were sent the flyer by employers) and a little less than the national ASLIA membership at the time (approximately 300 members) and therefore appears to be an accurate reflection of the number of working practitioners in early 2005.

Thus an estimated response rate of 42% (110 respondents / 257 estimated working practitioners) was considered more than adequate. This is considered a higher than average return rate in a survey methodology, whereby an average and acceptable return rate is deemed 20–30% (Jackson 2003).

Materials

The survey instrument was a 10 page questionnaire, including a carefully planned construction of questions based on the literature. A total of 22 questions were presented, with questions arranged in related subsets of four main sections – demographic information; skills gap information; perceptions of performance; interpreter education programs/training options; and rating scales, which were drawn from existing psychometric measures used in the field of organisational psychology.

[9] Approved by the ethics committee of Macquarie University and subject to standard requirements for data collection.

In summary, a combination of open ended, close ended, partially open ended, and various Likert rating scales (with 5 alternatives to obtain interval data) were to be completed by participants. The rating scales pertained to overall competency as an interpreter, a detailed skills gap analysis, and various self-reporting personality measures of self-efficacy, positive and negative affectivity, and goal orientation.[10]

The first 10 questions were for the purposes of collecting sociological data on practitioners in order to develop a profile of the profession in Australia. These included closed questions and partially open ended questions regarding work status, accreditation level, year of accreditation, work setting, age group, gender, state or territory of residence, first language, secondary schooling, post-secondary schooling and extent of formal interpreter education completed.

Question 11 listed fifty defined skills and areas of knowledge that may be relevant to signed language interpreters, as drawn from the literature. Participants had to rate the importance of each skill, knowledge or ability, and correspondingly, offer their assessment of their own competence in that particular skill or knowledge domain. This information provided the researchers with a quantifiable skills gap.

Questions 12–17 were open ended questions asking participants to express their thoughts on additional skills, knowledge or abilities of an Auslan interpreter not listed in question 11; the effectiveness of interpreter training; reasoning for decisions in regard to work selection; and perceptions of performance. These questions were designed to provide qualitative data, which could then be cross-referenced with the quantitative data collected.

Respondents were then asked to rate themselves on a scale assessing their overall competence as an interpreter on question 18. This information would provide a key variable to examine in relation to interpreting responses on other sociological and psychological variables.

Questions 19, 20 and 21 were scales with an established history of use in the field of organizational psychology, as established psychometric tools assessing social-cognitive personality constructs such as self-efficacy, positive and negative affectivity, and goal orientation respectively. Finally, at question 22 participants could make some open ended comments and add anything further if they so wished.

The first draft of the questionnaire was piloted with two Paraprofessional interpreters and one Professional Interpreter to obtain feedback regarding the comprehensibility of the material, and suitability of the line of questioning. Following

[10] The results of the latter three psychological self-report measures are reported in Bontempo (2008).

the review and feedback by colleagues, some minor amendments were made to the preliminary version before it was released to participants in the study.

Procedure

Participants in the research study completed the survey instrument after receiving it in the mail, or via email. Questionnaires were estimated to take up to 40 minutes to complete, and respondents completed the survey in English (handwritten or typed responses were possible) at their leisure and in their own chosen environment. Participants had access to information about the study and potential possession of the questionnaire for up to 8 weeks, and posted or emailed their questionnaires back upon completion. On receipt of the completed questionnaires, the figures were analyzed to note any areas of significance, using descriptive, parametric and non-parametric inferential statistical analysis.

Results and discussion

A total of 110 signed language interpreters returned completed questionnaires. No unusable surveys were returned. A total of 67.3% of respondents were accredited at Paraprofessional level and 32.7% at Professional Interpreter level.[11]

Table 1. Level of NAATI accreditation held by respondents

Accreditation level	Respondents (n)	Percent
Paraprofessional	74	67.3
Interpreter	36	32.7
Total	110	100.0

Skills gaps reported by practitioners

Reported in detail in Bontempo and Napier (2007), and only summarized here, are the skills gaps identified by Paraprofessional and Professional Interpreter respondents. Significant gaps in skill were found when Professional Interpreters rated degree of "importance" of certain skills and attributes; and then their own "competence" on the same. The variables where gaps for Professional Interpreters were identified after analysis included the following: self-confidence; memory

[11] See Bontempo and Napier (2007) for detailed demographic information regarding respondents.

skills; concentration skills; self-monitoring skills; specialist knowledge; objectivity; public speaking skills; self-discipline; world knowledge; contextual knowledge; assertiveness and intuition.

On the other hand, significant gaps for Paraprofessionals were found in the following areas: Auslan skills; interpreting/translating skills; contextual knowledge; memory skills; concentration skills; listening skills; self-monitoring skills; self-confidence; world knowledge; reputation; objectivity; spelling skills; situational management skills; specialist knowledge; general intelligence; self-discipline; analytical skills; and assertiveness.

The different ratings ascribed by practitioners depending on their level of accreditation, and the resulting data demonstrate Paraprofessionals identified a greater number of gaps in their skills base across a wider range of skill domains, including fundamental skills for Auslan interpreters, such as language proficiency in Auslan.

Perception of interpreter education programs in Australia

Of primary interest in this paper is the qualitative data obtained from the survey on a question pertaining to interpreter education programs. This question was placed immediately after the section regarding the skills, knowledge and abilities that might be important for interpreters and a rating scale where respondents scored their own competence on each of the 50 variables after indicating the degree of importance of each variable. This position of the open ended questions served to draw the participants' attention to areas where they might feel they need additional training, and to reflect on their own level of competence as a practitioner.

The specific question of interest was: "How well do you think interpreter education programs in Australia prepare interpreters for effective performance in the profession?" Despite this being presented as an open-ended question allowing for a free form response, and placed after an intensive Likert scale regarding skills gaps, a total of 106 out of 110 survey participants elected to provide a written response to this particular question. This is a profound response rate to an open-ended question.

A total of 67% of respondents to this question (25 Professional Interpreters and 46 Paraprofessionals) noted deficits in interpreter education programs, resulting in feeling poorly equipped to function as an interpreter upon course exit, and entry into the profession. Common themes amongst these respondents were:

- course duration (not long enough);
- course content and complexity (not reflective of the real world of work);
- insufficient resources and materials for use on course;

- the varied qualifications and competence of the interpreter educators on courses;
- the lack of mentoring, internships and "buddy systems" available to students and new graduates;
- the inaccessible location of courses (in certain city centres only);
- the infrequency of programs in some states;
- and the apparent lack of a national standard in programs.

The last point is particularly interesting given interpreter educators in all programs throughout Australia are operating from the same national competency based curriculum.

Selected comments representative of the negative perceptions of the interpreter education programs are noted below:

> Currently many paraprofessionals are sent out to work in the "real world" severely under-equipped. I believe this contributes to the very high attrition rate in newly qualified interpreters (Participant #36, Professional Interpreter)
>
> Considering the type of work that paraprofessional Auslan interpreters need to do, the Diploma courses do not really equip them.(Participant #103, Professional Interpreter)
>
> From what I have seen, they are not adequately prepared. Fluency in both languages is essential and personal attributes need to be examined and worked upon to enhance the skills interpreters require (Participant #38, Professional Interpreter)
>
> They don't seem to prepare interpreters to be able to work with language variation; nor do they seem to prepare interpreters in the use of a high standard of English (Participant #83, Professional Interpreter)
>
> Poor in technical interpreting techniques (Participant #82, Paraprofessional)
>
> After exiting interpreter training programs you feel thrown in the deep end (Participant #51, Paraprofessional)
>
> Poor communication between lecturers, does not focus on the more important aspects of interpreting (Participant #109, Paraprofessional)
>
> TAFE courses train people how to sign and how to prepare for their NAATI exam but they don't teach students how to actually *be* an interpreter (Participant #88, Paraprofessional)
>
> I feel here in *(State omitted)* it is ineffective. I feel the standard has dropped – interpreters are not flexible, too rigid and have difficulty bridging between the two cultures. Generally I find some are over-confident and it shows in their attitude and work (Participant #25, Professional Interpreter)

> Entry level programs (e.g. TAFE) currently do not run for long enough to adequately cover even the essentials (Participant #70, Professional Interpreter)
>
> The course all depends on the teachers' own knowledge, and importantly, their ability to impart it to students (Participant #26, Professional Interpreter)
>
> Courses across Australia are widely varied and inconsistent (Participant #35, Paraprofessional)

A minority of survey participants, at 22%, were more positive about the grounding they had received in various interpreter education programs. This group of respondents consisted of 4 Professional Interpreters and 19 Paraprofessionals. Some representative comments from this group of participants are below:

> I believe it did prepare me for effective interpreting and professionalism in the field (Participant #74, Professional Interpreter)
>
> The courses I completed at TAFE and at Macquarie were both excellent. The problem is not all people have access to (or choose to access) these courses (Participant #57, Professional Interpreter)
>
> Interpreter training is thorough – a good foundation of skills (Participant #15, Paraprofessional)
>
> The training program in *(State omitted)* is excellent. It has an excellent teacher who is an Auslan interpreter with a vast amount of knowledge. It contains role plays for specific subject areas which is highly valuable – almost real! Also includes various forms of interpreting, e.g. platform, consecutive, simultaneous… (Participant #42, Paraprofessional)
>
> I found the training program excellent in preparing interpreters for the profession (Participant #89, Paraprofessional)
>
> I found the course to be varied, interesting and challenging. Interpreting skills were addressed systematically, practically and very professionally, with many opportunities given for interactive learning… (Participant #95, Paraprofessional)

The final group of 11% (6 Professional Interpreters and 6 Paraprofessionals) were non-committal in their response indicating they didn't feel they had any insights to offer, or they provided a mixed response that could not be considered a wholly negative or a positive comment on the state of interpreter training in Australia. For example, participant #1 (a Professional Interpreter) noted:

> I think they're great – the problem is (a) they're are not compulsory and (b) paraprofessionals should NOT be doing most of the work they're doing!

Of interest with this latter group, is that with further analysis it was found that 8 out of 12 of these respondents had never participated in an interpreter education program themselves. This may have accounted for their unwillingness to provide a firmer opinion on the efficacy of interpreter education programs. Despite the lack of formal education amongst this group, 11 out of 12 of this group rated themselves as "more than competent" or "extremely competent" as an interpreter. Notably, all except one respondent had more than 10 years of practical experience in the field, however.

Returning to the first group of participants – those expressing concern about the deficits of interpreter education programs in Australia – some respondents took the opportunity to offer suggestions for improvements when responding to the question about interpreter education, including some specific, unprompted, references to program admission standards on TAFE programs, and the need for a entry level degree course at university, rather than entry level programs only being available at TAFE:

> We need either higher standards at entry or more units so students can achieve higher competency in many areas (Participant #105, Paraprofessional)
>
> Pre-requisites for entry need to be improved (Participant #9, Paraprofessional)
>
> Entry level requirements need to be strictly maintained (Participant #60, Professional Interpreter)
>
> Effective preparation is impossible without a full degree program and higher standards (Participant #36, Professional Interpreter)
>
> I don't think a 1 year part time TAFE course is appropriate for this high demand profession. I don't think we can call it a profession in that instance either. (Participant #23, Paraprofessional)
>
> I feel the most pressing issue is lack of baseline university training. (Participant #44, Professional Interpreter)
>
> Ideally, degree courses should be the basic training, but I don't see degree courses able to become essential because of (a) lack of numbers in Australia, and (b) lack of appropriate remuneration for practicing in the sign language field. (Participant #55, Paraprofessional)

A total of 13% of the overall respondents to the question (n = 106) on interpreter education made reference to raising entry level education standards to university level, believing this would better equip interpreters upon entry to the world of work. It is certainly logical to expect education at a higher level and over a longer duration would reduce the "readiness to work" gap, as observed in graduates by practitioners responding to the survey.

Given the trend in the data, of criticism leveled at interpreter education programs by 67% of survey respondents, and the calls for increased quality and higher standards coming not only from participants in the study, but also from the Australian Deaf community (Napier and Rohan 2007), the researchers turned their attention to interpreter education programs. If interpreter education programs in Australia are not perceived to be preparing students for effective performance in the field, how can we do better?

Developing a program admission test

A thread generated by survey respondents regarding standards at entry level on interpreter education courses, prompted the researchers to consider the development and introduction of a screening tool that could be used at program admission. The importance of being able to measure the tool and compare against course outcomes was paramount.

In developing such a tool, the survey data provided some key skills, knowledge and abilities for consideration and inclusion. As already noted, the quantitative data confirmed significant skills gaps in Paraprofessionals. The major skills gaps for Paraprofessionals revealed by statistical analysis as reported by Bontempo and Napier (2007) were:

- Auslan skills;
- interpreting/translating skills;
- memory skills;
- concentration skills;
- listening skills.

Understanding that the above mentioned areas were significant weaknesses as identified by practitioners in their survey responses, we sought to develop a formal program admission test for interpreter education programs that would specifically tap into these areas of concern. The premise in doing so was that if we could select quality students who demonstrated greater existing skills, knowledge and abilities in these domains at the time of course commencement, we would be setting students up for success in the course and presumably in the profession of interpreting. Indeed, Patrie (1994: 56) recommends formal entrance screening specifically as a method of dealing with the "readiness to work" gap, noting that as the demands of the job are continuing to increase, "these demands call for a reasoned response, the crux of which may rest in developing parameters for interpreter preparation programs that are in line with well-developed and articulated standards for entry and exit criteria which interface appropriately with job requirements."

A standardized testing tool for program entry does not exist in Australia, so a pilot admission test was developed to address that gap, and in response to the findings of the survey, with the intention of recommending national application of the measure pending evaluation of its reliability and validity.

Study 2: Program admission test

Methodology

The pilot screening measure was developed based on comments by respondents in the research study questionnaire, and informed by the literature regarding admission testing, screening and selection of interpreters for interpreter education programs.

Paraprofessional respondents to the questionnaire specifically identified significant gaps between the importance of certain skills applicable to the task of interpreting, and their degree of competence in a particular skill domain. To that end, measuring some of these skills formed the basis of the admission test. Admission test content was further influenced by recommendations arising from spoken language screening research, particularly the findings of Moser-Mercer (1985); Gerver et al. (1984, 1989); Lambert (1991) and Sawyer (2004), given the lack of conclusive research on admission testing on signed language interpreter education programs available at the time of test development.

On the basis of data provided by Gerver et al. (1989), tests of text memory, logical memory, cloze exercises, and error detection appear to be quite predictive of future success in trainee interpreters. In addition, recommended exercises such as shadowing, paraphrasing, sight translation/interpreting, processing of numbers, and candidate interview (Moser-Mercer 1985; Lambert 1991; Pippa & Russo 2002), were considered for inclusion in the pilot test with signed language interpreters in Australia.

Participants

The pilot of the admission test was administered to 18 applicants to a Diploma of Interpreting program in Australia. Due to the nature and scale of signed language interpreter education in Australia, with typically only four programs running annually at TAFE colleges around the nation, the location and name of the TAFE college and the year of intake will not be revealed to protect the identities of participants. Furthermore, as numbers of students and practitioners around the nation are small, only general information will be given regarding the participants who did successfully gain entry to the program, with participant numbers allocated to exam results and qualitative data only and not matched with personal information about the participant.

Of the 18 applicants, a total of 11 students gained entry to the interpreting program. All were female, aged between 18–51 years of age. The mean age was 29 years of age. All students had English as their native language, with one exception. All had studied Auslan formally at TAFE, with 8 of the 11 completing the Diploma of Auslan at TAFE. Three of the 11 students held an undergraduate level university degree.

Materials

1. Admission test

After careful consideration of the "adaptability" of some of the spoken language interpreter screening items, the resulting pilot screening tool consisted of: an essay in written English (choice of 2 topics); a candidate interview conducted in Auslan; and four practical activities relating to language skills, pre-interpreting skills, and cognitive processing skills. It was considered the range of items selected would allow examiners insight into the candidates' command of English and Auslan, listening skills, memory skills, concentration, and basic ability to transfer meaning from one form into another – either intralingually or interlingually. These were all key skills gaps identified in Paraprofessionals by the survey data, and covered the range of comments from survey respondents about what should be tested for entry into an interpreter education program. Table 2 outlines the skills, knowledge and abilities we expected would be evidenced by the particular test items selected. The tests were to be administered in one sitting, and would take approximately one hour and fifteen minutes to complete.

Some tests found useful by Gerver et al. (1989) were not used for our pilot as they were peculiar to spoken language and, although it may be possible, probably would not easily convert into meaningful measures in a signed language (such as cloze sentences, synonyms etc). Also of relevance was the fact that 83% of Auslan interpreter survey respondents had English as their first language (Bontempo 2005). It is anticipated therefore that the vast majority of applicants to a Paraprofessional interpreter education program in Auslan/English interpreting in Australia have English as their first language, and a program requirement to even be considered for interview was to have successfully completed at least Year 12 English (final year of secondary school – English skills are graded according to state-wide tests at this level). The emphasis therefore in screening needs to be on Auslan skills, and this was the language highlighted by survey participants as a concern for Paraprofessionals. The NAATI description of a Paraprofessional as

Table 2. Program admission test items

Admission test item (presented/assessed in this order)	**Skills, knowledge and abilities expected to be evidenced by this test item**
Essay	Fluency in written English; motivation; goal orientation; attitude; evidence of ability to manage time; interests; ability to express thoughts
Interview	Fluency in receptive and productive Auslan; presentation skills; discourse cohesion and general communication ability; general knowledge; personality; motivation
Shadowing	Selective attention; ability to "listen and speak" simultaneously (in Auslan); processing speed relating to language manipulation; intralingual skills; contextual knowledge
Paraphrasing / identification of main ideas	Comprehension of Auslan; "listening" skills in Auslan (notes permitted); text processing; recall of main points; summarizing; discourse cohesion; language skills (in English and Auslan); spoken English skills (oral production and fluency, vocal quality etc); interlingual skills; knowledge of Deaf culture/education
Dual task	Comprehension of spoken English source material; speech discrimination; memory skills; stress management; parallel processing skills; processing digits; speed; listening skills; intralingual English skills; spoken English skills (oral production and fluency, vocal quality etc); discourse cohesion
Consecutive interpreting	Comprehension of spoken English source material; speech discrimination; Auslan skills (specifically, use of constructed action due to text chosen); basic message analysis and transfer skills (at meaning unit level as text is chunked in short idea units); semantic processing and reconstruction; discourse cohesion; interlingual skills
***Individual traits	Not a test item per se. A subjective assessment of non-language based factors (confidence, resilience in testing process, personality etc.)

able to work within a range of "conversational" level discourse was also a consideration in selecting tasks.

A written essay in English about the candidate's interests and goals (i.e. non-academic in nature) and an interview conducted in conversational Auslan were designed to elicit information from the candidate. More detail appears below. Shadowing was strongly recommended by Lambert (1991), while dual tasking and paraphrasing are cited as Moser-Mercer (2008b) as "first level" cognitive skills needed by prospective interpreters in the stages of skill acquisition. Another important pre-requisite for interpreting is comprehension, so this was assessed in multiple ways across the various test items. A simple consecutive interpreting task was selected primarily on the grounds that they are commonly included as part of a screening process in spoken language interpreter programs (Timarova & Ungoed-Thomas 2008); Humphrey (1994) also included a basic interpreting task in her comprehensive screening tool; and we were also influenced by the fact that all interpreter programs in Australia were already using a basic interpreting task in their admission testing. It was considered that these various admission test elements would offer a glimpse into the applicants' readiness for the interpreter education program. On face value at least, the admission test elements appeared to reflect the complex sub-tasks and components required in the act of interpreting, as well as revealing some aspects of aptitude for interpreting.

Applicants to the Diploma of Interpreting program were interviewed and graded by a panel of examiners, who would also form the teaching team on the course the following year. The panel consisted of one native signer (deaf) and two native English speakers (both accredited and experienced Auslan interpreters). All of the panelists had completed the minimum qualification for teaching at TAFE (Certificate IV in Assessment and Workplace Training), as well as at the time holding between them 21 years of experience in teaching signed language interpreters. Two of the three panelists held formal qualifications at postgraduate / higher degree level in either linguistics or interpreting, and one held a postgraduate degree in adult education. The panel had worked together in previous years, determining admission based on more informal and intuitive measures. The panel had opportunity to have input into the admission test developed by the researchers, and were sent a copy of the proposed tool for discussion in advance of the testing date. No changes were made, and a one-hour meeting took place prior to the admission testing to discuss the tool and grading in more detail. A fourth person from college administrative staff remained outside the interview room to coordinate the arrival of candidates, set them to task with the essay, answer any questions the candidate might have had, and escort candidates into the interview room at the appropriate time. The specific test items are elaborated on below:

Essay – English. The English essay was designed to elicit attitudes, values and motivation/commitment indicators from the candidate. It was an opportunity to assess the written English skills of applicants, but also to gain insight into "who" they are, and their reasons for undertaking interpreting studies, as well as their commitment to the program and the Deaf community. This provided some "soft skills" evaluative information about candidates such as goal orientations, attitude, and views of the Deaf community. Two essay options were presented to candidates and they could select one. The essay questions were adapted from those used by Gallaudet University in their Department of Interpretation. Gallaudet University is the only liberal arts university in the world for deaf people (hearing students who meet admissions criteria can gain entry also), and it is the only university in the world that conducts both an undergraduate degree program and a graduate degree program in signed language interpreting. On this basis it was considered a good model from which to draw the foundation of some admission testing material. Essay options after adaptation were:

Short Essay One. Explain how you may have juggled the competing demands of studies, paid employment, family/personal commitments and/or voluntary activities in the past. Articulate how this demonstrates your capacity to commit to the Diploma of Interpreting and its extracurricular requirements of attendance at Deaf community events, and observing interpreters at work. Provide any information you believe will help us better evaluate you as an applicant for this program of study.

OR

Short Essay Two. Describe why you want to become an interpreter and what you hope to achieve from the profession – what are your goals and aspirations for work in the field? Highlight the academic and life skills you possess that will help you succeed in achieving your goal/s, and what you consider your greatest asset as a future professional interpreter.

Applicants were given 30 minutes to write the essay, with two pages the minimum acceptable response. All candidates regardless of educational background were required to complete the English essay. This aspect of the admission test was worth 25 points (out of 100) and grading criteria included: content (addressing and answering the essay topic / providing evidence, examples); clarity and register of language (including correct grammar, vocabulary, spelling, punctuation); logical coherence and organisation of text; evidence of thought and analytical skill; and insights offered into personal traits / motivation / interests.

Interview – Auslan. The aim of this aspect of the admission test was to evaluate the conversational competence of the applicant in Auslan. This test essentially measures both Auslan comprehension and production by way of an interactive process between the candidate and the examiners, where the examiners draw the candidate on different topics and issues in accordance with the interview purpose (i.e. pre-prepared prompt questions were asked about motivation for studying interpreting; personal interests; current affairs; experience of the Deaf community, etc.). This test was fluid to the extent that candidate comments may generate a spontaneous question by the examiner unrelated to the partially scripted range of questions. Also, examiners might seek clarification on a comment made by a candidate for example, so the emphasis was not so much to work through a prescribed list of questions with the candidate to assess knowledge, but more so to get him/her "talking" as much as possible so language skills could be observed and rated. Invariably, however, the measure also provided non-linguistic insights too, due to the nature of the test and some of the pre-prepared "prompt" questions. This was another measure based on an existing tool employed by Gallaudet University, this time for their wider university admissions screening (not just for signed language interpreters) that we adapted for our use. The Gallaudet University American Sign Language Proficiency Interview (GU-ASLPI) is an evaluation tool used by the university for admission screening across various courses on campus, to determine linguistic fluency in American Sign Language. It is modeled after a language proficiency test developed by the US Foreign Service.[12] The GU-ASLPI is holistically scored by assigning a proficiency level of "0 to 5" by considering the candidate's performance in five areas: visual-gestural production, American Sign Language grammar, sign vocabulary, fluency and comprehension. In the adapted version of the interview the grammar and vocabulary of Auslan was evaluated rather than American Sign Language. By way of example, a proficiency level of "3" on the test would mean the candidate demonstrated "with some confidence, the ability to use some Auslan grammar along with use of signs, fingerspelling, and numbers, in everyday communication needs related to social demands, work and/or study situations. In spite of occasional hesitations, there is fair to good control of everyday sign vocabulary with which to narrate and describe topics in some detail. In spite of some noticeable imperfections, errors rarely interfere with understanding. Comprehension is fairly good as repetition or rephrasing is needed only occasionally".[13] The proficiency level (graded from 0–5) was then converted to a score out of 10 for this part of the test (test total out of 100).

Shadowing – Auslan. A short pre-recorded monologic text in Auslan was viewed with a brief introduction to set the context. Candidates were to simultaneously "phonemically shadow" the signer, copying the signer as they produced a text, matching their signs production, prosody, etc. as closely as possible, i.e. repeating each phoneme (handshape, orientation, location, movement, facial expression) as it is seen. The text selected was a female native signer in her 50s talking about her holiday around Australia. No technical vocabulary was present, although con-

[12] Refer to http://www.ntid.rit.edu/slpi/documents/FAQSLPIHistory.pdf

[13] GU-ASLPI functional descriptions at http://deafstudies.gallaudet.edu/Assessment_and_Evaluation_Unit_(AEU)/American_Sign_Language_Proficiency_Interview_(ASLPI)/ASLPI_Functional_Descriptions.html

textual knowledge of place names would have assisted the student. Given that all students were residents of Australia, this should not have been a contextually difficult text. The narrative nature of the discourse lent itself to significant use of classifier (general) signs of depiction, use of space, constructed action and constructed dialogue. For most second language learners of Auslan, these can be difficult linguistic features to acquire and in this sense the text would have been challenging. This test measures command of the students second language, with Lambert (1991) noting that one cannot shadow what one does not understand, and if a student is unable to shadow in his or her "B" language, they do not have the linguistic competence for program entry. Considerations in grading were resilience in maintaining phonemic shadowing; ability to keep pace with the signer's speed across the length of the text; clarity of production; and adoption of prosody from the source text. This aspect of the admission test was graded out of 15 total points possible.

Paraphrasing/identification of main ideas – Auslan to English. A short pre-recorded monologic text in Auslan was viewed. Notes could be taken. Upon completion, candidate was to offer a summary of the main ideas of the passage in English. The text selected was a male signer in his 30s talking about his experiences in using interpreters during his university studies. No technical vocabulary or jargon was present in the text and many of the concepts within the text should have been familiar to most course applicants if familiar with Deaf culture, education of deaf people and the potentially uneasy relationship between the Deaf community and interpreters. The signer related some positive and negative experiences of his interpreted education. This test required the applicant to visually process and comprehend the source message and to recall and reformulate in a paraphrased form in English, a summary of the main ideas presented in the source text. As this was potentially a difficult task testing pre-interpreting skills, the passage selected was very clear and simple, and was presented at a slow pace. Considerations for grading included number of main ideas presented; the coherent articulation of the ideas in English; and quality of oral output in English (audibility, clarity, etc.) This part of the admission test was graded out of 10 total points possible.

Dual-task exercise/memory – English. A short pre-recorded monologic passage in English was played, and while listening to the primary text the candidate had to write down the numbers from 100 to 1 (backwards) on paper. At conclusion of the text, the candidate was to render the text again in English. The text selected was a particularly touching story about the intent behind giving a gift, by way of a particular example given. The speaker was an American woman in her 40s. No technical vocabulary appeared in the text. This text was drawn from Patrie's (2000) exercises for the development of cognitive processing skills in English and measures ability to selectively attend to the primary task (listening and comprehending the source text) while performing a distracting activity (the number writing backwards – which adds cognitive load). At the end of the passage the candidate had to recall and present the text in English. Because the task of interpreting is a "divided attention" task (Gile 1995; as cited in Patrie 2000: 200) due to having to listen (or watch) a source text at the same time as reformulating and reproducing a target text, the interpreting student should have some capacity to manage multiple simultaneous cognitive tasks at the time of course entry. Considerations for grading included textual fidelity, coherent presentation of story in chronological order, and quality of oral output in English. This part of the admission test was graded out of 10 total points possible.

Consecutive interpreting – English to Auslan. A short pre-recorded monologic passage chunked into units of meaning was to be interpreted from English into Auslan consecutively. The text selected was drawn from Patrie (2004) and was delivered by an American woman in her 40s. She described two children bathing a dog. The text was extremely simple and contained no technical vocabulary. This task required the comprehension and analysis of the source text and the reformulation of the message from English into Auslan. This text was selected due to its contextually familiar content, brevity, and the simple chunking already built into the recording. In addition, the text required candidates to spontaneously produce classifier (general) signs of depiction, constructed action, and to spatially indicate the relationship between the parties involved in the story. The separation of listening, then reformulating each chunk, allowed candidates time to include these grammatical features in Auslan, so that Auslan skills could be assessed as well as message transmission. Considerations for grading included Auslan production, classifier use, facial expression, use of space, role shift and capacity to convey the message from one language to another. This part of the admission test was graded out of 15 total points possible.

Individual traits. An additional score was recorded by the interviewers based on impressions of candidate's overall performance from a "personal" perspective. This involved considering the interpersonal skills, presentation, and manner of the candidate – evaluating traits and behavior rather than technical skills. The evidence from the field of organizational psychology suggesting a relationship between disposition and occupational performance prompted this inclusion in the program admission test, as well as some comments from survey respondents regarding personal traits of interpreters, and the skills gap data from study one.

At the time of conducting the survey and the screening test in 2005 and 2006 respectively, the work of Stauffer and Shaw (2006), Shaw and Hughes (2007) and Lopez Gomez et al. (2007) had not yet been published. However, the preliminary results of research by Bontempo (2005) regarding the potential impact of personal traits on interpreter competence were available. Financial considerations and time factors prevented the inclusion and trialling of reliable and valid psychometric tools that could test some of the following factors in the pilot admission test at TAFE, so examiners allocated a subjective score based purely on impression and individual performance during the interview and practical tests. Specific considerations included: confidence, maturity, demeanour and presentation, stress response to screening situation/testing dynamic, cultural behavior, social skills, resilience, general behavior and professional manner. This aspect of the admission test was worth 15 points (out of 100).

Candidates had to score a minimum of 65% overall on the test; however it was expected that candidates should pass each and every section of the test, achieving at least 50% of the points allocated for each section (that is, 5/10 etc). Candidates needed more than a bare pass in each section of the test in order to reach the minimum 65% required for program entry however.

The overall results for the pilot program admission test will be compared with the exit results of the end of year examination for the same cohort of students.

2. *End of year exam*

At the end of the one year program, students undertake a final examination. The result of the final examination determines program outcome. The final exam is developed and administered by the TAFE institution, but as the course is approved by NAATI, the test format follows the standard expected by NAATI. If students do not pass the final examination, they cannot obtain their Diploma of Interpreting or their NAATI accreditation as a Paraprofessional. If they pass the final examination by 70% or greater (this benchmark is set by NAATI) they are eligible to receive their Diploma of Interpreting, assuming all other assessments across all other modules on the course have been successfully completed and deemed competent. Once they are notified that they have passed the Diploma of Interpreting, students can apply to NAATI to recognise their qualification and to be awarded the Paraprofessional level of accreditation. This only applies if NAATI has approved the TAFE (or university) as a training provider recognised by NAATI. Institutions have to apply to NAATI every 3 years to obtain ongoing approval of their program content, format and lecturing staff.

The end of year Paraprofessional interpreter examination consists of a test on DVD (to ensure standardized delivery to all students). The test has three sections, as follows:

1. Cultural and social questions. Candidates are asked four questions, two in English and two in Auslan and must answer in the same language.
(5 min – 5 points)
2. Ethical issues – as above (5 min – 5 points)
3. Dialogue interpreting (2 × 300 word dialogues in Auslan/English to be interpreted between a hearing and a deaf person):
   – Consecutive mode
   – Simultaneous mode (20 minutes – 45 + 45 = 90 points)

The overall passing grade is 70%; however, candidates must have a minimum pass in all sections of the test (i.e. a minimum of 63/90 in section three, with at least 29/45 per dialogue, and a minimum of 2.5/5 in both sections one and two). The test takes up to 40 minutes to complete in its entirety.

The data from the final examination is contrasted with the program admission test data in the results and discussion section.

Procedure for admission test

Prior to arriving at the college for admission testing, applicants received a letter advising the entrance screening process would take approximately one hour and 15 minutes to complete. They were advised they would be required to participate in a testing process that would evaluate their Auslan and English skills, as well as tests that would ascertain their readiness to participate in an interpreter education program. Screening interviews were scheduled with the 18 applicants over two days, with staggered interviewing reducing the time commitment required by the examination panel (i.e. while one candidate was doing his/her essay, the panel would be interviewing the candidate who had just finished his/her essay, and so on).

Upon arrival at the test venue applicants received an information sheet articulating the instructions for each of the tests in the entrance examination, and these instructions also clarified what the examiners would be assessing, and how, for each test item. For example, in relation to scoring the interview, candidates were given the "0–5" proficiency scale scoring mechanism so they knew before entering the test room what the examiners would be looking for. Applicants were required to read and review the test instructions / information guide for 15 minutes before commencing any part of the test. The testing period would then commence, with the first 30 minutes spent on the English essay, and the latter 30 minutes spent on the more practical elements of the admission test. The interview / practical screening aspects of the program admission process were videotaped so examiners could return to the footage later if they needed to review the performance of any candidate.

During the screening sessions over two days the scores of each panel member were collated and averaged to offer a final result for each candidate, which also included the result of the English essay. As noted, candidates had to pass every section of the measure, as well as achieve an overall minimum of 65% in the admission test to be admitted for program entry. The 65% overall minimum was set as an achievable figure to allow progression through the course to the NAATI benchmark of 70% on the final test. However, only the top twelve students were expected to be selected for course entry, so achieving the minimum score was no guarantee of course acceptance. The overall time commitment from each candidate was up to one hour in total for all parts of the test (plus an additional 15 minutes for reading time).

Results and discussion

As already mentioned, of the 18 people that applied for program entry, 11 were accepted on the basis of admission test results, meaning 61% of presenting applicants gained program entry. Student admission test results were compared with their end of year final examination scores. Details are highlighted in Table 3, with ranking based on final examination score.

As noted in Table 3, students 7 through to 11 did not pass the final examination, as the pass mark for the final examination was 70%, a prescribed pass mark set by NAATI. The mean admission test score for the students who passed the

Table 3. Comparison of admission test score and final examination result

Candidate	Admission test result (%)	Final examination result (%)
1	71.3	81.75
2	74.8	79.25
3	77.2	77.50
4	77.5	72
5	74.1	71
6	66.6	71
7	74.1	65.75
8	75.5	64
9	74.0	63.75
10	73.0	62.5
11	69.9	60.75

final examination was 73.58%, with a mean final examination score of 75.42%. Of the group of students who failed the final examination, the mean admission test score was 73.30% and the final examination mean score was 63.35%. In total, only 55% of candidates admitted to the program successfully completed it.

Similar program exit results were identified by Timarova & Ungoed-Thomas (2008) in their review of 18 spoken language interpreter education programs. They found that, on average, admission tests accept only 24% of applicants (61% of applicants gained entry in this study), and of the admitted candidates, only 56% successfully completed the interpreting program. This pilot study produced similar end results, with just over half the accepted candidates who gained program entry passing the final examination (55%).

Our findings suggest the admission test results from this small scale pilot study were not predictive of final examination performance. The mean admission scores for the students who passed and the students who failed the final examination differed by only 0.28%. In hindsight, the admission test we developed leaned towards testing pre-interpreting skills (i.e. existing ability and declarative knowledge) and only vaguely tapped into individual aptitude per se, in an ill-defined fashion due to many of the test items actually testing several different aspects of skills and abilities even within one test.

The development and administration of this pilot test shed light on the need for greater emphasis on objectively assessing aptitude in signed language interpreter program entrance screening via psychometrically valid tools, measuring cognitive and affective factors rather than performance on a series of tasks that may be variants of interpreting skills as such. Such tests may assist educators in more effectively selecting students who have the *capacity to learn and transfer* new skills and knowledge across different environments, rather than only selecting students who have existing basic technical skills.

With general mental ability recognised to be the single most significant predictor of occupational performance (Schmidt & Hunter 1998), and to be more successful in recruiting suitable people into the interpreting profession, perhaps we need to seriously consider introducing general intelligence testing in some form for screening purposes, alongside measures of language proficiency and temperament. In addition to its role in predicting on the job performance, Ree and Earles (1992) confirm that general intelligence is the strongest predictor of job training success also, adding further weight to the suggestion to apply cognitive ability tests to interpreter program applicants.

It is unclear whether the program admission test resulted in allowing people program entry that in fact should were not ideal candidates for the program. Of greater concern, however, is the possibility that program admission test results may have also excluded people from the course who in fact would have been competitive students, if given an opportunity to gain entrance.

Many of the exercises incorporated into our program admission test were measures used internationally with spoken language interpreters accessing conference interpreter level programs and had never been used with signed language interpreters in Australia for program entry testing purposes before. Specifically, shadowing, paraphrasing, and dual tasking were new assessment items not used by any educator previously for program admission testing in Australia. We were able to adapt these to suit our purposes, and we obtained texts that were authentic and appeared well-suited for the tasks. However, the validity of these particular exercises for entry level Auslan interpreters, who will work primarily with dialogic discourse in community settings rather than in a conference environment, remains uncertain. The test results did not show any particular test item as an important predictor, and the inclusion of a consecutive interpreting task (albeit already in place in all the Australian screening tests, and used in international screening tests also) also may not be an effective way to evaluate interpreter-potential.

Prior to their final examination students were asked to elaborate on their perceptions of the usefulness and relevance of the program admission test, in hindsight. Of the 11 students in the program, only four volunteered to provide feedback on the screening tool. Of the four respondents, two ultimately passed the final examination and two were unsuccessful in passing the interpreter program.

All the respondents were in favor of screening at program entry. Each candidate gave feedback on each of the admission test items, and trends in the data supported the use of an essay in English and a candidate interview to assess language proficiency in Auslan. The shadowing; paraphrasing/identification of main idea; and the dual task test items did not prove popular overall, with respondents reporting these as daunting test items at the time, noting they'd never been exposed to such exercises in their language acquisition classes. However, candidates also noted that in hindsight they could see the value in such exercises in terms of their application to the interpreting process. Respondents were all in favor of the consecutive interpreting task.

Some direct quotes representative of the feedback include:

> *English essay* – The essay choice of two thought questions were good. Really made students think about their future goals and their commitment to the Deaf community (Candidate #7 – unsuccessful student)
>
> *Candidate interview* – This task should be relatively easy to a person who wants to enter the interpreter's course (Candidate #5 – successful student)
>
> *Shadowing* – Once you get the sense of where the story is heading the exercise becomes a lot easier (Candidate #5 – successful student)

> *Identification of main idea* – This type of assessment is essential to show the students' ability to first understand what they are seeing, to remember details, and to then give a summary (Candidate #9 – unsuccessful student)
>
> *Dual task* – I can see this would be a valuable tool to assess students' ability to remember details while processing something else (Candidate #9 – unsuccessful student)
>
> This task is just like multi-tasking – a skill which I now realise an interpreter must have. You must be able to hold and listen to something in English and deliver it in Auslan a few seconds later (depending on your time lag) whilst still listening to the next lot of information that will require interpreting (Candidate #5 – successful student)
>
> *Consecutive interpreting* – The passage shown was a very good choice as it involved the use of space and many classifiers (Candidate #7 – unsuccessful student).

Candidate #6 (successful student) who had actually unsuccessfully attempted program entry in a previous year (when entrance testing had been more intuitive) noted of the whole process:

> …the interview process was strange but I remember on the way home thinking how much better it was because it tested my individual skills and if they were good enough to handle interpreting. For example, testing my memory, and my ability to multi-task in the dual task exercise. Overall this latest method was a lot more effective in testing my abilities.

The low response rate to the call for feedback on the admission test (only 4 respondents out of a cohort of 11 students) is a limitation of this aspect of the study, and needs to be considered when interpreting the comments. In terms of using self-report data, the veracity of reports from participants can sometimes be of concern, and can be influenced by social desirability bias. This is a flaw of all social research survey design, and not unique to this study however.

Limitations of study 2

A number of confounding variables could have impacted student progress and performance between the time of program admission testing and the final examination. For example, the quality of instruction in the program over the duration of the year would be important, and is not measured in this study. As noted both in the survey responses, and in the literature, the role of the educator is very powerful (Hattie 2003; Robinson 2002). Furthermore, issues surrounding the transition from language student to interpreting student can throw learners off track (Shaw, Grbic & Franklin 2004); family and faculty support play a part (Shaw & Hughes

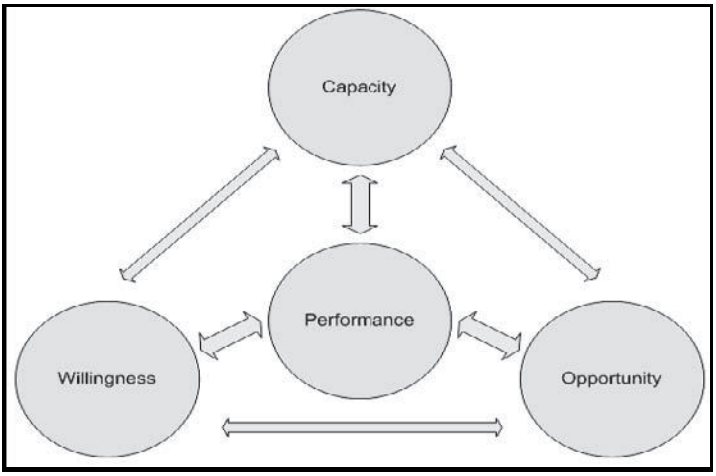


Figure 1. Performance = Opportunity × Capacity × Willingness – Determinants of human performance and their interaction (adapted from Blumberg & Pringle 1982 by Moser-Mercer 2008a: 3)

2006); and in particular, the student's learning style, attitude, motivation and willingness to engage and improve is also critical. Moser-Mercer (2008a) has adapted a visual representation of the determinants of human performance, as follows in Figure 1. Such a model clearly shows the interaction between an individual student's aptitude/learning style/intelligence/abilities (capacity), motivation and attitude (willingness) and opportunity to practice, and performance.

With individual capacity and willingness suggested to play such a significant role in determining performance, it is vital for interpreter program admission tests to start evaluating the aptitude of program applicants more effectively than is the case currently.

Additionally, a consideration that would have affected the pilot admission test outcomes is that the panel of examiners were using this tool for the first time. A lack of training and experience in administering a complex screening tool and in understanding how to allocate scores may well have influenced admission test results for the pilot.

Another potential limitation of the pilot is that it is precisely that – a preliminary study. Such a small scale preliminary study prevents any serious treatment of the results at this time, and the admission test therefore remains a work in progress. Developing expertise in administering and grading an admission test; and in collecting data from more Diploma of Interpreting cohorts from different programs around the country would be useful, as a greater sample will allow for a more comprehensive study, generalization of the results, and firmer recommendations arising from the results.

A final note of reservation could be raised about the reliability and validity of the examination used to assess students at the end of the interpreter education program. Although the examination has good face validity, concerns about NAATI test format and content have been flagged by Campbell and Hale (2003). As the final examination was modeled on the NAATI Paraprofessional test (as required by the Diploma curriculum), the researchers were restricted in being able to develop an alternative final examination. Thus we acknowledge that the reliability and validity of the final examination could be a possible confounding factor in the study, in that perhaps the final examination did not measure what it is supposed to, and the admission test may not actually be the problem.

The caution from Sawyer (2004) regarding course duration is also noteworthy here. It may be that some of the candidates accepted into the program based on their performance on the admission test pilot could not sustain and improve performance to the standard required by the end of the program. However, had the course duration been longer, it could be speculated that perhaps these candidates would have met the exit standard required. The program admission test, in largely evaluating ability (rather than aptitude), may have correctly pegged candidates abilities at the time of course entry; however, the capacity of candidates to learn and transfer the necessary new skills within the period of the course was not measured.

Despite the challenges evident in the admission testing process outlined in this pilot study, the notion of standardized screening needs to remain on the agenda. Although this preliminary attempt to administer an admissions test was not conducive to predicting candidates' success in passing the program of study, the fact that some kind of admissions screening is needed (and research on such screening tools and their link to program outcomes is most definitely needed) by interpreter practitioners, interpreter educators, interpreting students and researchers, cannot be denied. This study is therefore a first step in attempting to more effectively recruit quality candidates into entry level interpreter education programs in Australia.

In the absence of any other hard data provided on interpreter education program admission and exit outcomes for signed language interpreters in Australia, this study breaks new ground. Evidently much more work needs to be done in exploring admission test options, streamlining admission processes nation-wide, collecting and reporting data, and in addressing the reliability and validity of the tools used to admit and exit students from programs; and to certify candidates via direct NAATI testing.

By way of a postscript, although full details are unavailable, the researchers were advised that the interpreter education program involved in the test pilot employed the same screening tool again a year later with the next prospective cohort of interpreting students (before knowing the outcome of final examinations for the first pilot group). The same panelists were involved in entry test administration (although one of the three panelists did not remain on the teaching team – a new teacher joined the program), and of 20 applicants to the course, 10 were selected for program admission. A 100% pass rate on the final examination for this later cohort of 10 students was reported. Full data is not available to the researchers for this cohort of students; however the significantly improved results in many ways simply muddies the water further, and warrants further research.

Conclusion

In conclusion, we return to the specific research questions for the two related studies:

1. Are signed language interpreter education programs in Australia perceived by practitioners to be preparing interpreters for effective performance in the profession?
   According to the data collected via the survey, the majority of practitioner respondents (67%) do not perceive the interpreter education programs in Australia for signed language interpreters to be preparing interpreters for effective performance in the profession.
2. Can the interpreter education program admission tests commonly referenced in the literature for spoken language interpreters be adapted and applied to signed language interpreters for entry level screening purposes?
   A selection of the exercises commonly used for program admission testing purposes for spoken language interpreters were successfully adapted to suit the needs of signed language interpreters in this study.
3. Are the results of program admission tests developed and administered in this study predictive of final examination performance?
   The program admission test developed and administered in this study was not predictive of final examination performance.

The finding that entry level practitioners (Paraprofessionals) demonstrate significant skills gaps and that interpreter education courses are perceived to be inadequately preparing interpreters for the world of work should ring alarm bells for interpreter educators and program administrators in Australia. Evidently an urgent review into current practices and an overhaul of the national curriculum, instructional quality, resources, and so on, may be needed.

The signed language interpreting sector has much to learn from the path traveled by our peers in the spoken language interpreting field (and vice versa also). If developments and progress are occurring in the field and being documented in the literature, we can gain from this evidence-based approach to interpreting pedagogy, and in particular by reviewing and sharing forms of testing and assessment. Trialing methods and exercises that have proven useful to others is one approach to continuous improvement, and our efforts to adapt, adopt, and document the use of testing techniques should encourage others in our field to do the same.

A goal, and a challenge, remains for us in refining and further trialing a suitable admission screening tool for interpreter course entry to support standardized entry level competence in programs across the nation. It is an iterative process, and undoubtedly revisions will lead to a more robust screening measure.

Sawyer's (2004) assertion that screening instruments have to vary considerably from program to program, and that a single screening instrument may never be possible to develop, may be less valid in the Australian context. Given that TAFE interpreter education is delivered according to a national competency based curriculum, which defines the number of program hours available and resources for programs around the nation, it may in fact be possible to develop a single screening instrument to be used to assess all Auslan interpreting program applicants throughout Australia (and an adapted tool for spoken language applicants, as spoken and signed language interpreter education programs at TAFE in Australia all adhere to the same curriculum).

At present, colleges duplicate processes around the nation, with incumbent coordinators developing and trialing entrance examinations with little or no moderation with colleagues, and no collection of data to determine the predictive validity of the screening procedures employed. Preliminary discussions with program coordinators suggest considerable support for a national standardized approach to program screening, in the hope that not only will it lead to better student outcomes and improved professional standards, but that the administrative load on program coordinators will be somewhat alleviated by a national approach to admission testing.

Acknowledgments

The participants in the initial national research study regarding interpreter competence and performance, and the participants in the later pilot admission testing process and interpreter education program targeted for this study, are all acknowl-

edged with sincere gratitude. Without the enthusiastic contribution of their time and their thoughts, we would not have gained this further insight into the signed language interpreting profession. Their views and experiences will help shape future developments in interpreter education and program admission testing in Australia. Similarly, the support of the interpreter educators and the institution involved in piloting the admission test is greatly appreciated – their willingness to entertain a new way of doing things, and embracing a new process in the interests of improving program quality and course outcomes, has been inspiring.[14]

References

Anderson, Glenn B. and Stauffer, Linda K. 1990. *Identifying Standards for the Training of Interpreters for Deaf People*. University of Arkansas, Rehabilitation Research and Training Center on Deafness and Hearing Impairment.

Angelelli, Claudia. 2007. "Assessing Medical Interpreters: The Language & Interpreting Testing Project." *The Translator* 13 (1): 63–82.

Bachman, Lyle and Cohen, Andrew. (Eds.). 1998. *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Bernstein, Jared and Barbier, Isabella. 2000. "Design and Development Parameters for a Rapid Automatic Screening Test for Prospective Simultaneous Interpreters." *Interpreting* 5 (2): 221–238.

Bontempo, Karen. 2005. *A Survey of Auslan Interpreters' Perceptions of Competence.* Unpublished manuscript. Macquarie University, Australia.

Bontempo, Karen. 2008. "Personality Matters! Measuring Performance Predictors in Signed Language Interpreters." Unpublished research study presented at the Conference of Interpreter Trainers, San Juan, Puerto Rico, USA. October 22–25, 2008.

Bontempo, Karen and Levitzke-Gray, Patricia. 2009. "Interpreting Down Under: Signed language interpreter education and training in Australia." In *International perspectives on signed language interpreter education,* Jemina Napier (Ed), Washington, DC: Gallaudet University Press.

Bontempo, Karen and Napier, Jemina. 2007. "Mind the Gap: A Skills Analysis of Sign Language Interpreters." *The Sign Language Translator and Interpreter* 1 (2): 275–299.

Borum, Randy, Super, John and Rand, Michelle. 2003. "Forensic Assessment for High Risk Occupations." In *Handbook of Psychology,* Irving B. Weiner, Donald K. Freedheim, John A. Schinka and Wayne F. Velicer (Eds.), 133–147. John Wiley and Sons.

Bozionelos, Nikos. 2004. "The Relationship Between Disposition and Career Success: A British Study." *Journal of Occupational and Organisational Psychology* 77: 403–420.

Brisau, Andre, Godijns, Rita and Meuleman, Chris. 1994. "Towards a Psycholinguistic Profile of the Interpreter." *Meta* 39 (1): 87–94.

[14] This research was conducted by Karen Bontempo, PhD candidate in the Department of Linguistics, Macquarie University, under the supervision of Dr Jemina Napier.

Button, Scott B., Mathieu, John E. and Zajac, Dennis M. 1996. “Goal Orientation in Organisational Research: A Conceptual and Empirical Foundation.” *Organisational Behaviour and Human Decision Processes* 67 (1): 26–48.

Campbell, Stuart and Hale, Sandra. 2003. “Translation and Interpreting Assessment in the Context of Educational Measurement.” In *Translation Today: Trends and Perspectives*. Gunilla. M. Anderman and Margaret Rogers (Eds). 205–220. Clevedon: Multilingual Matters.

Choi, Namok, Fuqua, Dale R. and Griffin, Bryan. W. 2001. “Exploratory Analysis of the Structure of Scores from the Multidimensional Scales of Perceived Efficacy.” *Educational and Psychological Measurement* 61 (3): 475–489.

Clifford, Andrew. 2005. “Putting the Exam to the Test: Psychometric Validation and Interpreter Certification.” *Interpreting* 7 (1): 97–131.

Colina, Sonia. 2008. “Translation Quality Evaluation: Empirical Evidence for a Functionalist Approach.” *The Translator* 14 (1): 97–134.

Dodds, John. 1990. “On the Aptitude of Aptitude Testing.” *The Interpreters’ Newsletter* 3: 17–22. EUT – Edizioni Università di Trieste.

Doerfert, Karen and Wilcox, Sherman. 1986. “Meeting Students Affective Needs: Personality Types and Learning Preferences.” *Journal of Interpretation* 3: 35–43.

Finton, Lynn. 1998. Pre-Interpreting Skills: “Laying the Foundation Curriculum Considerations and Instructional Strategies.” In *The Keys to Highly Effective Interpreter Training: Proceedings of the 12th National Convention of the Conference of Interpreter Trainers* Alvarez, J. (Ed.). USA: CIT.

Frishberg, Nancy. 1986. *Interpreting: An Introduction*. Silver Spring, MD: RID Publications.

Gerver, David, Longley, Patricia E., Long, John, and Lambert, Sylvie. 1984. “Selecting Trainee Conference Interpreters: A Preliminary Study.” *Journal of Occupational Psychology* 57 (1): 17–31.

Gerver, David, Longley, Patricia E., Long, John, and Lambert, Sylvie. 1989. “Selection Tests for Trainee Conference Interpreters. *Meta* 34 (4): 724–735.

Goff-Kfouri, Carol A. 2004. “Testing and Evaluation in the Translation Classroom.” *Translation Journal* 8 (3).

Hale, Sandra and Campbell, Stuart. 2002. “The Interaction Between Text Difficulty and Translation Accuracy.” *Babel* 48 (1): 14–33.

Hattie, John. 2003. “Teachers Make a Difference: What is the Research Evidence?” Australian Council for Educational Research, October 2003. Retrieved on 12 June 2008 from http://www.education.auckland.ac.nz/uoa/fms/default/education/docs/pdf/arts/john%20hattie/influences/Teachers_make_a_difference_-_ACER_(2003).pdf

Humphrey, Janice. 1994. “Effective Screening Procedures for Entering Students in Interpreter Education.” In *Mapping our Course: A Collaborative Venture: Proceedings of the 10th National Convention of the Conference of Interpreter Trainers.* Elizabeth Winston (Ed.) USA, CIT.

Jackson, Sherri L. 2003. *Research Methods and Statistics: A Critical Thinking Approach*. Belmont: Wadsworth Thomson

Kozaki, Yoko. 2004. “Using GENOVA and FACETS to Set Multiple Standards on Performance Assessment for Certification in Medical Translation from Japanese into English.” *Language Testing* 21 (1): 1–27.

Kurz, Ingrid. 2001. “Small Projects in Interpretation Research.” In *Getting Started in Interpreting Research: Methodological Reflections, Personal Accounts and Advice for Beginners*. Daniel Gile (Ed). Philadelphia: John Benjamins Co.

Kurz, Ingrid. 2003. "Physiological Stress During Simultaneous Interpreting: a Comparison of Experts and Novices." *The Interpreters' Newsletter* 12. EUT – Edizioni Università di Trieste.

Lambert, Sylvie. 1991. "Aptitude Testing for Simultaneous Interpretation at the University of Ottawa." *Meta* 36 (4): 586–594.

Lauscher, Susanne. 2000. "Translation Quality Assessment: Where Can Theory and Practice Meet?" *The Translator* 6 (2): 149–168.

Lopez Gómez, María José, Teresa Bajo Molina, Presentación Padilla Benítez and Julio Santiago de Torres. 2007. "Predicting Proficiency in Signed Language Interpreting." *Interpreting* 9 (1): 71–93.

Losier, Gaetan F. and Vallerand, Robert J. 1994. "The Temporal Relationship Between Perceived Competence and Self-Determined Motivation." *Journal of Social Psychology* 134 (6).

Maurer, Todd J., Wrenn, Kimberley A., Pierce, Heather R., Tross, Stuart A. and Collins, William C. 2003. "Beliefs About 'Improvability' of Career-Relevant Skills: Relevance to Job/Task Analysis, Competency Modeling, and Learning Orientation." *Journal of Organisational Behaviour* 24: 107–131.

Monikowski, Christine. 1994. "Issue – Proficiency." In *Mapping our Course: A Collaborative Venture: Proceedings of the 10th National Convention of the Conference of Interpreter Trainers*. Elizabeth Winston (Ed.). USA, CIT.

Mortensen, Diane. 2001. Measuring Quality in Interpreting. A Report on the Norwegian Interpreter Certification Examination (NICE). Retrieved 5 June, 2007.

Moser-Mercer, Barbara. 1985. "Screening Potential Interpreters." *Meta* 30 (1): 97–100.

Moser-Mercer, Barbara. 2008a. "Skill Acquisition in Interpreting: A Human Performance Perspective." *The Interpreter and Translator Trainer* 2 (1): 1–28.

Moser-Mercer, Barbara. 2008b. "Developing Expertise in Interpreting" [on line] retrieved 15/9/08 from http://www.emcinterpreting.org/resources/module1.php

NAATI. 2007. "Levels of Accreditation" [on line] retrieved 10/10/07 from http://www.naati.com.au/at-accreditation-levels.html

Napier, Jemina and Rohan, Meg. 2007. "An Invitation to Dance: Deaf Consumers' Perceptions of Signed Language Interpreters and Interpreting." In *Translation, Sociolinguistic, and Consumer Issues in Interpreting*. Melanie Metzger and Earl Fleetwood (Eds). Washington D.C.: Gallaudet University Press.

Niska, Helge. 2005 (Retrieved 28 August 2005). "Testing Community Interpreters: A Theory, a Model and a Plea for Research." From http://lisa.tolk.su.se/00TEST.HTM

Nisula, Marjukka & Manunen, Juha. 2009. "Sign Language Interpreter Training in Finland." In *International Perspectives on Signed Language Interpreter Education.* Jemina Napier (Ed.) Washington, DC: Gallaudet University Press.

Oakes, David W., Ferris, Gerald R., Martocchio, Joseph J., Buckley, M. Ronald and Broach, Dana. 2001. "Cognitive Ability and Personality Predictors of Training Program Skill Acquisition and Job Performance." *Journal of Business and Psychology* 15 (4): 523–548.

Onwuegbuzie, Anthony J., Bailey, Phillip and Daley, Christine E. 2000. "Cognitive, Affective, Personality and Demographic Predictors of Foreign Language Achievement." *Journal of Educational Research* 94 (1): 3–15.

Orima. 2004. *A Report on the Supply and Demand for Auslan Interpreters across Australia*. Commonwealth Department of Family and Community Services. Commonwealth of Australia. http://www.facs.gov.au/disability/auslan_report/contents.htm

Ozolins, Uldis and Bridge, Marianne. 1999. *Sign Language Interpreting in Australia*. Melbourne, Language Australia.

Patrie, Carol. 1994. "The Readiness to Work Gap." In *Mapping our Course: A Collaborative Venture: Proceedings of the 10th National Convention of the Conference of Interpreter Trainers.* Elizabeth Winston (Ed.) USA, CIT.

Patrie, Carol. 2000. *The Effective Interpreter Series: Cognitive Processing Skills in English*. San Diego, Dawn Sign Press.

Patrie, Carol. 2004. *The Effective Interpreter Series: Consecutive Interpreting from English.* San Diego, Dawn Sign Press.

Pippa, Salvador and Russo, Mariachiara. 2002. "Aptitude for Conference Interpreting: A Proposal for a Testing Methodology Based on Paraphrase." In *Interpreting in the 21st Century: Challenges and Opportunities.* Giuliana Garzone and Maurizio Viezzi (Eds) Philadelphia: John Benjamins, 245–256.

Ree, Malcolm J. and Earles, James A. 1992. "Intelligence is the Best Predictor of Job Performance." *Current Directions in Psychological Science* 1 (3): 86–89.

Ree, Malcolm J., Earles, James A. and Teachout, Mark S. 1994. "Predicting Job Performance: Not Much More than "g"." *Journal of Applied Psychology* 79 (4): 518–524.

Reeve, Charlie L. and Heggestad, Eric D. 2004. "Differential Relations Between General Cognitive Ability and Interest-Vocation Fit." *Journal of Occupational and Organisational Psychology* 77: 385–402.

Riccardi, Alessandra, Marinuzzi, Guido and Zecchin, Stefano. 1998. "Interpretation and Stress." *The Interpreters'Newsletter.* 8. EUT – Edizioni Università di Trieste

Roberts, Roda P. 1994. "Student Competencies in Interpreting: Defining, Teaching and Evaluating." In *Mapping our Course: A Collaborative Venture: Proceedings of the 10th National Convention of the Conference of Interpreter Trainers*, Elizabeth Winston (Ed.). USA, CIT.

Robinson, Peter. 2002. "Researching Individual Differences and Instructed Learning." In *Individual Differences and Instructed Language Learning.* Peter Robinson (Ed). Philadelphia: John Benjamins Co.

Rudser, Steven F. and Strong, Michael. 1986. "An Examination of Some Personal Characteristics and Abilities of Sign Language Interpreters." *Sign Language Studies* 53: 315–331.

Russo, Mariachiara and Pippa, Salvador. 2004. "Aptitude to Interpreting: Preliminary Results of a Testing Methodology Based on Paraphrase." *Meta* 49 (2): 409–432.

Sawyer, David. B. 2004. *Fundamental Aspects of Interpreter Education*. Philadelphia: John Benjamins.

Schmidt, Frank L. and Hunter, John E. 1998. "The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings." *Psychological Bulletin* 124 (2): 262–274.

Schweda Nicholson, Nancy. 2005. "Personality Characteristics of Interpreter Trainees: the Myers-Briggs Type Indicator (MBTI)." *The Interpreters' Newsletter*. 13. EUT – Edizioni Università di Trieste.

Seal, Brenda C. 2004. "Psychological Testing of Sign Language Interpreters." *Journal of Deaf Studies and Deaf Education* 9 (1): 39–52.

Shaw, Risa, Collins, Steven, and Metzger, Melanie. 2006. "MA to BA: A Quest for Distinguishing Between Undergraduate and Graduate Interpreter Education at Gallaudet University Bachelor of Arts in Interpretation Curriculum: Gallaudet University." In *New approaches to Interpreter Education*. Cynthia Roy (Ed.) Washington, DC: Gallaudet University Press.

Shaw, Sherry, Grbic, Nadja and Franklin, Kathy. 2004. "Applying Language Skills to Interpretation." *Interpreting* 6 (1): 69–100.

Shaw, Sherry and Hughes, Gail. 2006. "Essential Characteristics of Sign Language Interpreting Students: Perspectives of Students and Faculty." *Interpreting* 8 (2): 195–221.

Slatyer, Helen and Carmichael, Andrew. 2005. *NAATI Rater Reliability Report* (Unpublished research report). Sydney: Macquarie University & NAATI.

Slatyer, Helen, Elder, Catherine, Hargreaves, Marian and Luo, Kehui. 2006. *An Investigation into Rater Reliability, Rater Behaviour and Comparability of Test Tasks* (Unpublished research report). Sydney: Macquarie University & NAATI.

Solow, Sharon N. 1998 . "A Highly Effective Sequence of Training Experiences." In *The Keys to Highly Effective Training: Proceedings of the 12th National Convention of the Conference of Interpreter Trainers* Alvarez, J. (Ed.). USA: CIT.

Stansfield, Charles. W and Hewitt, William. 2005. "Examining the Predictive Validity of a Screening Test for Court Interpreters." *Language Testing* 22 (4): 438–462.

Stansfield, Charles W., Scott, Mary L., & Kenyon, Dorry M. 1992. "The Measurement of Translation Ability." *The Modern Language Journal* 76 (4) 455–467.

Stauffer, Linda K. and Shaw, Sherry. 2006. "Personality Characteristics for Success in Interpreting Courses: Perceptions of Spoken and Signed Language Interpretation Students." *Journal of Interpretation*: 11–24.

Timarova, Sarka and Ungoed-Thomas, Harry. 2008. "Admission Testing for Interpreting Courses." *The Interpreter and Translator Trainer* 2 (1): 29–46.

Standards as critical success factors in assessment

Certifying social interpreters in Flanders, Belgium

Hildegard Vermeiren, Jan Van Gucht
and Leentje De Bontridder
University College Ghent / COC – Lessius University College,
University of Louvain / COC

This article gives an account of a professional procedure for the assessment of interpreters, namely the certification exams for social interpreters in Flanders, Belgium. After developing a professional profile for social interpreters, the authors align the profile of the graders with the certification exam. Special attention is given to standards, both for gate keeping of social interpreters and for the graders themselves. The introductory sociological framework underlines the importance of legitimization for institutions for gate keeping in the professional domain. Using the concrete test procedure as a starting point, the authors exemplify how competency based evaluation grids and expert knowledge of graders are determining factors for the legitimacy of the certifying institution.

Introduction

This chapter deals with the certification process of social interpreters in Flanders, Belgium. Certification is a specific kind of assessment. Broadfoot (1996: 68) states that assessment is one of the most central features of the rationality that underpins advanced industrial society itself. Rationality is "the quality of being reasonable or of being acceptable to reason," which means "be[ing] based on, or in accordance with or justified by principles of reason or logic" (VandenBos 2007). We will concentrate on the actual certification process and the challenges it poses to an objective assessment, i.e. a judgment in accordance with or justified by principles of reason and logic. Such judgment should be impartial, uninfluenced by personal feelings, interpretation or prejudice (VandenBos op.cit.).

Flanders is one of the two autonomous regions comprising the federal state of Belgium, covering an area of 13,684 km^2 and with a population of about 6 million. The official language in this area is Dutch, and its capital is Brussels. During the past few decades the Flemish regional government has faced the challenges of immigration and the emerging reality of a multicultural and multilingual society, causing the government to develop a broad policy for civic integration. One key element of civic integration policy is the provision of social interpreting to immigrants. Social interpreting is a concept used only in Flanders and refers to the above-mentioned federal structure of Belgium. Social interpreting is the "faithful, complete and neutral transfer of oral messages from a source language into a target language in the sphere of public and social services and public and social care"(SERV 2007: 7). Thus, social interpreting is community interpreting, excluding interpreting in the legal, police and asylum contexts. Like community interpreting, most Flemish social interpreting consists of liaison interpreting in the consecutive mode (Salaets et al.: 2008). It covers both interpreting in face-to-face situations and interpreting provided over the telephone (Wadensjö 1989: 33). Social interpreting service providers are subsidized non-profit organizations or governmental organizations. In 2004, the COC, Central Support Cell for Social interpreting and translation, http://www.sociaaltolkenenvertalen.be, was founded to support and develop this relatively new sector.

Social interpreting in a context of increasing rationality

Audits, assessments and other evaluation procedures have become commonplace in modern Western society. They are the exponents of the need for rationality (Weber 1923, 1947) and control (Durkheim 1947; Foucault 1977; Lianos 2003) surging at the beginning of the 20th century and which are still increasing in importance in our society. Consequently, the most adequate overall frameworks that explain the rise of audit, assessment, and evaluation procedures are those of the sociology of organizations.

Lianos (2003) provides us with a useful framework: a theoretical update of Foucault's reflections on the issue of control and the subject in today's society. Lianos views institutions as any source of mediating activity between human beings and particularly as an important source of normativity (Lianos op.cit.: 413). Institutions create a regulating universe and monitor and validate highly specific aspects of citizen behavior (Lianos op.cit.: 414). This institutional control is more often than not perceived as beneficial and sometimes even as liberating, rather than as constraining. This type of control is part of a service offered to the public as "users" (Lianos op.cit.: 415, emphasis in the original). Lianos (op.cit.: 416) also stresses that institutional control is by definition impersonal in its origin and atomized in its reception, stating that there is no interaction between the institution and the user. A specific mechanism of this kind of institutional control is the dispersion of 'discipline' (Cohen 1979 as cited in Lianos op.cit.: 425) through the injection of values or norms into the subject. Assessment plays an important part in this process of controlling and dispersing values and norms. It is foremost a tool for control of the efficient division of labor, for quality assurance and for improvement through vertical and horizontal diversification of (prospective) workers. Broadfoot (1996) studies how modern educational systems and their assessment procedures play a part in the rationalization and the overall control of society. Assessments may be used not only in an educational context, but also in professional selection, e.g. by means of credentialing. The purpose of credentialing is to provide the public with a dependable mechanism to identify practitioners of a certain profession who have met certain standards (APA op.cit.: 63).

In addition, Broadfoot (op.cit.: 107) stresses that assessment and other similar rational procedures provide ideological self-legitimization to qualified authorities. Authorities (certifying or other) thus legitimize themselves versus numerous stakeholders: the government, the educational system, professional organizations, private companies, service providers and their clientele, employees, and students. These stakeholders rightfully expect certification (or credentialing) procedures to be valid and reliable, and consider these authorities accountable to them.

Another major factor in both the quality of an assessment procedure and the legitimization of the monitoring authority is the role and qualification of the graders. Weiler (1981: 16–17) stresses the importance of expertise as a source of legitimization.

Certification and assessment as a means of legitimization

The certification of social interpreters in Flanders must be viewed in the above-mentioned context of increasing rationality and control. This control expresses itself through a growing demand for accountability from, and the search for legitimization by, the certifying authorities.

Interpreting, and more specifically, the domain of community interpreting (and consequently Flemish social interpreting) are activities that are closely linked to the general evolution of society. Traditionally, liaison interpreting in hospital contexts, public services, schools etc. was characterized by informality and ad-hoc interpreting. The job was usually done by non-professionals such as children, friends and neighbors. Interpreting was not considered a profession, but merely a service to friends, family or members of one's own ethnic group.

Nevertheless, the job done by "good but unskillful Samaritans, self-appointed experts and unscrupulous fixers who often 'helped' their less linguistically gifted compatriots for a 'fat fee,' as stressed by Niska (1991: 98), was a clear proof of a lack of rationality and control in the domain, and, not in the least, of a lack of awareness of accountability.

Due to growing migration issues in the last few decades, concerns arose for the welfare of minority, immigrant and refugee populations. Health authorities, for example, show an increasing concern to ensure the provision of services to people who are unable or unwilling to communicate (Wadensjö 1998: 37). Gradually, awareness of the role of interpreters in ensuring equal access to social services has increased. Wadensjö (op.cit.: 37) additionally stresses that civil rights and civil responsibilities are two sides of the same coin.

This awareness has further resulted in the creation of specific institutions that oversee the organization of the sector. Such agencies are accountable to their stakeholders: policy makers, but also financiers and the public in general. Certifying agencies or authorities (Lianos op.cit.: 413) do not create a formal institutionalized group with the interpreters they certify. Lianos' loose concept of organizations is particularly relevant for the situation of social interpreters. Social interpreters usually work as independents, and the certifying organization monitors its members and socializes with them by dispersing norms among them under the form of guidelines and interpreting deontology. As stressed by Lianos (op.cit.: 415), the beneficial effects include the interpreter obtaining a certificate, serving as an asset on the labor market, and providing the user with a professional interpreter.

Emerging organizations have gradually set up a structure to gain control over an informal sector. These organizations share a number of characteristics. Katz and Gartner (1988: 432) provide us with four fundamental founding characteristics for organizations in general: *intentionality, boundaries, resources,* and *exchange*. When applied to an emerging certifying organization, *intentionality* means the organization will have to target gate-keeping purposes - i.e. to compare the aspirant's attributes or competencies with predetermined criteria and make a decision on his/her selection (Broadfoot, op. cit.: 32). Second, the organization has to determine the *boundaries* of the certification. The precise delimitation may differ from one country to another. For example in Flanders the concept of social interpreting is limited to social and public services, and is predominantly performed in the consecutive mode. The third characteristic concerns *resources* (Katz and Gartner 1988: 432), i.e. the number and quality of interpreters taking the certification exam, exam material, the number and qualification of graders. Finally, since social interpreting involves a service, a certifying agency has to get acquainted with the profile and needs of users, service providers and interpreters.

The *exchange* between the organization and its users and interpreters is materialized through testing procedures and certification. In our case, the exchange presupposes the implementation of testing that combines a construct (a profession) with content (its tasks), criteria (levels), and a cut score (standard). This exchange requires tools for specialized assessment, such as grids based on frames of reference for evaluation.

Decisions on test design and testing methods

The American Psychological Association (APA) (op.cit.: 9–11) stresses the importance of the threefold evidence needed for validity in testing. The 'holy trinity' (Guion, 1980) of content-construct-criteria should be reflected in course material and in later assessments.

Regarding content, test or certification developers have to decide whether to include knowledge tests in the certification exam, i.e. terminology or culture. Developers of interpreter certification exams know that a major part of the content consists of performances: speaking in different languages and applying interpreting techniques. This implies the design and development of performance-based tools and grids to measure specific indicators related to performance. It is perfectly possible to combine both types of content, knowledge and performance, in a certifying exam. When both concur in a test, the assessment developers must decide upon the weight of each of the contents as they arrive at a final decision about scoring. Usually, a weak performance on the skills part is eliminatory in interpreting exams, even if the knowledge part was excellent.

Regarding construct, test or certification, developers have to decide on the domain or mastery they want to test, and on which tasks they consequently have to focus. Regarding criteria, they have to decide upon guidelines for scoring and a standard (cut-off score) for passing.

Next, certification developers have to decide upon the specific assessment method to be applied. They have to decide how they will grade the performance: whether they will use a norm-referenced or a criterion-referenced approach. Norm-referenced testing is "based on the comparison of a test taker's performance to the performance of other people in the specified group" (APA op.cit.: 92). This definition entails that norm-referenced testing cannot serve the purpose of certification organizations, given that the evaluation of interpreting performances has to be based on external standards and cannot be derived from the results of individual members of a group. Some of the definitions stated by APA (op.cit.) regarding criterion, criterion-referenced test and cut score clearly show that criterion-referenced testing is more suited to grading interpreter performance. APA (op.cit.: 90) defines a criterion as “an indicator of the accepted value of outcome, such as grade point average, productivity rate, absenteeism rate, reject rate, and so forth. It is usually a standard against which a predictive measure is evaluated.” Furthermore, a criterion-referenced exam “allows graders to score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others” (APA op.cit.: 90). This method implies using a cut score, namely “a specified point on a score scale at or above which candidates pass or are accepted and below which candidates fail or are rejected” (APA op cit.: 90).

As stressed by Brown & Hudson (2002: 76) performance format can come close to eliciting actual, authentic communication, and consequently can predict future performance in real-life situations more validly.

The validity issue

The pioneering work of Cronbach and Messick broke with the tradition of an entirely cognitive and individualistic way of thinking about tests (McNamara & Roever 2006: 11). Messick’s (1989) unified theory of validity and his pioneering distinction between evidential validity (in support of interpretations) and consequential validity (which involves values and social impact) were especially important. Validity for Messick (op.cit.: 19) is “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment”. This stress on adequacy and appropriateness shows that Messick does not regard validity as a simple test feature, but rather as an argument for its effectiveness for a particular purpose (Brown & Hudson 2002: 240–241). Cronbach (1988, 1989) broadens this insight. He stresses that there is also a wider functional, political, economic context that decides on the validity of tests. A test should have not just functional but also political and economic value. A test must be made in such a way that it provides information to decision makers. Moreover, the cost of testing must be taken into account, i.e. costs should not be either too low or too high (Cronbach 1988: 5–12). Consequently, tests that do not account for political and economic value may jeopardize validity.

Validity is “the degree to which a test measures what it claims, or purports to be measuring” (Brown 1996: 231). For reasons of social accountability and legitimacy in all types of performance testing, and in the case of social interpreting in particular, there should be a confluence of several types of validity. Key criteria are content validity, criterion validity, and construct validity. Moreover alignment of evidential, consequential, and face validity is needed. Nitko (2001: 44) insists upon the recommendation by measurement specialists that validity be used as a unitary concept, and not as different kinds of validity.

Content-related evidence shows the extent to which the content of the domain of a test is appropriate to its intended purpose (APA op.cit.: 90). Content validation relies on expert judgment of the skills and knowledge measured by the tasks (Crocker 1997). A well-designed test will effectively measure the competencies it claims to test. When applied to social interpreting assessment, this implies both content validity at the level of specific competency clusters, and a *gestalt* content validity, expressing the relevance of the procedure in terms of the profession of social interpreting generally. Authenticity and meaningfulness (Linn 1991: 20) seem particularly important in the case of interpreting tests.

Criterion-related evidence shows the extent to which scores on a test are related to a criterion measure, i.e. a standard against which a predictive measure is evaluated (APA op.cit.: 90). An effective test scale should be able to differentiate between relevant levels of performance, but also needs to include a cut score or critical score level deciding on a pass or a fail (APA op cit.: 90). Brown & Hudson (2002: 253) insist that "when we talk about setting a standard, we are referring to setting that cut-point." In the case of interpreting, the different cut-points of the tests are established by the levels required by professional performance. An adequate choice of criteria and standards will be reflected in the overall test difficulty level. There should be a convergence between the criteria and standards set for different tests, both on types and on items.

The test construct is not directly observable; it is a conceptual framework. Construct-related evidence supports a proposed construct interpretation of scores of a test based on theoretical implications associated with the construct (APA op.cit.: 90). In other words, it is the measure of agreement between the test and the concept or domain it is derived from. In an educational context, this would be the relevant curriculum (knowledge and skills), in a professional context, the professional standard (related to specific knowledge and skills). As with the content and criterion validity, different sub-scales should be convergent. The issue here is to decide to what degree of detail the different knowledge and skills aspects should be dealt with in a specific test.

Predictive validity evidence refers to the extent to which individuals' future performance on a criterion can be predicted from their prior performance on an assessment instrument (Nitko 2001: 49). The issue in certification screenings is how well test items identify candidates that are potentially certifiable or to what extent individuals are excluded that could pass (see Stansfield & Hewitt 2005: 439).

Validity entails moreover that the certification exam, training curriculum and professional standard are developed in parallel, i.e. aligned (see Biggs 1999; Nitko 2001: 104).

The concept of consequential validity, which was introduced in Messick's unified model (1989) and further developed by Shepard (1997) and Linn (1997), stresses the value-bounded and particularly social nature of assessment (McNamara 2001: 335). Presently consequential validity is central to any interpretation of test scores, since it appraises the social impact of assessment. In certifying future service providers, this is probably the most critical concern. Will they perform adequately in real-life situations? A second matter of concern related to consequential validity and particularly social impact is what Lysaght & Altschuld (2000: 95) call *maintenance*. This concept essentially expresses the degree of persistence of the competencies measured in the assessment over a significant lapse of time and raises the question as to whether certification should be periodically renewed. On the one hand, Dreyfus and Dreyfus (1986, cited in Lysaght & Altschuld 2000: 97) describe a continuum of competency, from novice to expert. On the other hand, Fossum and Arvey (1990 cited by Lysaght & Altschuld 2000: 97) state that competency may be compromised by an eroding knowledge base, mental or physical maturation, changing attitudes and values, and environmental constraints. All professionals must be accountable for the on-going possession of a meaningful set of core competencies (Lysaght & Altschuld 2000: 98).

Following Gronlund (1985: 59–60), face validity is not really a type of evidential validity. Face validity refers to the appearance of the test. Based on a superficial examination of the items, does the test appear to be a reasonable measure? Even though this may seem spurious from a test construction point of view, this type of validity may be vital both for stakeholder accountability and legitimization and to facilitate transparency for candidates.

When applied to the gate-keeping performed by the graders, the relevant construct would be the cluster of competencies expected from the graders: the content would be their universe and the different tasks they have to perform, and the criterion would be the level of expertise we want to predict, including a cut score. The consequential validity would boil down to the question whether a grader will be able to identify competent interpreters. Finally, face validity would be the acceptance by candidates of their evaluation by the certification authority.

The reliability issue

Reliability is defined as "the degree to which test scores are consistent, dependable or repeatable, that is, the degree to which they are free of errors of measurement" (APA: 93). If they are not consistent, scores cannot be generalized beyond the sample of items or persons (APA: 91). The maxim that "Without reliability, there is no validity" quoted by Moss (1994: 6) is particularly relevant in performance assessment. Reliability of performance-based tests supposes a particular challenge (Moss 1994: 6). It pertains to form reliability, intra-rater and inter-rater reliability, test-retest reliability and subjectivity or bias. Form reliability concerns the internal consistency, i.e. the correlation among items or subtests (VandenBos 2007). Inter-rater reliability concerns the consistency of judgments made about people or objects between raters or sets of raters. Intra-rater reliability concerns the consistency of judgments made on people or objects by one rater. Finally test-retest reliability concerns the correlation among two or more occasions of measurement (VandenBos 2007).

Linn (1991: 17) warns against the misconception that through performance-based assessment possible bias, e.g. on race or ethnicity, can be avoided.

Graders face a complex task, not only because of cognitive challenges, but also because of the risk of emotional strain. More specifically, some problematic effects are: the significance effect (influence of another paradigm), the halo effect (when a judgment on a specific dimension is influenced by some other dimension), the sequence effect (lasting effect of a previous test taker), the contamination effect (influence of the grader's own agenda), the personal comparison (personal tendency to judge severely or in a compliant way) (Groot 1975 as cited in Dochy & Moerkerke 1995: 202), and impression management by the candidate (see Lievens & Peeters 2008). Under such conditions, it is difficult to be objective, i.e. "to establish judgments as true of false independently of personal feelings, beliefs and experiences" (VandenBos 2007). Finally, test-takers can also be affected by problematic effects, such as consistency over time (short or long term), over tasks, and psychological factors such as illness, fatigue, stress, emotional strain previous to the test, or exhaustion and breakdown during the test.

The feasibility issue

Finally, tests or assessments have to be feasible. Some of the tests cannot be organized because there are no graders available and factors such as space, time, infrastructure, funds, but also stress and fatigue have to be taken into account. Transgressions of feasibility easily deprive the credentialing authority of its legitimacy. Linn (1991: 20) also warns against the prohibitive costs of performance-based assessment.

Summary

To improve the validity and reliability of gate-keeping, there are several solutions. Niska's rule of "selecting those in need of the least training" remains fundamental, but is not sufficient in the case of the graders. Lievens (1998: 143) stresses that graders are limited in their role as information processors. After further research, Lievens (2001: 255–257) consequently concludes that the best way to promote accurate and consistent grading is the use of a frame of reference (specific norms and values espoused by the organization). The accuracy of the graders will increase, inter-rater reliability will improve and there will be more differentiation in all dimensions when graders are trained in the use of a frame of reference, i.e. particular norms or criteria, and analytical procedures for grader scoring. A clear mental framework and the use of checklists or rubrics can help to get graders into line. In his study of grader behavior in performance assessment, Eckes (2008:155–156) warns however against the fact that graders remain far from functioning interchangeably in spite of training. Graders may differ e.g. in their understanding and use of rating scales, rubrics, their interpretation of criteria and their degree of severity or leniency and consequently have a specific scoring focus or scoring profile (Eckes 2008: 177). Finally, feasibility also plays a part. It is definitely not easy to assemble all graders on one examination board; it may be more feasible to record the performances of the aspirants and let the graders do the evaluation work at home. This raises a new challenge to the accuracy and reliability of the graders.

We have now drafted an overall sociological framework that allows us to situate certification of social interpreters in a society founded on the activities or organizations that increase control, accountability, and legitimacy. Next, we will describe how the Flemish certification authority emerged, and how, by providing training and tests –with the help of external experts who act as graders- this authority reveals itself as accountable and attempts to gain legitimacy.

The certification of social interpreters in Flanders, Belgium

The current social interpreter certification exam and, more broadly, the qualification process for social interpreters in Flanders results from a multi-stakeholder process that includes social interpreting providers, university interpreting colleges, social interpreting services users, and social interpreters.[1] Comments and suggestions by these stakeholders were considered when developing and modifying training programs and test procedures for public service interpreters in Holland and the UK. The result of this process was a 90-hour training program (consisting

[1] Social interpreters are consulted by the COC throughout their training and testing process. This feedback is taken into account when the COC embarks on curriculum and test change and improvement. However, formal consultation of the social interpreter community as such does not take place as there is no official representative body for social interpreters in existence.

of an 18 hour introductory course and a 72-hour basic interpreting training program) which has been expanded to a 102-hour program, and a certification exam. The certification exam and training program were jointly developed.

The COC and interpreter service providers opted for training bilingual candidates, who often already were experienced interpreters but who had not yet received (sufficient) training.[2] The curriculum focuses on training interpreting skills and providing social interpreters with information, practice and discussions on the code of ethics and the contexts in which social interpreters usually work. Upon finishing this training, the interpreters take the certification exam. The pilot year taught the COC that in a significant number of cases Dutch proficiency was inadequate to pass the exam and even to fully comprehend the training provided. The curriculum was reviewed and currently consists of the following elements: a Dutch proficiency admission test, an 18-hour introductory course, an interpreting aptitude test, an 84-hour basic interpreting training module, a certification exam, and a 21-hour remedial training module for candidates who have failed the test.

The certification exam procedure

Developing the certification exam

Flanders has opted for performance assessment, testing competencies relevant to interpreter performance, unlike Sweden (www.kammarkollegiet.se/tolktrans/tolkauk.html; Idh 2007: 135–138) or the United Kingdom (www.iol.org.uk), whose interpreter tests not only assess interpreting performance, but also cultural and terminological knowledge.

The objective of the certification exam is to have relevant interpreting skills tested as objectively as possible by a professional examination board consisting of three persons: a chairperson (a representative of the social interpreting providers or the COC); a grader of Dutch and of interpreting techniques (a representative of one of the Flemish university interpreting colleges); and a foreign-language expert.[3]

[2] The development of a training program for the whole of Flanders by the COC does not mean social interpreters did not receive any training previously. Most social interpreting service providers provided short trainings and crash courses for their interpreters.

[3] Foreign language experts function in a variety of professions in their daily life. The COC selects them on the basis of set criteria and provides them with a basic on the job training consisting of a one day observation of an experienced examination board. In addition, the new foreign language expert can exchange with the graders on the examination board and he/she

The current certification exam is the result of a seesaw process. The stakeholders provided information on competencies and skills levels they considered a requirement for a professional social interpreter. These suggestions were then compared to test procedures by foreign counterparts such as the Dutch and the British certification agencies, and to theoretical test development frameworks.

The certification exam consists of the following tests: Dutch proficiency, other language proficiency, reproduction, transfer, and role play. The procedure, contents, and grading method will be described below. An example of an evaluation grid is included as an appendix to this chapter (See Appendix 1).

Language proficiency tests

Both the *Dutch language proficiency test* and the *other language proficiency test* are evaluated with the same standard. As a basis for the standard, both the norms of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) level B2, and the criterion *voice* were used.

Levels in the CEFR are defined by "Can-do" statements which "define levels of ability in terms of what language users can typically do at each level of the Framework, [and] make it easier for users to understand what each level means in relation to what language users actually do." (See also www.alte.org) The Can-do system comprises approximately 400 statements, sub-divided into 40 categories, which describe what typical language users can do in a particular language, at a particular level and in one of the skill areas (listening/speaking/writing/reading).

The language proficiency test in the certification exam for social interpreters only measures oral skills (listening and speaking). For the B2 level these oral skills are: participate with native speakers in conversations on general topics; understand native speakers without any problem when they are speaking the standard language; participate in a conversation with a native speaker; be able to state opin-

will first work in a team consisting of the most experienced graders. Foreign language experts are fluent in Dutch; and they are native speakers of the particular foreign language, or have at least native speaker command (e.g. a foreign language expert of Chechen also functions as a foreign language expert Russian as she received her secondary and university training in Russian and Russian is an official language in Chechnya; A foreign language expert of Russian or Spanish can be recruited from one of the universities or university colleges); they have a degree in linguistics, interpreting or communication sciences and have experience as language teachers and/or graders. For some more exotic languages, these criteria are hard to meet. The COC will then recruit candidates who receive on the job training before functioning on an examination board. When no foreign language experts meet the criteria, no certification tests for this particular language will be organized, but the COC will continue recruiting in order to meet this need as soon as possible.

ions and ideas when discussing a topic with native speakers; be able to present a topic clearly and be able to define cause and effect and pros and cons; be able to exchange complex information on his/her own field of work; and, overall, contacts with native speakers should be of such a nature that native speakers regard these as natural and correct. These Can-do statements require a mastery of grammar and vocabulary at a particular level. Pronunciation, accent, intonation and pace should be familiar to native speakers rather than distracting. A candidate also has to be able to speak loud and clearly (audibility).

A candidate can be graded a *fail*, a *pass*, or *excellent*. To pass the language proficiency test the candidate's oral language usage has to meet the criteria for the B2 language level. The weighing of the sub-criteria contained in *voice* is somewhat more complicated in that none of the sub-criteria should appear distracting to native speakers. To obtain the category *excellent*, a candidate masters a higher level than the B2 level and passes all sub-criteria within *voice*. The grader of Dutch and of interpreting techniques sets the final mark for Dutch and the foreign-language expert for the other language. Although some certification exams for interpreters do not contain separate language skills tests, the COC grades these because they are a prerequisite for interpreting skills. Furthermore, a candidate who fails receives a diagnosis of his/her performance and advice for further training. When a *fail* is due to a particular linguistic problem, the candidate will receive clear feedback to enable him/her to remedy his/her shortcomings. Finally, in the course of the certification exam, the complexity of tasks increases to allow candidates to gradually build up to the final role play performance.

Reproduction

The next test is *reproduction*. This test evaluates listening comprehension, note-taking, memory, and consecutive reproduction skills in Dutch, but not translation skills. The grader of Dutch and interpreting techniques reads out in spontaneous speech an informative text of about 200 words on a subject relevant for the context in which social interpreters work, e.g. breast feeding, child abuse, registering for an allowance, etc. Texts are based on real life material from social services adapted to the needs of the test, and the source material is directed at the general public, not specialists such as nurses, doctors, or lawyers. All texts contain approximately the same amount of words (200), numerical references (a date and a telephone number), textual and logical links, names (person, organization), and an enumeration of 5 units (e.g. five symptoms of a medical condition). While the grader reads out the text, the candidate is allowed to take notes. Next, the candidate can quickly look through his/her notes and ask two extra questions about a particular passage that was not clear. The phrase containing the unclear item will be repeated. However, the text will not be re-read. The grid for this test consists of two parts: on the left hand side of the grid the content of the information rendered is evaluated, and on the right hand side of the grid the quality of the reproduced message is evaluated. The criteria used here are: cohesion (are the sentences produced correctly or not?); coherence (is the logical structure of the original message retained?); completeness and presentation; additional information; and distortion of the message. A candidate can be awarded points or subtracted points according to his/her overall performance on these criteria. If a candidate uses correct idiomatic sentences he/she will be rewarded 5 extra points for cohesion. When his/her performance is substandard he/she might- receive up to 5 minus-points on this criterion, depending on his/her performance. The same principle holds for coherence: if the structure and logical coherence of the text is fully respected a candidate receives 5 extra points. When this is not the case, he/she might receive up to 5 minuses, depending on his/her performance. When a candidate performs well on completeness and presentation he/she can be rewarded 1 to 5 extra points. For the sub-criteria, additional information and distorting the message 1 to 5 points might be subtracted, according to the effect on the meaning of the original message by the information added or distorted. The extra points or minus points are added to the grades for the content rendered. The minimum score for a pass is 35. The grader of Dutch and of interpreting techniques sets the final mark. There is no *excellent* category for this test. However, when a candidate fails, the grader of Dutch and of interpreting techniques may decide – after careful deliberation – to award a deliberated pass to the candidate. In order to be considered for a deliberation, a candidate should have at least 30 points in total, should not have received any *minus 5*, and should have passed all other tests of the certification exam.

Transfer

Whereas *reproduction* evaluates purely consecutive reproduction skills, *transfer* tests language transfer abilities by means of a sight translation from Dutch into the other language when the candidate is a Dutch native speaker, and from the other language into Dutch when the candidate is not a Dutch native speaker. Translating into a foreign language is more difficult than translating into your mother language. Therefore, we have opted for this method as it offers more certainty with regard to the candidate's level of transfer skills. Again, text material for this test is taken from authentic, real life material. Texts consist of a passage of about 100 words from a newspaper article of general interest. Contrary to the text material for reproduction, texts are not adapted in any way. The candidate is allowed to scan through the text before starting the sight translation. This naturally requires reading skills. However, not all candidates are able to read because of physical (e.g. interpreters with a visual impairment) or linguistic barriers (e.g. because the language tested is either not written or the writing system is not mastered by the population, which happens to be the case of Berber languages from Northern Africa). In these cases, the grader of the other language will read out the text completely and will then repeat the text sentence per sentence in order to enable the candidate to translate the whole. The candidate's performance is evaluated according to the criteria translation ability, structure, and vocabulary. Translation ability contains the following sub-criteria: *faithfulness* (the message in the target language is the same as the message in the source language); *accuracy* (the intention of the original message is retained); *completeness* (no units of meaning have been omitted); *fluency* (no pauses or repetitions that cause irritation with the listeners); and *pace* (considered acceptable and normal by native speakers). Moreover, production in the target language has to be such that native speakers would accept it as correct language usage. *Structure* is evaluated through the coherence and logical structure of the translation rendered. *Vocabulary* is split up into general vocabulary and terminology (http://www.serv.be/Publicaties/1280.pdf). The other language expert sets the mark: a candidate can obtain a fail, pass, or excellent. When a candidate has failed the test, the grader of the other language may decide – after careful deliberation – to award a deliberated pass to the candidate. In order to be considered for a deliberation, a candidate cannot receive more than one fail on the criterion translation ability, structure, and vocabulary, and has to pass all other tests. Experts of the foreign language are above all experts in this language. Not all of them are familiar with translation and interpreting. The COC provides them with on-the-job training, and during the evaluation the chairperson of the examination board will ask specific questions to determine whether a sub-criterion has been met or not. Nevertheless, this fact might be the cause of tension which will be discussed below (see grader competency profile).

Role play

The final test of the certification exam is *role play*. This test comprises all skills tested in the previous tests: language proficiency skills, reproduction skills, and translation skills. In addition, it also tests interpreting skills and adherence to the code of ethics. The role play is situated in a real life setting. Scenarios for the role play are based on transcripts of real-life interpreting situations, but adapted to the needs of the test. Scenarios should be from one of the following social settings:

infant and child care, health care, civic integration, asylum, residency issues, mental health, counseling, social welfare, and education. All scenarios are approximately of the same length (1200 words), they contain the same amount of lexical problems (15 general languages usage terms, 15 words typical of terminology of the social context, 3 acronyms typical of the social context and 2 idiomatic expressions) which should be translated correctly. The role play actors – a Dutch native speaker acting, for example, as a social worker or a nurse, and a native speaker for the other language acting as a client in a social service – receive a scenario script, and are required to adhere to it as much as possible. However, they should do this in a natural and spontaneous way so as to simulate the circumstances of a real conversation as much as possible. This also entails that the role play actors have to improvise when an interpreter commits a translation error: e.g. in one role play a woman who is divorcing her violent husband says "my lawyer doesn't want me to attend the hearings in court because last time I lost my temper, because my husband and his lawyer speak [language] and I could understand what they said and this angered me tremendously …". A candidate translated this as "my husband is not allowed to go to the court hearings anymore because last time I lost my temper because my husband and his lawyer speak [language] and I …" The Dutch role play actor reacts by saying "Oh really? Is your husband not allowed to attend the court hearing?" All scenarios contain 6 instances where the code of ethics is violated, three times by the Dutch native speaker and three times by the native speaker of the other language.

The candidate's performance is evaluated according to the following criteria: correctness/accuracy of the interpreted rendition, faithfulness, interpreter attitude, assertiveness, fluency, management of the triadic relation, and attitude with regard to the code of ethics. These criteria are further defined. *Correctness/accuracy* of the interpreted rendition means the word use in the target language reflects the word use in the source language, and is evaluated according to the following sub-criteria: additions, omissions, completeness, correct transfer, and structure. *Faithfulness* means the interpreter translates expressions from the source language into the target language in a way that is considered acceptable and normal by native speakers. *Faithfulness* contains the sub-criteria register, style, nuances and empathy. The interpreter attitude is defined in the *standard introduction* by the interpreter before each interpreting performance,[4] by the fact

[4] The interpreter should state before starting to interpret in both languages: 'I am an interpreter of languages X and Y and I will translate everything you say, without adding, omitting or altering anything to your words. I am neutral and I am not allowed to take sides. Furthermore, I am bound to professional secrecy. Lastly, I will interpret in the 'I'-form, which means you can speak directly to the other person just as if I were not here.

that the interpreter does or does not respect the correct seating arrangements (triadic position), steadfastness in the interpreter role, and the use of the first person. *Assertiveness* means the interpreter asks for a clarification when a word or passage is not understood, when he/she is unable to translate a word, or asks for repetition when necessary, or asks for a pause in the discourse when the message becomes too long to be conveyed completely and accurately in the other language. *Fluency* is defined by the degree to which the pace of speech when switching from one language to another is acceptable for native speakers. *Managing the triadic relation* means that the interpreter repeats the relevant passage of the code of ethics when it has been violated by one of the role play actors. The criterion *ethical attitude* contains neutrality, avoidance of private chats with either role play actor, complete transparency, not voicing one's own opinion, personal interpretation and emotional involvement (See http://www.serv.be/Publicaties/1280.pdf).

A candidate has to pass all sub-criteria to obtain a pass for this test. However, a minor weak point might be compensated for by other strong points. For example, an interpreter might do an incomplete *standard introduction,* but still passes because he/she has stuck to the correct way of interpreting and behaving throughout the role play. When the candidate's performance exceeds the expectation, he/she will receive *excellent*. The result is reached through careful deliberation by members of the examination board. When the graders cannot agree upon the final result for the role play, the chairperson has the prerogative to decide which grade the candidate will receive for the role play.

Summary evaluation grid

When all tests have been graded, a summary evaluation grid will be filled in by the members of the examination board. It contains the results for all the tests and the final result. A candidate has to pass all tests to obtain the social interpreter certification, or pass all the tests and have a *deliberated pass* on *reproduction* and/or *transfer*. To obtain *excellent* as a final result, the candidate must have obtained a *pass* on *reproduction* and *excellent* on all other tests. When a candidate fails the exam, the examination board has to provide an explanation of the shortcomings of the candidate. Moreover, it is expected that the certification board include a note of advice for the candidate whenever possible.

Measures to ensure a fair assessment

In the case of the Flemish certificate, both the training curriculum and the certification test procedure were derived from an earlier version of the professional standard and were developed together, thus enhancing validity through alignment.

In addition, there are a number of built-in checks to limit subjectivity to a minimum. The certification exams are constructed in a balanced way, spreading the difficulties equally across the different tests included in the exam, and across every single test. For example, in the role play, all scenarios contain the same number of ethical problems and terminological difficulties. Furthermore, graders evaluate the candidate's performance in accordance with criteria set out on a standard evaluation grid, reducing the subjectivity by introducing anchor points in the evaluation process.

In addition, the jury approach allows for the management of individual subjectivity through inter-subjectivity. For any statement to be (scientifically) objective, it must be inter-subjectively testable (Popper 1934: 24–25). The fact that graders have to discuss their observations and assessments in a team, allows the other team members to invalidate subjective influences.

Graders work as members of a team enabling them to broaden their own view on the candidate's performance, and to prevent potential errors in evaluation. As a result of this, inter-rater reliability will increase. In addition to inter-subjectivity, a thorough training of graders is a must. Training of graders will be discussed below.

Finally, a candidate may appeal to a commission when he/she does not agree with his/her test result. The appeals procedure is free of charge. The candidate is first invited by the COC to view the exam tape and file and to discuss the exam. When the candidate still disagrees with the result, his/her exam or one or more tests of the exam will be reviewed. The COC will then convene a commission comprised of a representative of the COC, a representative of one of the social interpreting services, and a representative of one of the university interpreting colleges. The commission will review the exam both with regard to the procedure and to the grading. Wherever the commission feels the need to hear one of the parties (candidate, members of the examination board) its members will invite that party to a hearing. The commission may also decide to refer the file to an external[5] expert for re-grading. This might be the case when there is an issue concerning the other language. The decision of the commission is final.

[5] An external expert is called 'external' in the sense that he/she was not involved in the grading process of the particular exam he/she is asked to *re-grade*, and that he/she is not a member of the commission that reviews the exam.

The professional standard for social interpreters

The Professional Competency Profile for Social Interpreters published by the Social Economic Council of Flanders (www.serv.be) in December 2007 defines all competencies that a professional social interpreter should master in order to work in a reliable and effective way. This profile contains competencies describing interpreting skills, such as processing oral messages, reproducing oral messages, complying with the ethical code, and dealing with ethical conflict situations. More general skills required in other professions are included as well, such as time management and accounting. (Meyers & Houssemand 2006: 124). Furthermore, this profile provides the foundation for the Standard for the Social Interpreter. A standard defines the competencies that can be objectively measured in a test procedure: processing oral messages, reproducing oral messages, complying with the ethical code, and dealing with ethical conflict situations. These competencies are already tested in the current certification exam. However, the standard not only defines the competencies to be measured, but also outlines a test procedure and requirements for quality control that have to be met. This is done to guarantee a fair test procedure for all candidates, including complex competencies such as processing and reproducing oral messages. The Standard for the Social Interpreter will influence the test procedure as all criteria and sub-criteria will have to contain clear and unambiguous descriptors, leaving far less room for subjectivity. In addition, each (sub-)criterion will have to be tested at least twice and it will have to be observed and graded by two graders simultaneously, thus allowing for triangulation in all tests.

Implications and limitations

The assessment procedure described above poses several challenges: legitimacy issues, consequential validity issues, the effect of random events, processing numerous indicators in relatively short laps of time, and subjectivity issues. We will expand on each challenge, and we will try to formulate possible answers to these challenges.

Limitations

Although it is an important factor in its own right, conformity with the new professional standard for social interpreters is not the only driving force for improvement of the current assessment procedures. Day-to-day practice and feedback by graders and candidates have also brought to light a number of challenges that ought to be addressed in future versions of the assessment procedure.

Internal sources of legitimacy
The overall level of internal legitimacy is determined by three critical success factors: the procedure's validity, its reliability, and the graders' reliability. Since no systematic validity or reliability research has been conducted, the procedure's main claim to validity is based on the systematic alignment that has been used in its conception and construction.

The evaluation grid is completely congruent with the competency-based professional standard. Following the publication of the standard, this has become a formal requirement for the accreditation of the assessment center.

As is evident in Figure 1, the standard describes each specific competency in terms of a general definition, indicators, required knowledge, required attitudes, and general competencies.

However, more exhaustive descriptors and a uniform weighing system still need to be developed.

Adding new test components might improve overall reliability, but is not an option in the light of budget constraints, and cannot be introduced without also changing the official standard.

Language competency of foreign-language graders
Graders are not only the users of the procedure; they are, in a very real sense, part of the measurement apparatus, providing it with eyes, ears, and decision making capability. As such, they are an important success factor. Improved training for graders would seem to be a highly feasible option to enhance overall reliability. However, even this aspect is not without its challenges.

One of the main problems is that foreign-language graders are selected on the basis of their qualification and portfolio. The COC staff has to accept most of these competencies at face value because they cannot be expected to be proficient in every language they certify. Nevertheless, there are checks and balances: by introducing a second foreign-language grader, a form of mutual control is introduced, and, in addition, the candidate has the option of turning to the portfolio commission whenever he/she has reasons to doubt the competency of one of the graders.

Consequential validity issues
Social interpreting is a profession with a major impact on the individual well-being of clients and the quality and accessibility of the social care they receive. Social and medical professionals and their clientele are ill-prepared to judge the quality of interpreting. For reasons of patient or client confidentiality, third parties (observers, researchers, or peers) are generally not admitted, even for formative evaluation purposes. Most social interpreters are not in-house staff members but freelancers sent out by a social interpreting agency. In a social context where,

Description:

The social interpreter transfers the oral messages of both triadic parties into the required target language. The interpreter reconstructs the message as completely, correctly and clearly as possible, by means of his/her memory or of consecutive note-taking.

The transfer of oral messages into another language should be done in such a way as to minimize the difference between an interpreted conversation and one where no interpreting would be necessary.

Indicators:

The social interpreter:

- Uses the consecutive interpreting mode of his/her own accord.
- Articulates clearly.
- Speaks clear and loud.
- Sustains an acceptable pace of speech when switching languages.
- Uses the target language's intonation pattern.
- Asks the speaker for clarification if a term is not known to her/ him or whenever s/he is uncertain about the correct translation or paraphrase.
- Either sticks as closely as possible to the original wording of the source message or conveys the intent of the message without loss of meaning.
- Transfers language-specific expressions and constructions in the source language into expressions and constructions that are accepted as correct and natural by users of the target language.
- Conveys the tone and attitudes that are indicated by the (non-)verbal communication of the triadic party.
- Either interprets messages with a negative or insulting content or explains them without any loss of transparency of the message.

Underlying knowledge:

- Consecutive interpreting
- Voice and communication techniques
- Dynamics and complexity of communication
- European reference frame B2 level in the foreign language: speaking
- European reference frame B2 level in Dutch: speaking

Underlying attitudes and key competencies:·

- Verbal communication
- Accuracy
- initiative

Figure 1. Reproduction of oral messages

more often than not, the use of an interpreter still has to be advocated, agreements to systematically allow observers for scientific, high-quality monitoring or training purposes remain few and far between. In such a context, the certificate is quite often the only readily accessible quality label. The label, in turn, is backed by claims of external and internal legitimacy. The former being rather more important for professionals in the social sector, whereas the latter is of greater importance within the interpreting profession. A major external source of legitimacy resides in the continuous involvement of and backing by a number of relevant stakeholders (the regional authorities, the interpreting service providers, the interpreting colleges, a number of major end-users, and the interpreting community). In the current state of affairs, the client's perspective is absent from that multi–stakeholder dialogue, mainly because they are not organized as a group. Another important external check in terms of consequential validity is the degree of congruence with foreign certification procedures. In the Flemish example, two neighboring countries, the United Kingdom and the Netherlands, are the main reference for certification and registering. These countries have comparable structures, but enjoy an advantage of more than a decade in regulating and promoting community interpreting. From the outset, the Flemish model has been influenced by these examples.

Even though it has been argued that the other triadic partners cannot fully evaluate the interpreting performance, there are a number of cues in the job performance that clearly differentiate the professional social interpreter from ad hoc interpreters. Generally speaking, professionals and clients are aware of the importance of respect for the code of conduct, as well as of behaviors such as a *standard introduction* or consecutive note-taking, and tend to value these aspects.

The effect of random events

No procedure can ever be immune to incidents and accidents. Generally speaking, robust procedures and experienced and well-trained graders will suffice to compensate for most of these incidents. However, practice has shown that, over time, procedures have a tendency to become more intricate and complex and, thus, more accident-prone.

Although, as yet, this has not given cause to any formal complaints, on several occasions the assessment procedures have not been correctly recorded. Generally, this is due to human error. The chairperson may forget to start up the audio or video recording at all, or may compromise the recording quality by setting up the camera at a wrong angle or by placing the MP3-recording device to far from the candidate. In other cases, recordings were incomplete due to technical issues.

The human factor also plays an important part in the management of role plays. On several occasions, the actor playing the part of the client failed to show up in time for the exam. In these circumstances, the foreign-language grader is asked to play the part assigned to the actor. This is extremely challenging for the grader as he/she will have to simultaneously focus on acting and assessment.

Another problem is that some actors start to improvise. They either leave the bounds of the scenario (thus skipping predetermined behavioral and lexical cues) or overact the role, for example through shouting or by becoming very emotional. In one specific case, this prompted the jury to stop the role play and to restart the procedure with another role play sequence.

Issues regarding graders

The current test includes a large number of criteria, sub-skills and attitudes that have to be tested in a 1.5 hour span. The assessment is conducted in real time, generally without any replay of taped performances, and is documented and supported by the use of evaluation grids at criterion and sub-criterion level. Inexperienced graders tend to find it difficult to navigate these grids while observing at the same time. Their efficiency could be greatly improved by further developing the criteria into a set of descriptors of specific behavioral indicators. The use of these descriptors could then also be included in grader training or coaching. There has as yet, been no research into the validity and inter-rater or inter-jury reliability of the assessment procedure. There we find that pragmatic issues clash with technical requirements. The funding agencies for this procedure generally prefer to see their means invested in an increased output of certified interpreters, rather than in scientific research and quality assessment.

Subjectivity

As a rule, preference is given to graders with prior professional experience in grading or assessment. This policy, however, also has some weaknesses. Graders may be too much influenced by other frameworks (e.g. as teachers of conference interpreting or foreign languages), and this influence may color their ratings. Some graders may even want to ignore the analytic framework provided for the test, and produce a holistic evaluation. Neophyte graders, on the other hand, may lack the required grading skills. Again, grader training and coaching, as well as continuous monitoring by the chairperson of the adherence to the test protocol constitute important mechanisms to ensure the fairness of the process. Systematic adherence to the test procedure may also help to minimize the impact of psychological effects, such as halo-effects, stereotypes or prejudices, against which even experienced graders are not immune. A second way of addressing these issues is by ensuring inter-subjectivity through triangulation to balance out individual biases.

Another source of subjectivity sometimes observed in this type of examination and calling for jury-based assessment is that graders may want to express views and opinions concerning criteria beyond their specific assignment or rubric (e.g. a language grader influencing the final score for consecutive note-taking).

This requires constant monitoring and immediate correction by the chairperson. In these cases, he/she refers to the procedure and the standard.

Implications

Backstopping and wash-back on teaching through remedial training

Interpreters who fail the certification exam are required to take a 21-hour remedial training before taking the exam a second time. In this course, the recorded test material is used for corrective training purposes. The remedial training course is provided by staff trainers to groups of 6 to 8 candidates. The recordings of the reproduction, transfer, and role play tests are analyzed with the group and used as a basis for corrective and systematic training.

This training course is primarily meant for backstopping purposes, and thus aims at capitalizing on the assessment's wash-back effect on trainee performance by reinforcing required behavior.

However, since the course is managed by in-house trainers, there is also an opportunity for a wash-back effect on teaching. A good example of this is a new teaching policy that is currently under development because of a case of group superstition learning that was corrected through remedial training. Several remedial students who all had attended the same basic training were convinced that they should at all times avoid eye contact with the other triadic parties in order to maintain and demonstrate their neutrality as an interpreter. This particular belief was based on one single example from a mental health setting given by one of our interpreting trainers. The trainer had explained how a particular therapy had greatly improved by her refusing eye contact with the patient and the therapist. The student group had later on spontaneously generalized the example into "good practice." As a result, the COC is now developing a set of guidelines for the didactical use of examples, confining more complex examples to specific basic training classes such as "role definition" and "ethical dilemmas" or to post-graduate training.

Triangulation

An important element in compensating for limited grader training is to ensure inter-grader-triangulation: every criterion should be observed by more than one grader. In addition, it is standard practice to pair up novice graders with experienced ones, both as a learning opportunity and as an additional check during deliberation.

Because of the large number of languages that have to be assessed, usually only the foreign-language grader(s) will know both languages. This means that for a full appreciation of a candidate's role play performance, the other board members have to rely on their narrative recount and evaluation. Currently, the foreign-language graders are systematically queried by the chairperson, but a specific training of the foreign-language graders in this role would be useful.

An additional way of addressing this challenge is to ensure intra-grader triangulation and intra-rater reliability by repeatedly testing for every criterion. In the authors' view triangulation contributes to the test procedure's robustness by making it less incident or error prone, thus improving the assessment's general reliability.

Up to a certain extent, the graders' intra-rater reliability can be monitored through consistency indexes. Averaged out over a number of evaluations to allow for individual candidates' variation, these indexes show how consistent a grader is in his/her evaluation of the same criterion in different observations within the same assessment.

Graders can also be trained to improve their intra-grader reliability by introducing systematic test-retest exercises in grader training. In these exercises graders repeatedly grade the same taped performances until they attain a certain level of consistency.

One of the challenges for the future development of the certification procedure is the management of an expanding team of graders. During the procedure's first development phase, general guidelines, a code of ethics, and a limited coaching program were developed for graders. In order to develop a more comprehensive training program and to fine-tune the current selection criteria for graders, we need a specific grader competency profile.

In the following section, we will outline proposed criteria for such a profile.

Professional competency profile for graders

To start a dialogue about grader profile, we will discuss the following competencies below: (1) evaluate interpreter performance according to the method presented above; (2) act as a professional and loyal member in a team of graders; (3) act in compliance with the code of ethics for graders; (4) manage ethical conflict situations; (5) plan and organize; and (6) develop personal professional competencies. We will not go into competencies related to knowledge of social interpreting and linguistic competencies, as we consider these to constitute a prerequisite for being appointed to a team of graders. These competencies and behavior guidelines are part of the grader trainer manual and the ethical code for graders.

Grading interpreter performance according to the method presented above

Graders have to evaluate the interpreter's performance according to criteria defined in the professional standard for social interpreters mentioned above. Moreover, they have to maintain the same level of quality of grading throughout the successive tests of the exam. In the case of the grader of Dutch and of interpreting techniques this means evaluating oral Dutch during the Dutch oral proficiency test, evaluating the reproduction of a Dutch text into Dutch, and evaluating the role play. The grader of the other language will evaluate oral proficiency for this language, sight translation, and the role play. Throughout the exam, the complexity of tasks for the interpreter increases, but so does the complexity of grading. This means the grader not only has to be aware of this increasing complexity, but also has to deal with grading of complex performance. For example, during the Dutch oral proficiency test attention is paid to correct and comprehensible production of Dutch. During the reproduction test, however, the interpreter has to be able to give a fairly correct and comprehensible rendering of an oral message of about 200 words. Here, the grader does not only grade Dutch on a communicative level, but also the correct and complete rendering of an oral message the interpreter has heard. Furthermore, when the grader is grading the performance during role play all indicators mentioned above are being evaluated, along with appropriate interpreting techniques and attitude. This complex set of competencies require from the grader extreme concentration throughout the certification exam.

In order to grade the interpreter's performance thoroughly, firstly a grader has to understand the design and the objective of the tests used. Secondly, he/she will grade interpreter performances according to the linguistic, cognitive, interpreter, ethical, and socio-cultural indicators set forth on the grading sheets. This also implies the grader has to be able to use grading sheets and indicate weak and strong points of performance on them, as well as specific examples of performance. The grader is also expected to comment on these examples in a way that is comprehensible to a candidate.

Tension may arise during the grading of role play, as the grader of Dutch and of interpreting performance usually does not understand the other language, and the grader of the other language, in turn, is less familiar with grading interpreting performances. This tension is dealt with, on the one hand, by coaching and training graders (before they sit on an examination board), and on the other hand, by having the chairperson of the board closely involved in test development. In this way, the chairperson is able to monitor the correct procedures of grading, explain test items, and query graders' opinions.

Acting as a professional and loyal member of a team of graders

To ensure the objectivity and fairness of the procedure, each member of the team has a clearly defined role in grading the performance that has to be respected by the other members: the grader of Dutch and of interpreting techniques will be responsible for oral proficiency in Dutch and reproduction, the grader of the other language for oral proficiency in the other language and for sight translation. During the role play the team acts together. This, however, might bring about the above-mentioned tension. This tension is largely resolved in the way that was explained above. In addition, when both graders do not agree on the performance during role play, the chairperson has the final say.

A grader has to strike the right balance between his /her point of view and that of other team members. He/she does not force his/her opinions upon others, nor is he/she susceptible to manipulation by others. In practice, this means each member has to respect both seating arrangements in the examination room as well as follow the guidelines in the scoring rubric. The chairperson has to possess the necessary leadership and listening skills to manage the team according to the procedure.

Graders are not tested in any way to evaluate their ability to function in a team, but they are thoroughly briefed on the procedure and the code of ethics (discussed below).

In general, graders are not staff members of the organization, but they do collaborate on a regular basis, both with the organization and other graders. They are loyal, reliable and efficient team members, who will signal any possible problems or conflicts in time to find suitable solutions. They only accept assignments that are within their field of competency, and they give the organizer notice in good time if they are unable to attend the assessment session. This means the grader needs to find a balance between his/her role as a grader and his/her other activities. Quite often, these other activities (e.g. as a linguist or teacher) contribute to his/her knowledge and competency as a grader.

Acting according to the graders' code of ethics

The graders' code of ethics regulates graders' behavior before, during, and after the examination. Each grader has to sign the code of ethics before becoming a member of an examination board. Graders limit private contacts with candidates as much as possible. Furthermore, graders cannot practice as social interpreters at the same time. Candidates can challenge a member of the examination board if this is the case, or if they have any professional and/or private relationship to avoid conflict of interest. Graders are free to voice their opinions during the deliberation process, but once consensus is reached they will not divulge their private opinions on the deliberation to the outside world. Graders will also act in a neutral and impartial way towards all candidates and will not discuss any private, political or religious opinions a candidate might have. Members of the board do not divulge exam material.

Graders are not led by sympathy or antipathy towards the chairperson, fellow graders, or candidates. They remain unbiased by any feelings they may have concerning ethnicity, nationality, group membership, gender, language, culture, age, or background.

Managing ethical conflict situations

Graders are aware of ethical boundaries. They monitor events in such way that neither they, nor the chairperson, their fellow graders, or the organizer are discredited. They share responsibility for the image of the organizer and the entire profession. This implies that they respects the rules at all times, and they have to inform the organization whenever other parties do not. Therefore when faced with situations that challenge the ethical boundaries of the examination board, a grader states or repeats the rules set by the code of ethics. Whatever happens, a grader demonstrates respect toward the chairperson, the fellow graders, and candidate. He/she withdraws from the examination board when a conflict of interest arises after stating the nature of the conflict that has arisen. Moreover, graders keep all contacts and conversations with candidate and third parties strictly neutral and they do not accept gifts from candidates or third parties. Whenever a verdict by an examination board is challenged, the grader is expected to defend the board's decision in public and versus the candidate. The grader can refer the candidate to the COC for a review of the exam. The chairperson has an additional role here, in that he/she monitors the procedures, prevents ethical conflict situations, and when these do occur, intervenes by stating and clarifying the ethical code.

Planning and organizing

Graders plan their agenda so as to be present (on time) when exams are scheduled. Graders of the other language are also expected to prepare a sight translation and hand this in for review a few days before an exam takes place.

During the exam, graders have to keep to the time allotted to each separate test. However, the chairperson has an additional role in this; he/she will brief graders on what is expected of them and he/she will manage time throughout the whole procedure. The chairperson informs the candidate of the procedure before the exam starts and repeats instructions for each test at the beginning of the test.

Continuous professional development

In this program, graders are required to continuously develop their expertise in languages, interpreting techniques and (criterion-referenced) evaluation methods. Continuous professional development implies a dedicated investment of time, even when assessment for the social interpreter certification exams may only be an occasional activity to certain graders. Graders develop their expertise through feedback and training by the COC, as well as through self-study and training by other organizations. This particular competency implies that graders be open-minded about feedback on their performance by others, and that they are able to reflect on their own performance during tests. If distinct types of graders can be identified in the context of interpreting assessment, specific training can redirect their perceived importance of criteria towards a common agreement. In addition to the ability to develop graders' skills, the chairperson also possesses management and leadership skills.

Conclusion

As we have discussed in the introduction, consequential validity issues are a driving force in societal pressure for efficiency and accountability of the assessment procedure. As a government funded agency, the COC has to legitimize its certification model on the basis of internal and external determiners of legitimacy.

The Flemish Central Support Cell's (COC) model of certification is presented from the candidate's and the graders' perspective. The model is based on a human resources assessment approach. In other words, all testing is carried out by specialists and not by peers. These specialists act together on an examination board, not individually. The chosen type of measurement is criterion-referenced testing at a basic level of interpreting proficiency. This implies that the final evaluation is analytic in nature rather than holistic. This human resources approach is only possible because of the COC's ongoing partnership with the Flemish university interpreting colleges that provide language and interpreting teachers as graders. However, for the more exotic languages, it is far more difficult to find qualified graders.

To ensure the reliability and validity, but also the perceived fairness of the assessment, the test procedure introduces a number of objectifying elements such as criterion-based evaluation grids, guidelines for scoring, pre-determined cut-off scores, and triangulation.

There is a permanent need for improvement of test materials and procedures. Future developments to reinforce factors of internal legitimacy are likely to include: conducting validity and reliability research, improving evaluation grids by developing indicators into weighted descriptors, and introducing triangulation for every criterion.

We have also attempted to establish that graders' performance constitutes a critical success factor for the assessment procedure. Experienced academic interpreting and language trainers and neophyte graders both have their own specific training needs. The former group may have to 'unlearn' their traditional holistic evaluation framework to adopt the criterion-referenced grid-based approach. The latter may still have to acquire the ground rules of assessment. In the same way that the standard for social interpreters is the cornerstone for an objective and valid certification procedure, a standard for graders will allow for more efficient training and assessment practice.

References

American Psychological Association (APA) 1985. *Standards for Educational and Psychological Testing*. New York: American Psychological Association.

Biggs, John 1999. *Teaching for Quality Learning at University. Buckingham*: SRHE and Open University Press.

Broadfoot, Patricia M. 1996. *Education, Assessment and Society*. Buckingham/Philadelphia: Open University Press.

Brown, James D. 1996. *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall.

Brown, James D. and Hudson, Thom 2002. *Criterion-Referenced Language Testing*. Cambridge: Cambridge University Press.

Cohen, Stanley 1979. "The Punitive City: Notes on the Dispersal of Social Control." *Crime, Law and Social Change* 3 (4): 339–363.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR)*. Cambridge: Cambridge University Press. See also: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf [2008.01.18].

Crocker, Linda 1997. "Assessing Representativeness of Performance Assessment." *Applied measurement in Education* 10 (1): 83–95.

Cronbach, Lee J. 1988. "Five Perspectives on Validity Argument." In *Test Validity*. Howard Wainer and Henry I. Braun (eds), 3–17. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dochy, Filip and Moerkerke, George. 1995. "Selectie van Beoordelaars en Criteria in Assessment Centers." In *Assessment Centers: Nieuwe Toepassingen in Opleiding, Onderwijs en HRM*. Filip Dochy and T. de Rijke (eds), 201–208. Lemma: Utrecht.

Dreyfus, Hubert L. and Dreyfus, Stuart E. 1986. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: The Free Press.

Durkheim, Emile 1947. *The Division of Labor in Society*. New York: The Free Press.

Fossum, John A. and Arvey, Richard D. 1990. Marketplace and Organizational Factors that Contribute to Obsolescence" In *Maintaining Professional Competence: Approaches to Career Enhancement, Vitality and Success through a Work Life*. Sherry L. Willis and Samuel S. Dubin (eds). San Francisco: Jossey-Bass.

Foucault, Michel 1977. *Discipline and Punishment. The Birth of the Prison*. London: Allan Lane.

Gronlund, Norman E. 1985. *Measurement and Evaluation in Teaching.* New York: Macmillan.

Groot, de Adriaan D. 1971. *Methodologie*. Digitale Bibliotheek der Nederlandse Letteren. http://www.dbnl.org/tekst/groo004meth01_01/ [2008.01.18].

Guion, Robert M. 1980. "On Trinitarian Doctrines of Validity." *Professional Psychology* 11: 385–398.

Idh, Leena 1997. "The Swedish System of Authorising Interpreters." In *The Critical Link 4. Professionalisation of Interpreting in the Community*. Cecilia Wadensjö, Brigitta Englund-Dimitrova and Anna-Lena Nilsson (eds), 135–138. Amsterdam/Philadelphia: John Benjamins.

Katz, Jerome and Gartner, William B. 1988. "Properties of Emerging Organizations." *Academy of Management Review* 13 (3): 429–442.

Lianos, Michalis 2003. "Social Control After Foucault." *Surveillance and Society* 1 (3): 412–430.

Lievens, Filip 1998. "Factors which Improve the Construct Validity of Assessment Centers: a Review." *International Journal of Selection and Assessment* 6 (3): 141–152.

Lievens, Filip 2001. "Assessor Training Strategies and their Effects on Accuracy, Interrater Reliability, and Discriminant Validity." *Journal of Applied Psychology* 86 (2): 255–264.

Lievens, Filip and Peeters, Helga 2008. "Interviewers' Sensibility to Impression Management Tactics in Structured Interviews." *European Journal of Psychological Assessment* 24 (3): 174–180.

Linn, Robert L., Baker Eva L. and Dunbar, Stephen B. 1991. "Complex, Performance-Based Assessment: Expectations and Validation Criteria." *Educational Researcher* 20 (8): 15–21.

Linn, Robert L. 1997. "Evaluating the Validity of Assessments: the Consequences of Use." *Educational Measurement* 16 (2): 14–16.

Lysaght, Rosemary M. and Altschuld, James W. 2000. "Beyond Initial Certification: the Assessment and Maintenance of Competency in Professions." *Evaluations and Program Planning* 23: 95–104.

McNamara, Tim 2001. "Language Assessment as a Social Practice." *Language Testing* 18 (4): 333–349.

McNamara, Tim and Roever, Carsten 2006. *Language Testing: The Social Dimension*. Malden MA: Blackwell Publishing.

Meyers, Raymond and Houssemand, Claude 2006. "Comment Évaluer les Compétences Clés dans le Domaine Professionnel?" *Revue Européenne de Psychologie Appliquée* 56 : 123–138.

Messick, Samuel 1988. "The Once and Future Issues of Validity: Assessing the Meaning and Consequence of Measurement." In *Test Validity*. Howard Wainer and Henry I. Braun (eds), 33–45. Hillsdale, N.J.: Lawrence Erlbaum.

Messick, Samuel 1989. Validity. In *Educational Measurement* (3nd.ed.) Robert L. Linn (ed.), 13–103. New York: Macmillan.

Moss, Pamela A. 1994. "Can There be Validity Without Reliability?" *Educational Researcher* 23 (2): 5–12.

Niska, Helge 1991. "A New Breed of Interpreter for Immigrants: Contact Interpretation in Sweden." In *Proceedings of the Fourth Annual Conference of the Institute of Translation and Interpreting.* Catriona Picken (ed.), 94–104. London: ITI.

Niska, Helge 1999. "Testing Community Interpreters: a Theory, a Model and a Plea for Research." In *Interpreting in the Community*. 278–287. Mabel Erasmus, Lebohang Mathibela, Erik Hertog and Hugo Antonissen (eds). Pretoria: Van Schaik.

Nitko, Anthony J. 2001. *Educational Assessment of Students.* Upper Saddle River, NJ: Merrill Prentice Hall.

Popper, Karl R., 1934. *The Logic of ScientificDiscovery*. New York: Routledge.

Salaets, Heidi, Segers, Winibert and Bloemen, Henri 2008. *Terminologie van het Tolken.* Nijmegen: Vantilt.

Shepard, Lorrie A. 1984. "Setting Performance Standards." In *A Guide to Criterion-Referenced Test Construction. Motives, Models, Measures and Consequences.* Ronald A. Berk (ed.), 169–198. Baltimore, MD: Johns Hopkins University Press.

Sociaal-Economische Raad voor Vlaanderen (SERV) and Centrale Ondersteuningscel voor Sociaal Tolken en Vertalen (COC). 2007. *Beroepscompetentieprofiel Sociaal Tolk*. Brussels: SERV. See also http://www.serv.be/Publicaties/1280.pdf [2008.01.18].

Stansfield, Charles W. and Hewitt, William E. 2005. "Examining the Predictive Validity of a Screening Test for Court Interpreters." *Language Testing* 22 (4): 438–462.

VandenBos, Gary (ed.) 2007. *APA Dictionary of Psychology*. Washington: American Psychological Association.

Wadensjö, Cecilia 1998. "Community Interpreting." In *The Routledge Encyclopedia of Translation Studies*. Mona Baker (ed.), 33–37. London/New York: Routledge.

Weber, Max 1923. *General Economic History*. London: Allen and Unwin.

Weber, Max 1947. *The Theory of Social and Economic Organization*. New York: Free Press.

Weiler, Hans 1981. *Compensatory Legitimization in Educational Policy: Legalization, Expertise and Participation in Comparative Perspective*. Stanford: University of Stanford.

Appendix 1. Sample evaluation grid

EVALUATION GRID – DUTCH PROFICIENCY		
Part 1: voice	±	**Examples**
Articulates clearly / Natural intonation pattern		
Sufficient audibility		
Acceptable pace		
Part 2: European Framework B2 level	±	**Examples**
Vocabulary:		
General vocabulary		
Specific terminology		
Grammatical structures:		
Communicative skills:		
Is able to understand the interlocutors		
Is able to participate in a conversation		
Is able to formulate an opinion		
Is able to enumerate and to discuss pros and cons		
FINAL GRADE – DUTCH PROFICIENCY:		**fail**
		pass
		excellent

Assessing ASL-English interpreters

The Canadian model of national certification

Debra Russell and Karen Malcolm
University of Alberta / Douglas College

This chapter highlights the certification processes for signed language interpreters in Canada. The Association of Visual Language Interpreters implemented its first evaluation mechanism in the early 1990s. In 2002 AVLIC reviewed the testing system, examined the current test construction research, and determined a new model of certifying interpreters. The result was a comprehensive and responsive test process designed to support interpreters in pursuing certification. The model includes a written test of knowledge, mandatory participation in three professional development seminars, and a performance test. The seminars are designed to address the interpreting patterns that were most common when the results of unsuccessful test takers over the past 10 years were analyzed. Candidates also receive feedback on samples of interpreting, designed to guide test takers prior to taking the performance test.

Introduction

This chapter provides an overview of the development of a responsive national certification system for American Sign Language-English interpreters in Canada, exploring the past and current contexts of testing interpreters. The testing system in Canada will be contrasted with that of the processes used to test interpreters in the United States and Australia, highlighting the rationale and differences in creating the testing model developed by the Association of Visual Language Interpreters of Canada (AVLIC). The chapter will highlight the purpose of the test, test methodology and procedures, and test construction and piloting processes. In addition, test criteria, scoring procedures, rater training, and the method of reporting test results to candidates are described. Finally, future directions for AVLIC and the testing of interpreters are explored.

Finally, this chapter generally adopts the text conventions described by Janzen (2005) in which he refers to "interpreting" as the activity that interpreters undertake and participate in, and "interpretation" to refer to the product of the activity.

The Canadian context

The Association of Visual Language Interpreters of Canada (AVLIC) was founded in 1979 and has served as the national organization representing signed language interpreters in Canada since that time. AVLIC has eight regional affiliates across Canada, and signed language interpreters who are actively working in the field are required to hold dual membership in both their provincial/regional affiliate and the national organization (see http://www.avlic.ca). The association operates with voluntary direction from board members to conduct the day-to-day affairs of AVLIC, and two part-time staff members. With a membership of approximately 500 interpreters, the association represents interpreters on issues of national importance, such as working standards, guidelines for ethical and professional practices, and certification processes.

In the earliest stages of AVLIC, the organization was comprised of both ASL-English interpreters and French-Langue des Signes Québécoise (LSQ) interpreters, reflecting the multilingual nature of Canada. As a national organization, we worked to honor the commitment to provide information in both official languages of English and French. However, as AVLIC evolved, this responsibility brought tremendous financial costs and complex logistical concerns to the organization. In the late 1990s, discussions ensued with our French colleagues; it became clear that we were not effectively representing the concerns of French-LSQ members, and that AVLIC could no longer continue to operate as a multilingual organization, offering services in English, ASL, French and LSQ.

AVLIC has consistently engaged in collaborative efforts with the national organizations representing Deaf[1] people in Canada, such as the Canadian Association of the Deaf (CAD) and the Canadian Cultural Society of the Deaf (CCSD), seeking guidance and direction on issues that are of common concern among interpreters and Deaf Canadians. The nature of these collaborative relationships also had a significant impact on the shaping of the evaluation mechanisms created by AVLIC. For example, both CAD and CCSD have had representation on the Evaluations Committee since its inception in the early 1980s, and they continue to ad-

[1] This paper adopts the common convention of referring to culturally Deaf persons with an upper case "D" on Deaf, and using a lower case "d" when referring to audiological deafness, as suggested by Padden and Humphries (1988).

vise AVLIC on issues of test content, processes, proctoring standards, criteria and standards for success on the Canadian Evaluation System (CES), and so on. Both organizations have also contributed financially to the CES, and their input determined the performance standard that was set for attaining national certification.

Prior to the existence of AVLIC, and during the early years of AVLIC's development, some interpreters in Canada chose to be members of the U.S. professional association known as the Registry of Interpreters for the Deaf (RID), and to avail themselves of the testing and certification processes offered at that time in the United States. However, from the earliest inception of AVLIC, there was a desire for a Canadian evaluation system that would reflect Canadian content, and represent the language use of Deaf Canadians within the testing samples. The first AVLIC Canadian Evaluation System tests were developed in the late 1980s and offered for the first time in 1990.

The initial testing system

The initial test was developed in consultation with a series of teachers of interpreting and American Sign Language, interpreting practitioners, and persons deemed to have specialized expertise in language, culture, interpreting, and test construction. The result was a two-part test comprised of a Written Test of Knowledge (WTK) and the Test of Interpretation (TOI).[2] Two equivalent versions of both the written and performance tests were created.

The written test consisted of 75 questions, which focused on three areas of knowledge deemed necessary for a professional interpreter: interpreting, culture and language, and AVLIC and related organizations. Candidates needed to secure 70% or better on the written test in order to proceed to the performance test. Their pass status remained valid as long as they maintained their active membership status in AVLIC. The Written Test of Knowledge was revised in 2000–2001 in order to reflect current research and developments in the field. At present, the test is offered twice a year, in the fall and spring. Beginning in 2005, it was also made available online, so that candidates could opt to take the test in one location with a proctor present, and were able to obtain their results immediately. Any active member is eligible to take the WTK. Once a candidate has successfully passed, she maintains her pass status as long as she continues to be an active member.

[2] The Test of Interpretation (TOI) is the formal name given to the performance portion of the AVLIC certification process.

The performance test required candidates to demonstrate their interpreting skills with an ASL narrative,[3] an English narrative, and three interactive or dialogic segments, where both English and ASL are used. The test was offered annually, in March, at testing centers across Canada. Candidates interpreted one narrative text, twenty minutes in length, from the source language of English into the target language of ASL; another narrative, also twenty minutes in length, from ASL to English; and three interactive segments, each fifteen minutes in length, which they were able to select out of a list of five topics. The dialogic segments included scenarios that represented community interpreting, for example, a doctor-patient interview, a parent-teacher interview, or a workplace conversation between two colleagues. The register used in all of the dialogic segments was consultative in nature, and the assignments were such that, when booked in the community, would require an experienced interpreter. While none of the scenarios involved specialized or technical language, they would not have been viewed as "easy" assignments. Performances were videotaped by a Deaf proctor, and then sent to the CES office for copying and distribution to raters.

Setting the standard and identifying the criteria

A significant prerequisite for developing the CES was determining the criteria for passing. To this end, a Criteria Development Project (CDP) was undertaken, employing the services of two highly respected U.S. consultants, M. J. Bienvenu, a Deaf ASL expert, and Betty Colonomos, an interpreter expert. At the time they were the co-directors of the The Bilingual-Bicultural Center in Maryland. Both woman had expertise and experience in the creation of interpreter and signed language tests, and had consulted widely on the topic of assessment. Both possessed graduate degrees and were working on their doctorates at the time their services were contracted to AVLIC. They were also active in the professional organization known as the Conference of Interpreter Trainers (CIT), and had played a major role in creating the seminal document that was a task analysis of signed language interpreting, and the development of curricula suitable for language learning and interpreting. What follows is a description of the steps undertaken in the project.

Under the direction of the consultants, the chapters of AVLIC were asked to collect samples of interpreting by their members, from both ASL to English and English to ASL, to use as the materials to determine the standard. The interpreters who submitted samples of work were considered competent interpreters chosen

[3] Narratives are monologic presentations, delivered in a formal or consultative register: for example, a conference presentation.

to represent the diversity within the interpreting communities. The diversity included features of age, gender, education, native usage of ASL, later acquisition of ASL, and years of interpreting experience, and all had experience performing community-based and post-secondary interpreting.

A group of eighteen nationally identified language and interpreting experts, evenly divided between the domains of English, ASL, and interpreting, gathered to set the standard of acceptable performance for a national certification test. The ASL experts were recommended by the national organizations of Deaf people, and were either Deaf ASL teachers or Deaf ASL researchers. English experts were teachers of English and/or had interpreting experience, but were not actively working as ASL/English interpreters. Interpreting experts were those who were recognized and respected by their colleagues as experienced interpreters; these recommendations came from the AVLIC chapters. The group of experts brought a number of strengths to the task, as all of them had considerable experience with teaching and assessing ASL, English, and interpreting. They were knowledgeable about interpreting in Canada and the United States. In addition, all of them were part of North American professional networks of other interpreting and language-teaching colleagues, and possessed an understanding of interpreting that reflected the current state of knowledge at that time. In addition, they were familiar with the demands of community interpreting and the quality of interpreting service required. For all three groups, regional representation was also considered to ensure that the standard set would reflect a national perspective. However, in some regions the expertise sought was not available, so that experts from the closest region were then selected. The final group was comprised of persons representing the provinces of Quebec, Ontario, Manitoba, Alberta and British Columbia. It is interesting to note that in this period of 1988–1989, the association did not invite consultation from testing experts from the field of spoken language; if they had been involved, it might have strengthened the process. In addition, testing and measurement experts from outside of the field of interpreting were not considered, and their involvement might also have strengthened the process.

The external consultants first asked the group to consider what competencies a nationally certified interpreter would hold. This started the documentation of areas to address. The experts then viewed several samples of interpreting performed by one of the external consultants. After each sample, the group talked about what they had seen in the work, and why it would be considered a pass or fail. The presence of Deaf people in determining the standard was crucial in creating an evaluation system respected by the larger Deaf and interpreting communities.

The experts were then divided into three groups, ASL, English, and interpreting, and began to view the samples of interpreting that had been gathered by the chapters. Raters watched each performance without discussion, voted pass or fail, and then discussed the reasons for their decision. The experts were free to determine the linguistic and interpreting features based on their observations and discussions. While the consultants facilitated the conversations, they did not provide a standardized assessment tool or model that could potentially influence the experts. Consultants participated in assisting the teams in identifying the features or traits of the work that were deemed successful. They enriched the discussions by bringing forward knowledge of interpreting theory and research, drawing on the work of Danica Seleskovitch[4] with spoken language interpreters, and Colonomos's own development of a pedagogical model of interpreting. The Colonomos model addressed language transfer issues within a theoretical framework of comprehension, visualization of the target message, and production. The model also accounted for some of the multiple variables that affect discourse and interaction, such as setting, register, goal of the interaction, status and relationship of the participants, and intercultural views of power. Just prior to the development of the CES, the Conference of Interpreter Trainers (CIT) had published a task analysis document highlighting the cognitive tasks needed in order to produce effective interpreting, and this task analysis was helpful in shaping the criteria. In addition, both consultants represented a philosophical framework of intercultural communication, drawing attention constantly to the nature of working between two languages and cultures and the ways in which this impacts the interpreting product. In reviewing the processes through which Colonomos and Bienvenu led the organization, we found that their approach also incorporated the thinking of Canale and Swain (1980) with regard to communicative performance, which they defined as grammatical and sociolinguistic competence, along with strategic competence in order to use the language in a meaningful communicative situation. Canale (1983) expanded the previous Canale-Swain framework to include discourse competence, which was understood to reflect the ability to combine and interpret meanings and forms to achieve cohesive language texts across different contexts and language registers. The consultants encouraged the experts to see language and the interactional goals of the participants as key, and to avoid viewing interpretation at only the lexical level, but rather to see it as discourse-based.

Over the course of the two days, the rating process and ensuing discussions (sometimes in domain-specific groups of ASL, English, and interpreting, and sometimes as an entire group) led to the determination of what constituted a

[4] Seleskovitch's primary work has been examined and critiqued by others for its lack of empirical basis. It is noted here simply as acknowledgement of an early influence on the work of Betty Colonomos. Readers are referred to Gile (2006) and Wadensjö (1998) for further assessments of Seleskovitch's work.

pass. ASL and English experts identified features of conventional language[5] use needed, while the interpreting experts focused on the components of successful interpreting. Participants decided that interpreting performance needed to demonstrate meaning-based discourse, rather than form-based or lexical equivalent approaches. Given that the majority of signed language interpreters work from their L1 (which in most cases is English) into their L2 (in most cases ASL) the guidance of the ASL experts was crucial in setting a standard for ASL use as well as interpreting that met the needs of the Deaf community in Canada.

Criteria for successful demonstration of test

AVLIC's rating system is based on a set of linguistic and discourse features in American Sign Language and interpreting per the Message Equivalency domain. These items represent the key features an interpreter must demonstrate in order to be deemed successful in that domain (For a more detailed explanation of the ASL and Message Equivalency features, see Appendices 1 and 2).

The ASL raters examine the performance across three major bands: discourse strategies, linguistic form, and register. Within each band there are statements of standards; for example, overall discourse strategies used result in a coherent text. There are linguistic features mentioned, such as linguistic devices and discourse strategies conventionally seen when introducing a topic and transitioning to another. These are called strategies of "opening and closing" the message. Also required are essential elements of meaning with sufficient supporting details, appropriate use of topic transition and topic maintenance strategies, and avoidance of unwarranted restatement of ideas that are not present in the source message. In terms of the marking forms, raters have a standard form, and they note the presence, absence, or inconsistent use of the discourse traits that form the criteria.

The following examples serve to illustrate how the raters apply the criteria. If a candidate is able to use American Sign Language consistently, and is able to produce the language in ways that are conventional and register appropriate, they

[5] By Conventional Language Use, the experts adopted features that resulted in the language appearing natural and being able to be understood easily, such as topic cohesion and discourse cohesion, prosodic elements such as pausing and phrasing, lexical and grammatical constructions that were consistent for consultative and formal register demonstrated in the nature of the interactions, natural "openings and closings" within the boundaries of the utterances, and the ability to represent an unfolding conversation within an interaction. A consultative register is the register often found in a teaching or interview context, and formal register reflects the register used by a presenter to an audience where one does not expect to be interrupted, such as a motivational or technical lecture (Grice 1981; Halliday 1981; Joss 1961).

would be scored as having the presence of ASL grammar and choosing the correct register to match the speaker for whom they are interpreting. By contrast, if a candidate had very little control of producing ASL grammar and was not able to mark topics in the language, and yet offered some supporting detail, the scoring would indicate inconsistent for ASL grammar, and absence of the other targeted linguistic features.

Similarly, the Message Equivalency area has several interpreting areas that require the demonstration of ASL-English interpreting strategies to proficiently construct an effective and equivalent message[6] in the target language, including appropriate lexical choices, tone, grammar and syntax, with appropriate use of register, pausing and phrasing, rhythm, intonation, pitch, and other supra-segmental features. The interpreting area also looks at the candidates' ability to be able to comprehend the source message and provide an effective target language message. Finally, raters look at the gravity of errors of the message, noting miscues as omissions, additions, or interpreter anomalies. Interpreter anomalies are those features that are not part of the source message, but are unusual patterns that belong to the interpreter and detract from the overall message accuracy. For example this occurs when interpreters have a pattern of false starts, beginning a sentence several times, repairing the language used in the sentence, or the use of phrases such as "you know" that detract from accurately representing the presenter.

The ASL raters and Message Equivalency raters used the same scale when reviewing the features. Within this scale the raters noted whether the feature was consistently present in the interpreting work, or absent, or was inconsistently demonstrated in the interpreting.

The following examples will highlight the application of the criteria by the ME raters. The Target Language examples are exactly what the interpreter said or a translation of what was signed.

Example one
Source Language – English: So, Sam, what brings you into the office today? Did the medication prescribed last visit work for you?

Target Language: ASL: *Why are you here today? On the last visit, I gave you medication – did it work?*

[6] The term "equivalency" reflects the development of the criteria in the early 1990s: however the raters understand that an interpreted version is never "equivalent". It can be a very close rendition, but by the sheer nature of being a rendition, the text is different. An effective interpreted rendition realizes all of the features identified in the criteria, and AVLIC is preparing to adjust the language to more accurately reflect current understandings of interpreting.

The raters would view the opening construction to be effective in ASL and containing all of the essential elements of meaning. It presents a coherent message that is accurate and represents comprehension of the Source Message. Overall, the target message is accurate and there are no miscues that skew the interpreting.

Example two

Source Language – ASL: No the medication did nothing. I think the infection is worse now, so that is why I am back.

Target Language: *No, that medication did nothing for me. That is why I am back.*

The raters would view this as not demonstrating all of the essential elements of meaning, in that there is no mention of the infection and its worsening state. So, while the English grammar and register are appropriate, the omission of a key proposition results in an inconsistent rating per the Essential Elements of Meaning.

Example three

Source Language – English: I will need to check that, but before I do that, I want to review your blood work and ask you a few more questions about your reaction to the drug.

Target Language – ASL: *OK – I'll check your blood pressure and ask a couple more questions.*

In this example the raters would choose to score this as inaccurate work that does not demonstrate the features of complete and accurate content per the criteria.

While these examples are at the utterance level, as in a question asked and answered, the raters look at accuracy within the utterances and across the interaction. In addition, they examine the adjacent pairs offered, and seek to review the cohesion present within the interpreted interaction. The criteria for both domains are published on the AVLIC website (see http://www.avlic.ca), and the workshops that candidates take serve to demystify the features.

Within a text, interpreters may be required to exhibit strategies for comparing and contrasting ideas and referencing previously introduced information. This approach to identifying discourse features of some of the competencies needed to interpret is consistent with the approach suggested by Clifford (2001). Clifford argues this approach provides greater rigor in constructing valid interpreter assessment tools. By "form," the raters look for overall language use that is clear and intelligible; so for example, in ASL, the signs must be produced accurately and clearly, or in English, the words must be articulated clearly. Form also refers to the use of grammatical markers that are accurate per the norms of the languages. For example, the use of markers may include tense/time markers, plurals, pronoun use, and ability to represent the grammars of both languages through complete and grammatically correct sentence structure. Both pausing and phrasing need to observe the norms of the language, and in ASL, if fingerspelling is used it must be appropriate for the context, with the words spelled correctly and clearly.

The interpreting criteria require the interpreter to demonstrate the use of bilingual and bicultural strategies in English and American Sign Language. Successful interpreting performances in this domain demonstrate the application of a processed or sociolinguistic model of interpreting (Cokely 1992). The criteria are divided into areas such as Message Processing and Interpreting Sub-Tasks. The standard to be met is that the overall message processing results in coherent and effective rendition. This includes the ability to convey the goals of the participants in the interaction, representing the essential elements of meaning and supporting detail in order to convey the points and appropriate use of contextualization, to introduce new information or previously referenced information. The interpreting must be processed beyond the lexical level and must not be marked with false starts. In terms of Interpreting Sub-Tasks, the interpreters must demonstrate that they can comprehend the source message and produce an interpreted message that is accurate and grammatically correct. The criteria address the need for register-appropriate work, and the interpreter's ability to monitor their own work in order to make corrections appropriately. The raters observe whether the interpreter appears confident and demonstrates few or no personal distracting mannerisms. Finally, the raters also examine the miscue or mistake patterns that may arise in the work, noting where the impact of the miscues is excessive. A miscue can include omissions of content, additions of information not found in the source text, or substitutions that do not represent the meaning of the source text.

The ASL and interpreter specialists, the workshop facilitators, and the raters all use the same terms to describe the linguistic and interpreting features, and AVLIC has produced materials that show the interpreting work of certified interpreters as a model for test takers to review.

Rating processes

Subsequent to the CDP meeting, the first offering of the Test of Interpretation took place across Canada.[7] Raters were selected, including many who had participated in the CDP process. Raters were trained to identify the criteria, using the Canadian test samples that set the standard at the CDP meeting. The standard has remained constant throughout the history of the CES, and did not change with the revision of the testing process that took place in the early 2000s.

[7] "Test of Interpretation" is the official name AVLIC has given the test.

Candidates were rated in three areas: English, American Sign Language, and Message Equivalency (ME). In both the English and ASL domains, raters were solely determining if the language use was grammatically and semantically correct, used an appropriate register, had clear cohesion, and fulfilled the communicative functions within the discourse. The Message Equivalency raters rate the interpreting work separately.

In the English domain, raters listened to the spoken English on the tapes and determined if it met the criteria of being grammatically correct, clearly articulated, and cohesive. If the first rater deemed the performance a pass, the tape then moved to the next domain, ASL. If it was deemed a fail, it was sent to a second English rater who did not know whether the tape had been previously rated. If the second rater determined it was a fail, the candidate's tape proceeded no further and the performance was deemed unsuccessful. If the second rater deemed it a pass, the tape was sent to a third rater, and that rater's decision determined whether the tape continued on or exited the rating process.

It was important to have the language domains rated first, since competence in both languages is a prerequisite for Message Equivalency. The decision to rate English first was arbitrary; ASL could just as easily have been the first domain rated. There is no implication intended that English skills are a prerequisite for ASL skills.

The next domain to be rated was ASL. A similar process to that of the English assessment was followed. Once two raters determined a fail, the tape then exited the rating process. Therefore, only tapes that had passed English and ASL domains continued on to the Message Equivalency rating stage.

Message Equivalency raters were interpreters who were able to view the signed source message and listen to the spoken interpreted rendition; listen to the spoken source and watch the signed rendition; and determine whether the performance met the set criteria. Again, if the first rater deemed the performance to be a pass, the candidate was certified with the Certificate of Interpretation (COI). If the first rater determined it was a failing performance, two other raters needed to agree to a pass score or else the performance was deemed to fail. Raters were trained to assess borderline performances as a fail, so the decision to award certification with one passing vote would not skew the test results, and was also perceived in a positive light by test takers. However, the potential did exist in this system for a false pass, meaning that a person passed this domain even when their performance did not meet the criteria. This weakness in the testing system has since been addressed in the new testing system.

Candidates who were unsuccessful were notified where they exited the process, be that in the domain of ASL, English, or Message Equivalence. No appeals were permitted on the decisions reached, but appeals related to administrative process or technical difficulties were allowed. Successful test takers were awarded the Certificate of Interpretation (COI).

After a few years, the AVLIC Evaluations Committee determined that a change was needed in the rating of Message Equivalency. Because the population of signed language interpreters in Canada is small, candidates were often known to the raters, which created potential difficulties in rating. Raters were colleagues and sometimes friends to the candidates, which created tension in evaluating their performances. In addition, many of the performances included a mix of successful performance with unsuccessful performance, and raters reported finding it very difficult at times to determine a pass or fail on their own.

Consequently, the committee decided that Message Equivalency rating would be done by a group of three raters meeting face to face. A facilitator was also present, to cue tapes and record results, and also to ensure that discussion consistently focused on the criteria developed by the Criteria Development Project. Raters were separated by screens so they could not see each other, and instructed to make no vocal utterances that could indicate their opinion of the performance. Raters watched the whole test performance, and then individually voted pass or fail. The facilitator recorded the votes and then informed the raters of the result. If there were three passes, the candidate was granted the COI; three fails resulted in the candidate's exit from the system; and a split vote required the raters to discuss the performance and work on arriving at consensus. Reaching consensus can be a challenging process. Recognizing the potential for a strong personality to sway the decision of the other two raters, AVLIC retained the services of a facilitator for these discussions. The facilitator's role was to ensure that all comments on the performance referred to the criteria, and that raters were listening to each other in a mutually respectful manner. In this way, what may be perceived as problematic in reaching consensus has been addressed.

The divergent rater spoke first, noting areas of the performance relative to the established criteria (see Appendix One and Two for further description) that led to the decision, and then the other raters had a chance to respond with their observations. Raters were then asked if anyone wanted to change their vote, based on what they had heard. If not, more discussion ensued, and at times raters viewed parts or all of the test tape again. If still unable to reach a decision, raters could put the tape aside and move on to evaluating other performances, returning to the undecided tape later in the day, or on the following day.

Dealing with reactions

The initial offering of the COI resulted in some consternation among interpreters who failed, especially those who had expected to pass due to years of experience in the field. However, representatives of Deaf community organizations strongly upheld the standard, as did many interpreters. As statistics were gathered, analysis of the results of the performance test showed a pass rate of twenty-three percent. Over the years from 1991 to 2000, the pass rate increased to 31% and 38% during 1996 and 1997, and then returned to 25% in 1999. During 1996 and 1997, AVLIC and its chapter affiliates offered TOI Preparation Workshops in some cities across Canada, taught by a Deaf-Hearing team of educators, and it is suggested that the increase in the pass rate was related to those who had the opportunity to participate in the preparation workshops and receive feedback about their readiness to take the TOI.

The approach to setting an absolute performance standard is referred to as the "criterion-referenced" method (Bachman & Palmer 1996; Hudson 2005) and involves linking decisions about examination performance to criteria for acceptable certification standard of the relevant profession. The objective of these "test-centered" methods is to set a performance standard on the examination, with the expectation that those who meet the standard will be judged as meeting it, and those who are judged as not meeting the standard will fail the examination. With any criterion-referenced method of setting the pass rate of a certification or licensing exam, the goal and challenge is to set a standard high enough so that it reliably distinguishes between those who are meeting the standard for certification and those who are not, but not so high that the standard excludes those who are competent from meeting the standard (Johnson & Squire 2000). While many organizations use the Angoff Method (1971) to determine the minimal pass performance, this was a method that AVLIC chose not to use, in that it would have lowered the standard needed to pass, and this was unacceptable to the Deaf community organizations. The pass rate continues to be a matter of debate among the members of AVLIC, with some members advocating for a minimal pass that would allow for greater numbers of candidates to pass. To date, the AVLIC membership has continued to uphold the original standard.

Bachman (2007) has reviewed testing constructs, and argues that the language testing field has typically used three general ways of defining constructs that are assessed: (1) ability-focused, (2) task-focused, and (3) interaction-focused. As Bachman (2007) traces the past thirty years of language testing, he describes the earliest ability-focused testing as those tests that emphasized language skills of the test-taker. Task-focused testing drew upon an understanding of designing tests that looked at communicative competence that were situated in authentic tasks. Finally, he points to more recent research in social interaction and discourse analysis that has shaped testing to view the construct we assess not as an attribute of individual language users or of the context, but as jointly co-constructed by the participants and the patterns within the interaction. He suggests that test design and development, and the use of assessments must address all three of these approaches (ability, task, interaction) if we are to address some of the limitations or weaknesses of the assessment process.

From 1996 to 1998, expanded documentation of the criteria was undertaken by a group of interpreting experts. The criteria were outlined in documents available to members, and written in language that would make them understandable to all raters (See Appendix One and Two). In addition, feedback began to be offered to test takers who were unsuccessful, so that they learned where their performance did not meet the standard as outlined in the criteria. This documentation process represented attempts to be more explicit about what testing candidates needed to demonstrate in order to be successful in managing the content, construct, and interactional demands of the testing samples, and to strengthen the assessment process.

While the interpreting and Deaf communities continued to support the maintaining of the standard for passing the TOI, there was growing concern regarding the small number of certified interpreters in Canada. Discussions within the Evaluations Committee resulted in a proposal of changes to the system to assist interpreters in obtaining certification. This led to the development of new performance test materials during 2004, as well as a new four-step process for certification. The criteria and standards for passing did not change; but the new materials incorporated support for the test-taker to be able to demonstrate interpreting skills successfully. While the new test is just in the early stages of being offered, preliminary results would suggest that the changes have resulted in an increased passing rate.

One of the features that is unique in the Canadian context is that all of the test candidates have taken a formal interpreter education program, with a minimum of two years of full-time study. There are five ASL-English interpreter education programs at the current time in Canada, ranging in length from a two-year diploma program to a bachelor degree program. The programs have recently begun meeting regularly, and one of the issues discussed has been the gap between graduation from a program and readiness to practice. This gap is also part of the concern that has resulted in some of the changes to the AVLIC testing system, most noticeably the addition of workshops focused on discourse analysis, which appears to be an area that is not sufficiently addressed in the interpreter education programs.

What follows is a brief comparison of the testing processes for signed language interpreters in the United States and Australia. This information will provide a context for how signed language interpreters are tested in two other countries. AVLIC reviewed these other models of testing prior to revising their testing model: this was beneficial in leading to decisions that would be effective for Canadian interpreters.

The Australian context

Napier, McKee, and Goswell (2007) describe the accreditation process that is in place in Australia. One body, known as the National Accreditation Authority for Translators and Interpreters (NAATI), regulates the testing processes. Signed language interpreters in Australia use the signed language indigenous to that country, known as *Auslan*, which has been part of the NAATI test system since 1982. One of the unique features within the Australian context is that spoken and signed language interpreters are tested through the same system. NAATI has two approaches to certifying interpreters – one can take the examinations or one can pass a NAATI-approved training course (Napier, McKee & Goswell 2007). Two levels of certification exist – "Paraprofessional" and "Interpreter". The Paraprofessional designation is for entry level interpreting, which NAATI suggests is the competence required for general conversations and non-specialist dialogues. The Interpreter level of certification is described as the professional level of interpreting, where the interpreter is capable of working in a wider range of settings, including conferences, public events, and legal matters.

For interpreters who take training in order to gain their certification level, interpreters accredited at the Paraprofessional level may take part-time courses over one year, available at technical and further education colleges throughout Australia. Interpreters who have completed an advanced diploma or a postgraduate diploma at the university level qualify for the Interpreter level of accreditation without having to take the exams.

Napier (2005) reported that NAATI has recently reviewed the testing procedures for Auslan interpreters, and a number of changes have been recommended in the content and structure of the tests. Auslan interpreters will be tested on their ability to provide both a "free" and a "literal" simultaneous interpretation[8] to different target audiences, thus testing interpreters in both interpretation and transliteration,[9] with less emphasis on consecutive interpreting. The rationale for

[8] Free interpretation focuses on the meaning of the message, based on linguistic and cultural conventions, while literal interpretation focuses on the form of the message and results in producing a message that has more of the features of the source message. This is also known as transliteration in North American literature.

[9] Transliteration is the label used to describe the process used by interpreters to visually represent English words and grammar (Davis 2005).

this choice was that the test protocol was designed to test interpreters on what they realistically do on a regular basis.

The test has one section that tests consecutive interpreting and has three sections that test simultaneous interpreting. Within the simultaneous interpreting test segments, one section addresses a dialogue (professional context); another a monologue where the target audience is a bilingual deaf professional; and another a monologue where the audience is a monolingual Auslan user (personal communication, Jemina Napier, Sept. 22, 2008). Finally, NAATI is expected to introduce a process designed to test interpreters who have previously achieved certification, highlighting the need for all interpreters to demonstrate that they are maintaining their skills in order to be entitled to accreditation.

The United States context

The Registry of Interpreters for the Deaf (RID) is a national membership organization that provides the National Testing System for its members (see http://www.rid.org). RID began testing and certifying interpreters in 1972. Their testing process, until very recently, has always tested both interpretation and transliteration. During the 1990s the National Association of the Deaf developed its own interpreter assessment, which resulted in two different standards to regulate interpreters in the US. In recent years, RID and the National Association of the Deaf (NAD) have come together to collaborate on one testing system known as the NAD-RID National Interpreter Certification (NIC). The current NAD-RID test has three components: a written test, a videotaped interview, and a videotaped performance test. After December 2008, the multiple certificates will be streamlined, and RID will offer only the NIC process as a replacement for their Certificate of Interpretation and Certificate of Transliteration. Individuals achieving certification at the NIC, NIC Advanced or NIC Master level are all deemed professionally certified interpreters. The National Interpreter Certification (NIC) exam tests interpreting skills and knowledge in three critical domains:

a. General knowledge of the field of interpreting through the NIC Knowledge Exam.
b. Ethical decision making through the interview portion of the NIC Performance Test.
c. Interpreting and transliterating skills through the performance portion of the test.

In all three domains, certificate holders must demonstrate professional knowledge and skills that meet or exceed the minimum professional standards necessary to perform in a broad range of interpreting and transliterating assignments.

Testing materials remain the same for each level of certification on the NIC. For the interview portion, raters are trained to identify decision-making skills that meet or exceed basic professional standards. For the performance portion, they are trained to identify interpreting and transliterating performances that meet or exceed basic professional standards. Those candidates whose performances are at or exceed that standard are awarded certification. Those who pass as Certified have shown basic professional-level interpreting/transliterating skills. Those who pass as Certified Advanced have scored within the standard range on the interview portion, and high on the performance portion of the examination. Those awarded the Certified Master accreditation have scored high on both the interview and performance portions of the test. Finally, a unique feature of the RID testing process is that, beginning June 30, 2009, hearing candidates for certification must have a minimum of an associate's degree to take a performance exam. Deaf candidates must have a minimum of an associate's degree after June 30, 2012 (RID 2008). By requiring degrees as a pre-requisite to take the performance test, RID is taking an explicit stand on the minimal educational qualifications necessary to begin work as an interpreter. This decision has tremendous implications for the field, in terms of gaining recognition as a profession when its practitioners possess academic credentials. It is also a decision that has been controversial for experienced practitioners who do not possess formal education and have little desire to acquire it at later stages of their career. However, educational institutions have been creatively addressing education, and there are on-line degree options that are enhancing the access to education for interpreters, regardless of their location.

Comparing testing models in Australia, Canada and the United States

Examination of the Australian and United States models of accreditation verifies that there are both similarities and differences across the models. Each of the countries highlighted in this chapter has sought to advance the profession of interpreting in their country and, as Napier (2005) has suggested, has demonstrated leadership in the area of accreditation of interpreters. By offering standardized forms of assessment, each of the organizations has increased the awareness of the role of professional interpreters. The necessary interpreter qualifications and skills suggested by these organizations are widely based on the original standards set by RID in the United States. Although the assessment tools vary between organizations, the overall interpreting skills being assessed remain the same. Most national organizations now evaluate interpreters on their critical analysis skills (ability to self-assess personal qualifications for an assignment), the ability to prepare for an interpreting assignment, ability to understand and accurately convey the original message in the target interpreting language, use of appropriate communication mode and language, and ability to facilitate the flow of communication effectively and across a variety of settings. In addition, all organizations are establishing the requirement for certified interpreters to maintain competence within the field of interpreting (e.g., by participating in ongoing professional development activities). By registering with these national organizations, individuals are encouraged to maintain their input and participation in the interpreting field, and to develop their awareness of the changes occurring within the profession.

More recently, these national organizations are including cultural awareness and its effect on interpreting as necessary background knowledge for interpreters. AVLIC, RID, and NAATI all suggest that the linguistic variation found across geographical regions influence appropriate signing, and have now included these measures within their certification processes. This includes ensuring that the interpreter matches the linguistic preferences of the consumers while monitoring message comprehension and feedback during the interpreted event (and is able to modify accordingly) (RID 2008).

The next section of this chapter describes the current AVLIC testing model, beginning with the testing purpose and overview of test methodology.

The new AVLIC testing model

Clifford (2001) suggests that there are four steps in the assessment cycle: intention (purpose of the assessment); measurement (data collection and marking), judgment (creation of a common system for interpreting that is understood by all of the assessors or raters), and decision (fairness and equity of decision making based on a rigorous examination of the first three steps). The mission or intention of the Canadian Evaluation System is "to accredit interpreters who demonstrate competencies that reflect the diverse communication preferences of Deaf and hearing Canadians".[10] This goal has remained constant throughout the history of certifying interpreters in Canada. However, the new model provides much more support for test takers to demonstrate the competencies needed to perform simultaneous and consecutive interpreting in narrative and dialogic settings.

[10] Association of Visual Language Interpreters of Canada (2008). *The Canadian Evaluation System*. Retrieved July 9 2008 from http://www.avlic.ca/services.php?evaluation

During 2002–2003, AVLIC brought together an ad hoc committee that could review current research and practices related to the measurement or assessment of interpreting. The committee also included an invited interpreter and researcher who has published on the topic of measurement and evaluation, Dr Andrew Clifford, to offer expertise and guide the discussions. The first aspect of planning for a new model was to consider the purpose of the assessment process and how best to gather the testing data. One of the issues considered by the committee was the use of a portfolio-based assessment system. However, after reviewing current research and investigating the use of portfolio assessments in testing contexts, the committee chose not to recommend this as an option for AVLIC, given the reliability issues and lack of standardized work samples that would be submitted by interpreters. Additional rationale that guided the decision included the following.

a. While portfolio assessment is used as an alternative to traditional assessment methods (paper/pen tests), it would appear that portfolio assessment is best utilized in an educational setting where coursework is being challenged, or to demonstrate cumulative learning (Baume & York 2002; Lombardi 2008). While there appears to be a growing educational literature that promotes portfolios as assessment tools, it is clear from the interviews conducted by Baume and York (2002) that university professors were enthusiastic about portfolio assessment, often at the expense of their own extra time and effort in assessing. Baume also reports that it became clear that well-established practice is still difficult to find, and practitioners are engaging in trial and error in designing their systems. If portfolios are to continue to spread as an assessment method and fulfill their educational potential, more attention must be paid to the question of efficiency. Fields like nursing, teaching, and social work are using portfolio assessment in combination with the completion of an academic degree to verify competencies.
b. Portfolio assessment processes have acceptable validity but very low reliability in determining work standards for the purpose of certification (Gadbury-Amyot et al. 2000; Nystrand, Cohen, & Dowlin 1993). For example, it may be very difficult to establish reliability on different work samples where variables such as setting, language register, context, preparation and so on are not controlled. Schutz et al. (2004: 52) note a lack of clarity regarding what constitutes reflection and performance, and as to whether reflection can be said to have levels and, if it does, how these might develop over time. Schutz et al. also claim that the assessment of reflection is problematic for assessors who are uncertain as to what they are assessing, and whose uncertainty extends to the relationship of grading to levels of reflection and performance.

c. In order to address the reliability issue, standardized performance tapes would have to be utilized in the portfolio process. The introduction of sufficient standardized test tapes into the portfolio assessment to address the reliability issues would require interpreters to produce an extensive portfolio and take the same test that is currently part of the TOI, in order that decisions about candidates could be made consistently and fairly.

d. Portfolio assessments tend to be time consuming and labor intensive, which usually translate into increased costs (Schutz et al. 2004). In addition, given that portfolios contain more information about their candidates, there is a greater chance that raters would be biased in their judgments.

e. There is limited research in the area of using portfolios as a certifying mechanism (Wilkerson & Lang 2003). From a pragmatic perspective, the development, administration, and scoring of portfolio assessments appear to be more complex and costly, but candidates may be more satisfied with the assessment process overall. As Ingersoll and Scannell (2002) pointed out, portfolios are not assessments, but are instead collections of candidate artifacts that present examples of what candidates can do. The contents need to be evaluated individually as part of the portfolio process and therefore need to meet psychometric standards of validity, reliability, fairness, and absence of bias. These standards, along with US federal law, form the cornerstone for legal challenges to decisions when students are denied a diploma or license based on the results of the assessment. If an organization cannot demonstrate these standards, a court decision against the organization can result in financial damages and damages to the institution's reputation.

f. Interpreters have reported that creating live work samples for inclusion in a portfolio would be an impossible task for those specializing in medical, mental health, or legal settings (AVLIC Evaluation Systems Coordinator, Monique Bozzer, personal communication, August 12, 2008).

The ad hoc committee continued to explore solutions to concerns raised by interpreters and Deaf community members about the viability of the current TOI. To that end, the committee prepared a plan that included the development of new testing materials, additional support for test takers in preparing for the TOI, and a certificate maintenance process. This resulted in a four-phase model consisting of the Written Test of Knowledge, Test Preparation Workshops, the Test of Interpretation, and Certificate Maintenance. It was felt that this comprehensive model would best exemplify all steps of the assessment cycle, including what is designated by Clifford (2001) as the fourth step of the cycle: fairness and equity in decision making.

The new four-step testing process

What follows is a thorough description of the new Four-Step Testing Process used by AVLIC, beginning with the Written Test of Knowledge, followed by the TOI Preparation Workshops, the Test of Interpretation and finally, Certificate Maintenance.

Written test of knowledge (WTK)

The WTK, revised in 2000–2001, is now offered online. Two test versions exist with each consisting of 75 questions. Any active member of AVLIC can take the WTK. The test is offered twice a year, in the fall and spring, at various locations throughout Canada. Preparation for the written test is available on the AVLIC website, where a list of reading materials and other resources is posted. Once a candidate has passed the WTK, she/he is eligible to enter the TOI preparation stage.

TOI preparation workshops

Test candidates are required to take two workshops that have an emphasis on discourse analysis strategies when working with narrative and dialogic segments that are assessed in the actual test. During 2004, AVLIC hired a curriculum development team to prepare these workshops. AVLIC then recruited several workshop facilitators, who were trained in how to deliver the workshops. The facilitators are all certified interpreters, with proven abilities as educators and/or mentors. They are individuals who teach in interpreter education programs and/or provide in-service training, and are nationally recognized organizational leaders.

The first workshop that participants take focuses on interpreting narrative material. This was determined to be the first workshop needed, because it was the area on the previous test where most test takers were unsuccessful. After registering, participants are provided with pre-reading materials and online access to narrative and dialogic videos that simulate the actual Test of Interpretation. Narrative segments, often called monologic presentations, are tested because signed language interpreters work in the community where they often are called upon to interpret in conference or educational settings where narrative presentations are given. Hence, the test format includes both narrative and interactive segments.

Participants are to tape themselves, and these tapes are then sent to interpreter and ASL specialists. These specialists are familiar with the AVLIC TOI criteria, and they provide the workshop participant with written feedback about their linguistic and interpreting skills. This feedback is designed to help the participants decide if they are ready to take the Test of Interpretation, and offer guidance on areas to improve prior to taking the test. After that, the candidates attend a two-day workshop focusing on interpreting narratives, led by a trained facilitator. They discuss the pre-readings at the workshop, and also use practice materials similar to the actual test ones. In addition, candidates meet individually with the facilitator for feedback on their interpreting work, and to obtain suggestions for practice that the test taker can implement in their development plan.

The second two-day workshop focuses on interpreting interactive or dialogic material. Once again, candidates are provided with a stimulus tape that they interpret and then send to an ASL specialist and an interpreting specialist for feedback. One of the changes to the test materials is that the interactive segments were filmed with enough time to allow test takers to interpret them consecutively if so desired, or to use a combination of consecutive and simultaneous. This is a significant change to the testing process, drawing on research supporting the use of consecutive interpreting for this form of discourse (Russell 2002). This approach also allows interpreters to engage in practices that reflect real life, in that an interpreter may choose to perform some utterances within an interpreted interaction in consecutive form, and others in simultaneous form, depending on the demands of the discourse and the setting. Thus, participants in the workshop practice working with interactive materials that include long pauses, in order to familiarize themselves with the process.

Testing segments for the interactive segments are non-technical in nature, and represent common appointments as identified by the employer referral agencies in Canada. The test developers contacted the interpreter referral agencies using a standard set of interview questions to glean the information needed. Employers indicated the ten most common interpreting assignments and the range of topics addressed in those settings. Additionally, employers suggested topical areas they perceived as relevant and fair for a national test. Each province provided employment data and the responses were gathered in personal telephone interviews or by electronic communication. Scenarios such as a work place interview, a counseling appointment, or a parent-teacher interview were all identified as appropriate for a national test. The segments that are narrative in nature are also non-technical, and representative of formal presentations offered by Deaf and non-deaf professionals. Topics were identified that are representative of ones that interpreters frequently interpret in their community work. These include issues such as human rights and access, language and literacy issues among others. This approach to establishing content validity is one supported by Berger and Simon (1995) as reported by Clifford (2001), and is further described as simulating authentic language use with fidelity (Khattri et al. 1998).

Workshop participants are also provided with a third workshop that is offered in self-directed study format. The materials focus on test-taking strategies and managing test anxiety. The materials also address implementing preparation into the work: working from presentation outlines; taking notes during consecutive interpreting, if needed; using a blend of both consecutive and simultaneous interpreting during a dialogic segment, and dealing with the two-dimensional aspects of video testing materials.

Test of interpreting

This phase involves the actual interpreting performance evaluation. New test materials were developed in 2004, consisting of a Version A, and a Version B. Cam McDermid, one of the test developers, contacted all major referral services in Canada, and solicited suggestions for topics and settings to be used. Once the topic areas were confirmed, the content was balanced so that each test version had a similar topic, but was filmed using different participants. The participants who were filmed were prepared for the scenario at the same time and had opportunities to review the topic areas each would cover, thus ensuring the content and level of complexity was very similar between the presenters.

In addition, the participants who were filmed for the test segments were selected from provinces such as Newfoundland, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia to reflect the range of linguistic diversity across the country. An advisory committee of experts vetted names of participants who would be suitable for the videotaped scenarios, as well as appropriate topics. It should be noted that while the scenarios were created for the purpose of the TOI, the interactions were not scripted. Participants relied on their preparation and backgrounds to allow for the use of natural language throughout the scenarios. This contributed to the authenticity of the test samples and further supported content validity principles (Clifford 2005).

Test materials were filmed in a professional studio. They included signed or spoken introductions of the participants, as well as the introduction of a target audience member, to assist the candidate in envisioning the individual for whom she was interpreting.

Candidates interpret a fifteen-minute narrative from ASL into English; another narrative from English into ASL, also fifteen minutes long; and two interactive segments, which they choose out of a possible three topics (see previous description of potential topics). Candidates can interpret the segments in any order they choose, although the narratives must be interpreted as a unit. Prior to taking the test, candidates are granted access to presentations by the same narrative presenters speaking on different topics, in order to familiarize themselves with speakers prior to taking the actual test. This simulates the kind of preparation an interpreter in the field of signed language interpreting might engage in when interpreting in authentic real-world assignments. These presentations are available online through a password-protected website controlled by AVLIC. In addition, candidates are provided with outlines of the narrative presentations that will be on the test, so that they can conduct research and prepare for the segment as they would for some real-life interpreting assignments. This type of preparation is more commonly used by signed language interpreters working in formal settings with narrative discourse, and AVLIC had the desire to replicate this best practice strategy in their testing process. (This process is not available for the interactive segments, where the interpreters are much more able to rely on discourse strategies such as understanding adjacent pairs, and their own real-life experiences in similar settings to understand the messages of the participants.) Candidates may bring pen, paper, and presentation outlines into the testing room with them. Candidates may also take rest breaks between segments, totaling no more than 30 minutes of break time. Additionally, test takers are allowed to pause each of the segments and/or rewind to the beginning of a utterance, to a maximum of four times per segment. This is offered as a way to replicate a real-world interpreting situation, where the interpreter can often ask the speaker to pause or to repeat a concept.

A significant change to the TOI process is that candidates may choose to submit a videotape of local work that can be considered by raters as supplemental evidence of successful previous performance. This video consists of both an ASL to English and English to ASL narrative presentation, each no longer than 15 minutes in length. The linguistic register must be within the consultative to formal range, and the audience and context must be described in writing. The interpreting sample must have been done within the six months prior to taking the TOI. This change addresses the concern from some AVLIC members that they do not perform well on videotaped scenarios, and that the work they do in their local community is more reflective of their skill level. Message Equivalency raters will only consider the work if the performance on the TOI is borderline, and raters are seeking further input prior to arriving at a decision. To date, only one test taker out of thirty-five has used this option.

Piloting of new materials

Once the test materials were completed in 2004, they were pilot tested on a group of sixteen individuals. Eight of the test takers were COI holders who provided their feedback on the new materials, and eight were non-certified interpreters who were given the opportunity to gain certification during the pilot study.

One interpreter was granted certification from the pilot test. As a result of feedback from the pilot, two interactive segments were switched between Versions A and B in order to balance the level of complexity and to balance the gender representation per the two test versions. In addition, one presentation that had been used as the pre-test practice material became a test segment, with the former test segment becoming the practice tape. This decision was made based on the construction of the segment, and the coherence demonstrated within the monologue.

Issues of test validity and reliability

Like RID, AVLIC strives to maintain adherence to nationally recognized testing industry standards of validity, reliability, and equity (see www.itlaonline.com). As a result, an independent psychometrician serves as a consultant to AVLIC, and interpreting/test development consultants were retained to revise the test. The Test Development Consultants included three educators who had considerable experience in the construction of interpreting assessments. Cam McDermid had played a key role in the creation of the testing system used by an Ontario Interpreting Agency and is an established interpreter educator with a strong research background. Karen Malcolm had also been a consultant on the development of interpreter screening tools in British Columbia. In addition to being an established educator, she has had a lengthy involvement with AVLIC and has been the rater trainer and facilitator with the previous test model. Debra Russell had extensive experience creating interpreter assessment tools, both nationally and provincially, and brought a strong research and teaching background to the team. The experts also assembled a group of expert advisors who could offer guidance at major decision points. These advisors were Deaf and non-deaf people who had experience with assessment processes and were well respected in their provinces for their knowledge of key issues affecting interpreting. Finally, an ongoing psychometric analysis is performed on the written and performance tests to ensure that they remain valid and reliable instruments for measuring an interpreter's abilities. For example, AVLIC requires that data be organized on success/fail rates, inter-rater reliability, and interpreter participant feedback on the preparation workshops.

Raters for the performance test are trained to identify skills that meet or exceed basic professional-level interpreting standards. This training took place over a three-day period, and each time the raters are brought together to assess candidates, they review the standard, criteria, and marking forms prior to beginning their rating. The facilitator also has them examine previously rated tests, ensuring that all three raters are consistently determining pass and fail performances. Once the raters have reached the point that they are arriving at a common decision 95% of the time, the rating of the actual test samples begins. Psychometric procedures have been established to monitor the consistency of inter-rater reliability and comparison of test results across test offerings and across both versions of the test. Another process that has been established, so that all candidates are treated fairly, is the requirement that raters declare any perceived or real conflict of interest with a testing candidate. Once aware of this declaration of the potential of a conflict of interest, the rating facilitator ensures that the rater with the perceived conflict is solely addressing the criteria in their comments.

Test validity for the Test of Interpretation was approached in terms of content, and construct validity. Bachman (1990, 2005) argues that the field of language testing has largely relied on two distinct approaches to language testing: construct-based and task-based approaches to language testing. In his view, test design must employ both approaches in order to address the problems of generalizability and extrapolation. Construct-based approaches suggest that an instrument is valid if is actually measures what it was designed to measure, and allows raters to make inferences about the competencies targeted. In terms of content validity, the test measures interpreting abilities across a range of scenarios that were determined to be typical and common interpreting situations found across all provinces. Content validity is also a non-statistical judgment, but requires a more detailed examination of the test. As indicated in an earlier section of this chapter, the content of the test segments was created based on community consultation with interpreters and consumers of interpreting services, along with interpreter referral agencies, in order to plan test scenarios that are realistic and reflect the broad range of settings where ASL-English interpreters typically work. ASL-English interpreters in Canada are employed in diverse settings, serving Deaf consumers in all aspects of their lives. This includes interpreting for medical appointments, employment related matters such as job interviews, staff meetings, on-going staff development, and educational contexts from Kindergarten to Grade Twelve. It also involves post-secondary settings, theatre productions, travel and tourism programs, etc. Basically, interpreters working with signed languages find themselves working from birth to death with people who are deaf or have family members whom are deaf.

The Test Developers contacted each of the community-based interpreter referral agencies across Canada, and gathered data about the type and frequency of interpreter assignments. The data were examined and all assignments that were unique to a region, or were highly technical or industry specific, were not considered for inclusion as possible test scenarios. A slate of twenty possible scenarios was circulated to the Deaf and non-deaf people serving on the test advisory committee. Examples of the kinds of scenarios considered included: a routine doctor's visit, a parent-teacher interview, a staff development presentation on retirement planning, etc. This data-driven approach to determining content area is a similar approach to one described by Angelelli (2007) in the development of a test for health care interpreters. The scenarios chosen were verified by the panel of experts and also the Evaluations Committee, confirming that the scenarios are indeed representative of those commonly experienced by community interpreters.

Construct validity refers to the degree to which inferences can be legitimately made from testing performance to the theoretical constructs on which the criteria is based (Trochim & Donnelly 2006). It requires that there be a theoretical model of a trait, and that a series of studies supports the meaningful existence of the traits. In the area of signed and spoken language interpreters, a number of studies support the criteria used to score the test, and verify that interpreters work in settings requiring simultaneous interpreting of narratives, and simultaneous and consecutive interpreting within dialogic work settings (Angelelli 2007; Cokely 1992; CIT 2008; Napier 2002; Roy 2000; Russell 2002). Pöchhacker (2004) provides a comprehensive review of many of the models that have been published in both signed and spoken language interpreting fields of study. He reminds us that models can be used for various kinds of inquiry, from theorizing about a given phenomenon, to describing and explaining some aspect of a phenomenon, and to predicting the occurrence of the phenomenon, for example in a testing context. While all models are by their nature incomplete representations, in the field of interpreting, models can be useful in capturing some of the complexities of communication and discursive processes.

Pöchhacker (2004) identifies that over the course of history, our field has been introduced to models that have drawn on anthropological views of interpreting, socio-professional and institutional functions of interpreting, interactional aspects of interpreting, and interpreting as a communicative process, which gave rise to discursive and cognitive models of interpreting. Further, Hatim and Mason (1997) have drawn our attention to the concepts of text and discourse processing as a framework for analyzing interpreting. Hatim and Mason describe three key concepts of discourse processing as most significant when providing simultaneous, consecutive, and liaison interpreting, namely text, structure, and context. Specifically in the field of signed language interpreting, Cokely (1992) developed a model that addresses the modality of input and output (spoken or signed) and offers a sociolinguistic and cultural model. The model is based on seven major stages of cognitive processing, from initial message reception to production of the target language rendition.

When one reviews the criteria and rating forms that AVLIC has developed for the Test of Interpretation, it is clear that they have chosen constructs based on discourse analysis and have been influenced by several of the theorists described earlier, and have drawn on Cokely's sociolinguistic model of interpreting when creating the rating forms. Clifford (2001) posits that one way in which discourse theory can benefit interpreting assessment is in its explanation of meaning. He emphasizes that we must learn to see interpreting as a form of discourse, incorporating both lexico-semantic concerns and the socio-cultural context of delivering a target language message. AVLIC has chosen to draw upon the usefulness of discourse models to guide the raters in determining how well candidates demonstrate the traits on the simulated interpreting segments of the test. An example of one of the traits from the criteria described in Appendix One may serve to illustrate this point. The trait requires that interpreters demonstrate the use of overall discourse strategies that result in a coherent text. This is further defined as the use of appropriate and opening and closing comments, the essential elements of the source language message, found in both the language and the contextual variables of the setting and interaction, and topic transition and topic maintenance strategies. Lastly, when addressing matters of reliability, it is important to take steps to reduce the risk of error inherent in assessing professional interpreters (Resnick & Resnick 1992). That is, an interpreting test is a sample of performance, and based on that performance, the raters determine whether the candidate's interpreting skills meet the standard. There is always a risk that the sample performance does not give an accurate picture of the person's actual skills. Berger and Simon (1995) as reported by Clifford (2001) suggest that the adherence to four principles of evaluation be used to minimize the possibility of error – validity, reliability, equity, and utility. Clifford (2001) has added a fifth related principle, that of comparability. AVLIC has considered each of these principles carefully, planning for content and construct validity so that raters are able to make inferences about the target competencies. The principle of content validity was addressed by using a data-driven process to determining test content. The theoretical constructs guiding the criteria and test traits are grounded in research within the field of interpreting, and are embedded in discourse theory. Reliability data is gathered for each round of testing, ensuring stable results from one administration to another. Testing results over the years of the previous test have shown a great deal of stability, with a high level of consistency of the pass rate across a ten-year span. The new test data will be tracked by the AVLIC Canadian Evaluation System and monitored by the contracted psychometrician. The conditions of offering the test do not vary, and the pool of Deaf proctors who administer the test have been well trained, as they are guided by explicit step-by-step instructions in both English and in American Sign Language on how to structure the testing environment. All raters use the same rating forms, and there is a common understanding of the criteria. All of these features contribute to reducing the risk of error in the area of reliability. The equity principle is the one that AVLIC has addressed by ensuring that the rating facilitator manages the rating process, and consistently ensures raters are following the criteria and that personal biases do not have a place in the rating discussions. Utility refers to assessments that may be valid, reliable, and equitable, but high cost or elaborate procedures may prevent the test from having broad use. This is an area that AVLIC will continue to monitor, given the costs associated with the test preparation workshops and rating processes, and the limited resources found within the organization. By comparability, Clifford (2001) means that the assessment is administered consistently, there is a common understanding of the test criteria, and the performance is evaluated fairly. Again, AVLIC has ensured that its policies and practices support a consistent administration of the test, the raters have a common understanding of the criteria, and performances are rated fairly.

Certificate maintenance program

Certificate Maintenance is the final stage of the certification model. Maintenance requirements for certified members have always been to abide by the Code of Ethics and Guidelines for Professional Conduct, and to hold consistent active membership in AVLIC. In addition, AVLIC members have expressed the desire for a mechanism to ensure that interpreters who achieve certification continue to demonstrate the skills and abilities required for a COI interpreter. As a step towards developing a Certificate Maintenance Program, COI interpreters are required to document their professional development activities when they renew their membership annually. This process will be followed for three years, and the data gathered will be used as the basis for establishing requirements for certified members to follow in order to maintain certification. There has been some discussion about the need for a separate testing body, similar to what teachers have. However, given the small numbers within our field, an arms' length organization is not possible at this time. Ongoing membership is one way to demonstrate commitment to professional growth, and to abiding by a professional code of ethics and guidelines for professional conduct. Membership also ensures consumers and members alike have the opportunity to resolve issues through a formal set of processes known as the AVLIC Dispute Resolution Process. However, the organization also notes that holding membership does not ensure that an interpreter's skills are maintained or even enhanced, and hence the need for development of a certificate maintenance system that will require documentation of professional development activities.

Revised rating process

Suggestions had been made that the rating of English and ASL separate from ME might not be necessary, since it would not be possible to achieve message equivalency without using grammatically correct, complete ASL and English. Therefore, in the Message Equivalency rating that took place in 2002, a pilot test was conducted with the ME raters. The raters evaluated eight tapes from previous years. The raters' task was to rate the tapes and determine the candidate's result, which could have included: candidate exits at English or ME, or candidate demonstrates a passing performance. While there were several instances where raters stated that a tape would have exited at ME when it actually exited at English, there was complete consistency with previous test results, in that no candidates who had previously exited the system prior to certification were deemed to be passes. On this basis, the Evaluations Committee decided to eliminate the English domain for rating.

The committee also considered eliminating the ASL domain; however, Deaf representatives on the committee raised concerns. Given that ASL is a first language for very few ME raters (only those who grew up with Deaf parents have ASL as a first language), Deaf community members were not convinced that raters would accurately and consistently recognize the ASL standard. Therefore, a motion at an annual general meeting was passed, mandating ASL rating to continue for a three-year period during the implementation of the new testing process. This will be re-evaluated in 2009 at the end of the three years.

ASL rater training

ASL raters are all Deaf individuals who are recommended by the Deaf representatives on the Evaluations Committee. The ideal raters are experienced users of ASL and have some background in analyzing ASL. Many of them are ASL teachers or tutors, and they are all familiar with the work of interpreters. Variation in geographical location, as well as a mix of male and female raters, is sought. Raters commit to five years of involvement, and are paid an honorarium for each tape rated.

A team of consultants conducted the initial ASL rater training. Raters were oriented to the overall process by watching an introductory video in ASL presented by the Deaf rater trainer. They were introduced to the criteria and were able to clarify any confusion regarding the criteria through conversations with either of the rater trainers. Because this training was conducted at a distance rather than face-to-face, these conversations were held either via email, or by videophone.[11]

Raters were presented with video samples of the pilot test tapes, the rating forms, and criteria, and are asked to rate the test tapes. Results were shared, and then raters debriefed with one of the consultants. Gradually, as raters continued to watch performances and rate tapes, their determinations became more and more similar. AVLIC does not expect 100% agreement, because raters are evaluating language in a holistic fashion, while relying on the criteria to guide their decisions. The requirement that two raters agree on the test results balances the slight variations that arise in rater decisions. The CES office also monitors rater results in both domains, and tracks if one rater is consistently rating differently than the other raters: differences may indicate the need for re-training or even for the eventual removal of that particular rater if they are unable to rate per the criteria and standard established by AVLIC.

Message equivalency rater training

Initially, a call for raters was sent out to the AVLIC membership. All Message Equivalency raters needed to be COI holders. Message Equivalency rater training was conducted face-to-face, with one consultant training the group.

Raters were first given the opportunity to familiarize themselves with the test materials by viewing the narratives and interactive segments. In addition, each rater was given their own performance test tape to review prior to attending the training. This activity has been very useful because by noting some of the miscues that occur in their own performances, raters learn to recognize the errors and skews that exist in interpreting, particularly in test situations. There is a danger that, as raters become more and more familiar with the source stimulus materials, their expectations for the interpreted renditions become higher and higher, which could lead to a gradual unwanted raising of the bar for a pass. Watching the errors in their own performances, which were deemed passes, reminds them that errors are to be expected, and do not automatically lead to failing results being determined. While the potential danger exists that raters might expect candidates to adopt the same successful strategies they themselves employed in producing the interpreted rendition, the reality is that raters have viewed many candidates interpreting the

[11] Videophones employ equipment that works through Internet access so that two people can converse in ASL, viewing each other's image either on a computer screen or on a TV. In this way, conversations are conducted in a Deaf person's first language.

same source tape, and are well aware of the variety of successful means of interpreting the source, so that they do not expect to see the same strategies used.

Raters review the criteria and ensure that they understand them. Then, they view performances that have been deemed successful and unsuccessful in the past. They practice recording their comments, voting individually and then discussing the results with the other raters. They view a minimum of two tapes, and then are offered the opportunity to view one more tape in order to feel confident about the standard. Once raters feel ready, they begin to rate actual test tapes.

There are six Message Equivalency raters, but only three come together to evaluate the tapes for each test offering. The decision of the three raters is final. However, all the test tapes are sent out to the other Message Equivalency raters to view on their own, so that they continue to stay current in their application of the criteria in assessing performances. The Evaluations Committee alternates the raters who participate in the face-to-face rating each year, to assist raters in staying current.

Message Equivalency raters also complete a conflict of interest form if they believe there is any real conflict, or the potential for a perceived conflict, whether positive or negative. In this way, the CES can track potential conflicts or any perceived conflicts, and can examine individual raters' votes. Nonetheless, the three ME raters still need to come to agreement and the facilitator continues to keep them focused on responding to the criteria and the performance at hand.

Reporting results to test takers

Candidates typically receive their written results within four to six months of taking the Test of Interpretation. Candidates who have not been successful are provided with written feedback about their performance. The results are considered final and there is no appeals process available except in instances of administrative or technical concerns.

However, unsuccessful candidates do have the option of viewing their performance tape within a six-month period after receiving their results. They arrange to have a guarantor who has custody of the tape for the entire period, and can arrange for a diagnostician to view the tape as well, in order to apply the written feedback received to actual examples demonstrated on the test tape.

Comparison of test results

The first offering of the new TOI occurred in 2007. The Evaluations Committee is collecting statistics regarding pass/fail rates and rater validity and reliability, but more data is required before being able to make any kind of comparison between the old and new systems. In keeping with appropriate measurement conventions, a psychometrician has advised AVLIC on the types of record-keeping systems and statistical analyses required to monitor the inter-rater reliability and the pass rates of both versions of the Test of Interpretation (Clifford 2005). Data will continue to be analyzed in the same manner that it was prior to the implementation of the new Test of Interpretation. Given the recent implementation of the new Test of Interpretation, data was not available at the time of the writing of this chapter; however, the AVLIC Evaluations Coordinator (Monique Bozzer, personal communication, June 8, 2008) has indicated that the pass rate appears to be higher with the new version, and feedback offered from workshop participants indicates that the workshop content is enhancing their interpreting abilities and confidence to take the test.

Scoring procedures

One of the decisions made in revising the Test of Interpretation was to retain the criteria and overall standard from the original test, ensuring the standard was not lowered for the new test. Our consultations with interpreters and Deaf consumers, in particular, supported the existing standard, and stressed the need to find other ways to support candidates in passing the test, as opposed to lowering the performance standard. The scoring is not based on counting errors or omissions as is common on some translation exams (ATA 2008), or on a checklist, as some other tests are, for example the Educational Interpreter Performance Assessment (EIPA) or the Ontario Interpreting Services Screening Tool; rather, the raters are trained to identify the criteria, and make evidence-based decisions about the consistent representation of those features across all four test segments. AVLIC, in collaboration with the test development experts, reviewed quantitative and qualitative scoring approaches used on other employment screening tools and tests (Barik 1971; OIS 2005; Taylor 1993). While each approach has strengths and advantages, drawbacks are also evident. The disadvantage of counting errors, for example, is that it limits raters to assessing the interpretation at a lexical level, versus looking at the interpreting from a discourse-based approach (Clifford 2001; Pöchhacker 2004). It also means that the agreed-upon test standard is based on a norming pool that may or may not be representative of the interpreting community as a whole. For example, if the group of interpreters selected as the norming sample was comprised of six very experienced certified interpreters, with four of them having ASL as their first language, the number of miscues in their performance would be expected to be far fewer than for interpreters with less experience and later language acquisition of ASL. After reviewing the options, AVLIC, in conjunction with the test developers, made the decision to approach the rating and scoring from a qualitative approach that suits the examination of interpreting data.

The Message Equivalency facilitator is responsible for recording each rater's initial vote, final vote, and the criteria-based examples. These examples provide the evidence needed to reach consensus among the raters, and ensure that the raters are consistently referring to the criteria and the standard of performance necessary to achieve a pass. After a sample has been determined as a pass, the Message Equivalency facilitator views the ASL rater's decision, and confirms that the person has also passed the ASL domain.

Considerations arising from the new testing model

The new testing model is entering its second year since implementation, and AVLIC and its members are continuing to learn how this model will impact practice. AVLIC is committed to supporting interpreters in their pursuit of national certification, and has expended considerable resources on the four-phase model. AVLIC is constantly seeking feedback from workshop participants, workshop facilitators, ASL and interpreting specialists, ASL, and Message Equivalency raters and facilitators, as the organization implements continuous learning and improvement practices.

Feedback from participants in the TOI preparation workshops

Some of the themes that have emerged from the feedback from TOI preparation workshop participants include:

a. Using research to inform practice is still new for some interpreters and interpreter education programs.
b. The pre-readings on discourse analysis that were distributed prior to the workshops were not part of the participants' Interpreter Education Programs and are therefore intimidating.
c. Having direct and honest feedback about their use of ASL and interpreting skills is very helpful in setting a plan of action and determining readiness to take the test.
d. Preparation for the test is a much larger and longer process than anticipated.

Feedback from participants in the TOI workshop facilitators and specialists

Some of themes that have emerged from the feedback from workshop facilitators and ASL and interpreter specialists include:

a. Interpreters often lack a full understanding about ASL grammatical features and overestimate their abilities to use the language.
b. Interpreters often struggle to see their interpreting work through an accurate lens, either being too critical of their work or unable to see the extent of the problems in it.
c. The workshop materials are challenging for interpreters who have not been exposed to current studies and readings.
d. There is a need for interpreter education programs to examine their curricula per discourse analysis principles and practices.

Accessing resources

What is also interesting to note is that some test takers are not taking advantage of the support offered in the new testing process. For example, prior to taking the TOI, candidates are given access to narrative presentations given in ASL and English by the same presenters they will view on the test. While these are on different topics, the presentations provide the interpreter with a sense of the presenter's style, language use, organization and pacing, and are excellent preparation for the test. These samples are online, and after the first test offering, the tracking mechanism showed that some test takers chose not to view these samples. These same test takers were not successful on the Test of Interpretation. This invites the question: Is it that interpreters in their day-to-day work are not engaging in preparation work, or are they over-confident in their abilities to manage the test environment? There are other AVLIC members who believe that the preparation materials, in particular the narrative presentations for viewing prior to the test, as well as the one-page presentation outline, are giving the test taker too many advantages. However, the early test results are not showing significant gains in pass rates, and there is some preliminary data revealing that some candidates did not choose to access the preparation resources, but rather went into the testing "cold," which may then contribute to the fail rate. So while the preparation materials are designed to be helpful to the test-taker, not all interpreters are choosing to use them. AVLIC views the offering of preparation for the narrative presentations as creating an authentic testing environment that mirrors best practices in community interpreting work.

Test utility and resources

One of the future directions that AVLIC will need to consider is the resource base needed to support this comprehensive test model. The resources required to deliver the workshops are significant for a non-profit organization, and creative solutions will need to be found if this phase of the model is to be offered in its current form. Additionally, AVLIC will need to examine the costs of delivering the testing system, given the relatively small numbers of test takers per year. There is a risk in pricing the test in a way that supports the management of it but makes it prohibitive for test takers.

It will also be interesting to watch the development of the Certificate Maintenance Program and to gauge membership endorsement of the process. The challenge for AVLIC will be to create a model that allows for the tracking of authentic learning experiences that demonstrate that a certified member is remaining current in the field.

Limitations and future research

As AVLIC continues to work on ensuring their testing processes are appropriate, it will be important that some of the limitations are addressed. For example, there is a need to revisit the training provided to ASL raters in order to obtain the same inter-rater reliability levels that have been reached with the ME raters. By addressing the inconsistency in grader training, AVLIC will have taken an important step in ensuring testing principles, such as validity and reliability, are addressed. Another limitation of the current test model is that, while the criteria is well-understood by the ME raters, there are opportunities to further refine the tools used by the raters to include more detailed marking rubrics than the ones presently used. The rubric could assign score points in addition to using the scale of "consistently demonstrated, inconsistently demonstrated, and not demonstrated". The current system identifies the significant trait or dimension to be assessed and the level of performance required; however more detailed descriptions should be developed to reflect whether and to what extent the key requirements of the performance have been demonstrated. Further, it would be helpful for AVLIC to employ the services of a testing and measurement specialist when considering any changes to the existing or future testing processes. Finally, there are questions to be explored about the nature of consensus grading in the message equivalency domain, and whether this in fact constitutes a "best practice" approach. The consensus approach seems to have emerged early in the development of the testing process, and may reflect cross-cultural sensitivity in working with a minority linguistic community that operates as a collective community, preferring this form of decision-making.

An additional concern for AVLIC is the need for continual monitoring of the knowledge and skills of the test proctors. Given that the Test of Interpretation is offered once per year, AVLIC will need to ensure that proctors are adhering to policy and procedures that protect the integrity of the test. Finally, AVLIC will be challenged to continually update the Written Test of Knowledge, given the increased research about interpreting available to us. The information can quickly become dated, requiring regular updating and piloting prior to releasing new versions of the WTK.

As Bachman (2007) has pointed out, language testing practitioners may be held accountable for the decisions based on the basis of a test, and likewise, AVLIC bears a burden of accountability. As such they will need to be diligent in collecting evidence to support their testing decisions. One area of focus will need to be on the established constructs, and how they reflect the balance of test focus on language and interpreting abilities, tasks, and interactional skills required for success on the Test of Interpretation. As a practice profession, AVLIC will need to examine the impact of their decisions on stakeholders, while tracking the reliability of the scoring system, the validity of the decisions, and the fairness and appropriateness of the decisions that are made.

AVLIC will continue to engage its members and stakeholders in dialogue about the new testing model, and implement continuous learning practices that will build on the responsive and comprehensive nature of the testing model. As the fields of signed language and spoken language interpreting interact more frequently, there are opportunities for shared research agendas to address some of the complexity of testing language and interpreting abilities within the broader social context. What are the most effective testing and assessment approaches? How will the field expand to include specialized exams for settings such as medical and legal contexts? There are also opportunities for research to explore the gap between the skills and knowledge of interpreters graduating from interpreter preparation programs at the college and university level, and the skills and knowledge required to obtain national certification. Finally, research is also needed to explore the impact of national certification processes at the local level, and how such decisions may impact stakeholders (positively or negatively), and impact access to quality interpreting services.

Summary and conclusion

The previous sections have discussed the past and current testing models that lead to certification for ASL-English interpreters in Canada. While there are similarities to the testing processes used in Australia and the United States, the model implemented by AVLIC has some unique features designed to support test takers in achieving certification. The four-phase model requires a commitment to learning on the part of test candidates, beginning with the Written Test of Knowledge. The online nature of the WTK testing allows for immediate results. The Test Preparation Workshops offer candidates personal feedback about their interpreting work in both narrative and constructed dialogue settings, while also introducing them to current research in the field. The online preparation materials for candidates taking the Test of Interpretation provide test takers with additional strategies to prepare for the narrative test segments. Finally, the Certificate Maintenance Program will provide consumers and interpreters alike with an approach to ensure that certified members continue to demonstrate evidence of the skills and knowledge required for a Canadian certified interpreter.

The strengths of the current model reflect the development of test content that is situated in authentic discourse and grounded in research from the field and in the work contexts that interpreters find themselves in while working in Canada. The test criteria draw on theoretical models of sociolinguistic and discourse analysis, and evidence-based research that highlights the tasks and sub-tasks of linguistic use required by interpreters.

As AVLIC continues to work on ensuring their testing processes are appropriate, it will be important that the limitations discussed in the previous section are addressed. As the fields of signed language and spoken language interpreting interact more frequently, there are opportunities for shared research agendas to address some of the complexity of testing language and interpreting abilities within the broader social context.

While the information presented in this chapter reflects the current model of the testing interpreters in Canada, the resolution of issues raised here, along with the constantly evolving field of interpreting will inevitably lead to further refinements of the certification model.

References

American Translators Association. 2008. *ATA Certification Program.* Retrieved June 27, 2008 from http://www.atanet.org/certification/aboutexams_overview.php

Angelelli, Claudia V. 2007. "Assessing Medical Interpreters: The Language and Interpreting Testing Project." *The Translator* 13(1): 63–82.

Angoff, William H. 1971. "Scales, Norms, and Equivalent Scores." In R. L. Thorndike (ed), *Educational Measurement,* 508–600. Washington, DC: American College on Education.

Association of Visual Language Interpreters of Canada. 2008. *Evaluation Information.* Retrieved June 23, 2008 from http://www.avlic.ca/services.php?evaluation

Bachman, Lyle. 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, Lyle and Palmer, Adrian. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Baume, David and Yorke, Mantz. 2002. "The Reliability of Assessment By Portfolio on a Course to Develop and Accredit Teachers in Higher Education." *Studies in Higher Education* 27(1): 7–25.

Barik, Henri. 1971. "A Description of Various Types of Omissions, Additions and Errors in Translation Encountered in Simultaneous Interpretation." *Meta* 16: 199–210.

Berger, Marie-Josée and Simon, Marielle. 1995. *Programmation et évaluation en milieu scolaire: notes de cours, PED 3713 95/96*, Université d'Ottawa, Unpublished manuscript.

Canale, Michael. 1983. "From Communicative Competence to Communicative Language Pedagogy." In J. Richards and R. Schmidt (eds), *Language and Communication*, 2–27. London: Longman.

Canale, Michael and Swain, Merrill. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics* 1: 1–47

Clifford, Andrew. 2005. "Putting the Exam to the Test: Psychometric Validation and Interpreter Certification." *Interpreting* 7(1): 97–131.

Clifford, Andrew. 2001. "Discourse Theory and Performance-Based Assessment: Two Tools for Professional Interpreting." *Meta* 16(2): 365–378.

Clifford, Andrew. 2003. *A Preliminary Investigation Into Discursive Models of Interpreting as a Means of Enhancing Construct Validity in Interpreter Certification.* Unpublished doctoral dissertation, University of Ottawa.

Cokely, Dennis. 1992. *Interpretation: A Sociolinguistic Model.* Burtonsville, MD: Linstok Press.

Conference of Interpreter Trainers. 2008. *Interpreter Education Standards.* Retrieved June 23, 2008 from http://www.cit-asl.org/ccie.html.

Davis, Jeffrey. 2005. Code Choices and Consequences: Implications for Educational Interpreting. In M. Marschark, R. Peterson, and E. Winston (eds), *Sign Language Interpreting and Interpreter Education,*112-141. New York: Oxford Press.

Gadbury-Amyot, Cynthia Holt, Lorie Overman, Pamela and Schmidt, Colleen. 2000. "Implementation of Portfolio Assessment in a Competency-Based Dental Hygiene Program." *Journal of Dental Education* 64 (5): 375–380.

Grice, Paul H. 1981. "Presupposition and Conversational Implicature." In P. Cole (ed), *Radical Pragmatics*, 167–182. New York: Academic Press.

Hatim, Basil and Mason, Ian. 1997. *The Translator as Communicator.* London and New York: Routledge.

Halliday, Michael. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* London: Edward Arnold.

Hudson, Thom. 2005. "Trends in Assessment Scales and Criterion-References Language Assessment." *Annual Review of Applied Linguistics* 25: 205–227.

Ingersoll, Gary and Scannell, Dale. 2002. *Performance-Based Teacher Certification: Creating a Comprehensive Unit Assessment System*. Golden, CO: Fulcram Publishing.

Johnson, Robert and Squire, Jan. 2000. *Setting the Standard for Passing Professional Certification Examinations*. Retrieved Sept. 16, 2008 from http://207.36.165.114/FMAOnline/certifications.pdf

Khattri, Nidhi, Reeve, Allison and Kane, Michael. 1998. *Principles and Practices of Performance Assessment*. Mahwah, NJ: Erlbaum.

Lombardi, Judy. 2008. "To Portfolio or Not to Portfolio: Helpful or Hyped." *College Teacher,* 56(1): 7–10.

Nystrand, Martin, Cohen, Allan and Dowling, Norca. 1993. "Addressing Reliability Problems in the Portfolio Assessment of College Writing." *Educational Assessment* (1)1: 53–70.

Napier, Jemina, McKee, Rachel and Goswell, Della. 2007. Sign Language Interpreting: Theory and Practice in Australia And New Zealand. Sydney, NSW: The Federation Press.

Napier, Jemina. 2005. "Training Sign Language Interpreters in Australia: An Innovative Approach." *Babel* 51(3): 207–223.

Napier, Jemina. 2004. "Sign Language Interpreter Training, Testing, and Accreditation: An International Comparison." *American Annals of the Deaf* 149(4): 350–360.

Napier, Jemina. 2002. *Sign Language Interpreting: Linguistic Coping Strategies*. Coleford, England: Douglas McLean.

National Authority for the Accreditation of Translators and Interpreters. 2008. *NAATI Accreditation Levels.* Retrieved June 18, 2008 from http://www.naati.com.au/at-accreditation-levels.html

National Authority for the Accreditation of Translators and Interpreters. 2008. *NAATI PI Assessment Format.* Retrieved June 18, 2008 from http://www.naati.com.au/at-testing-procedure-I.html

Ontario Interpreting Services. 2005. Registration Process Overview. Toronto: Canadian Hearing Society.

Padden, Carol and Humphries, Tom. 1988. Deaf in America: Voices from a Culture. Cambridge, MA: Harvard University Press.

Pochhacker, Franz. 2004. Introducing Interpreting Studies. London, UK: Routledge.

Registry of Interpreters for the Deaf. 2008. *CI/CT Certification Guidelines.* Retrieved June 23, 2008 from http://www.rid.org/education/testing/index.cfm/AID/87

Registry of Interpreters for the Deaf. 2008. *NIC Certification Guidelines.* Retrieved June 23, 2008 from http://www.rid.org/education/testing/index.cfm/AID/86

Registry of Interpreters for the Deaf. 2008. *NIC Performance Criteria.* Retrieved June 23, 2008 from http://www.rid.org/education/testing/index.cfm/AID/86

Registry of Interpreters for the Deaf. 2008. *NIC Interview Domains and Rating Scales.* Retrieved June 23, 2008 from http://www.rid.org/edcation/testing/index.cfm/AID/86

Resnick, Lauren and Resnick, Daniel. 1992. "Assessing the Thinking Curriculum: New Tools for Educational Reform." In B. Gifford and M. O'Conner (eds), *Changing Assessments: Alternate Views of Aptitude*. Boston, MA: Kluwer Academic Publishers.

Roy, Cynthia. 2000. *Interpreting as a Discourse Process*. Oxford England: Oxford University Press.

Russell, Debra. 2002. *Interpreting in Legal Contexts: Consecutive and Simultaneous Interpretation*. Burtonsville, MD: Linstok Press.

Schutz, Sue, Angrove, Carrie and Sharp, Pam. 2004. "Assessing and Evaluating Reflection." In C. Bulman & S. Schutz (eds). *Reflective Practice in Nursing* (3rd ed.), 47–72. Oxford: Blackwell.

Taylor, Marty. 1993. *Interpretation Skills: English to American Sign Language*. Edmonton, AB: Interpreting Consolidated.

Trochim, William and Donnelly, James P. 2006. *The Research Methods Knowledge Base* (3rd ed.). Mason, OH: Atomic Dog Publishing.

Wadensjö, Cecilia. 1998. *Interpreting as Interaction*. London; New York: Longman.

Wilkerson, Judy and Lang, William. 2003. Portfolios, The Pied Piper of Teacher Certification Assessments: Legal and Psychometric Issues. *Education Policy Analysis Archives* 11(45). Retrieved from http://epaa.asu.edu/epaa/v11n45/

Appendix 1

Association of Visual Language Interpreters of Canada

Canadian Evaluation System

avlic

Association of
Visual Language Interpreters
of Canada

Test of Interpretation

American Sign Language Criteria

Introduction

Within the domain of American Sign Language, the Canadian Evaluation System (CES) examines TOI candidates' ASL production. Unlike the Message Equivalency domain, the ASL domain focuses on production of ASL as the target language, not interpreting performance. To a certain degree, the content of the source text is irrelevant. Raters are asked to view candidates' ASL and determine whether this sample of their work meets the TOI standard. The actual source language message is not available to the raters for reference.

AVLIC Evaluation Committee 2007

I DISCOURSE STRATEGIES

1. Standard: Overall discourse strategies used result in coherent text.

a. Appropriate use of opening/closing comments
b. Essential elements of meaning with adequate supporting detail
c. Appropriate use of topic transition and topic maintenance strategies
   – exhibits strategies for comparing and contrasting ideas
   – references within the text to previously introduced information
d. Avoids restatement of ideas that do not add meaning to the text

II FORM

1. Standard: Overall sign production is clear and intelligible.

a. Sign production is clear and accurate
b. Fingerspelling is clear and appropriate for the context
c. Pausing is appropriate

2. Standard: Overall use of grammatical markers is accurate.

a. Cohesive use of markers (e.g., tense/time indicators, plurals, etc.)
b. Use of space appropriate
c. Effective use of classifiers
d. Accurate use of pronouns
e. Accurate use of non-manual sign modifications (e.g., mouth movement, eyebrows, sign movement/intensity, etc.)

3. Standard: Overall use of sentence structures is appropriate.

a. Use of complete sentences
b. Sentence structures are appropriately marked (e.g., eyebrows, eye gaze, mouth movements, used to indicate negation, questions, etc.)

Appendix 2

Association of Visual Language Interpreters of Canada

Canadian Evaluation System

avlic

Association of
Visual Language Interpreters
of Canada

Test of Interpretation

Message Equivalency Criteria

Introduction

Within the domain of Message Equivalency, the Canadian Evaluation System (CES) examines interpreting work that demonstrates the use of bilingual and bicultural strategies (ASL/English/cultural adjustments). Successful interpreting performances in this domain demonstrate the application of a processed or sociolinguistic model of interpretation.

The domain of Message Equivalency (ME) is rated by using criteria outlined in the following document. These features are not discrete entities where the presence or absence of any one feature will determine the success or failure of a particular segment. Rather, raters examine the interpretation within the holistic context of language usage, examining patterns of success or patterns of miscues. The Message Equivalency raters have used these criteria since the CES was implemented, and all raters participate in training review prior to annual rating.

AVLIC would like to acknowledge the time and talents of Debra Russell, Karen Malcolm, Greg Evans, Marty Taylor and Terry Janzen for their integral part in the creation of the Message Equivalency Rating Form. In addition, we would like to acknowledge and thank Douglas College Sign Language Interpretation Program, New Westminster, BC, for allowing us to draw upon their resources and to adapt their Preceptor's Guide toward the development of this rating form.

AVLIC Evaluation Committee 2007

I MESSAGE PROCESSING

1. Standard: Overall message processing results in coherent and accurate interpretation.

a. Understands and conveys speaker/signer {source} goals
b. Essential elements of meaning/main points conveyed
c. Appropriate detail conveyed to support main points
d. Appropriate use of expansions and reductions
e. Overall discourse strategies used result in a coherent target text[12]
f. Successful management of processing levels[13]
g. Interpretation is not marked with numerous false starts
h. Interpretation is generally successful. If not, is there a pattern (deceptive, intrusive or dysfunctional)?[14]

II INTERPRETING SUB-TASKS

1. Standard: Overall, interpreter comprehends the source message.

a. Analysis of source message,[15] syntax and grammatical features
b. Monitoring of own work demonstrated/makes corrections appropriately
c. Effectively mediates culturally-laden elements of the message
d. Conveys cultural (and other) gestures; verbal and non-verbal cues
e. Demonstrates awareness[16] of the register for that given situation

[12] Opening/closing comments, cohesion and discourse markers, topic transition and topic maintenance, etc.

[13] Not typically operating at the lexical or phrasal level, but rather at the sentential and discourse levels.

[14] **Deceptive**: surface appearance of being successful in the target language, however the interpretation actually conveys a message or intent other than the source.
Intrusive: interpretation deviates from the expected norms and may retain much of the source language; consumers may be able to use closure skills and or knowledge of the source language to recover the original intent.
Dysfunctional: impossible to retrieve the source message through the interpretation.

[15] Semantics, contextual knowledge and associated relations (previous knowledge/experience with the topic), cultural norms, etc.

[16] Matches source affect, matches linguistic features that reflect the affect.

2. Standard: Overall target message is accurate.

a. Target language output: overall interpretation grammatically correct
b. Target language output: overall interpretation semantically accurate
c. Target language output: appropriate use of discourse markers

III MISCUE PATTERNS

1. Standard: Overall impact of miscues on interpretation is minimal.

a. If miscues are excessive, is there a pattern (omissions, additions, substitutions, anomalies)?[17]

IV ADDITIONAL OBSERVATIONS

1. Delivery and flow[18] look natural.
2. Interpreter looks confident.
3. Interpreter demonstrates few/no personal distracting mannerisms.

References

Colonomos, B. 1983. Interpretation: A Working Model. Riverdale, MD. The Bicultural Centre. Unpublished workshop materials.

Cokely, D. 1992. Interpretation: A Sociolinguistic Model. Burtonsville, MD: Linstok Press.

Douglas College Program of Sign Language Interpretation. 1996. Preceptor's Guide & Student Guide. Vancouver, BC: Douglas College.

Humphrey, J. and Alcorn, B. 1995. So You Want To Be An Interpreter: An Introduction To Sign Language Interpreting. Amarillo, Tx: H & H Publishers.

Livingston, S., Singer, B. & Abramson, T. 1994. Effectiveness Compared: ASL interpretation vs. transliteration. Sign Language Studies, 82 Spring.

[17] **Omissions**: deletions that cause significant loss of meaning from the source message;
Additions: additions that are not found in the source message that significantly alter the source message;
Substitutions: substitutions that are not found in the source message that significantly alter the source message;
Anomalies: idiosyncratic linguistic and non-linguistic behaviours that are attributed to the interpreter and not the source language.

[18] Demonstrates appropriate use of: breathing, voice quality, sign production quality, pausing strategies, ability to change the pacing according to the source message.

For further information contact:

Association of Visual Language Interpreters of Canada
Canadian Evaluation System
P.O Box 29005 Lendrum
Edmonton, Alberta
T6H 5Z6
Phone: 403-430-9442 V/TTY
Fax: 403-430-9489
Email: ces@avlic.ca
Web: www.avlic.ca

Author index

Subject index

In the series *American Translators Association Scholarly Monograph Series* the following titles have been published thus far or are scheduled for publication:

XIV **ANGELELLI, Claudia V. and Holly E. JACOBSON (eds.):** Testing and Assessment in Translation and Interpreting Studies. A call for dialogue between research and practice. vi, 386 pp. *Expected October 2009*

XIII **DUNNE, Keiran J. (ed.):** Perspectives on Localization. 2006. vi, 356 pp.

XII **BAER, Brian James and Geoffrey S. KOBY (eds.):** Beyond the Ivory Tower. Rethinking translation pedagogy. 2003. xvi, 259 pp.

XI **SPRUNG, Robert C. (ed.):** Translating Into Success. Cutting-edge strategies for going multilingual in a global age. Co-editor: Simone Jaroniec. 2000. xxii, 240 pp.

X **FISCHBACH, Henry † (ed.):** Translation and Medicine. 1998. viii, 180 pp.

IX **LABRUM, Marian B. (ed.):** The Changing Scene in World Languages. Issues and challenges. 1997. 160 pp.

VIII **MORRIS, Marshall (ed.):** Translation and the Law. 1995. viii, 334 pp.

VII **HAMMOND, Deanna L. (ed.):** Professional Issues for Translators and Interpreters. 1994. viii, 216 pp.

VI **WRIGHT, Sue Ellen and Leland D. WRIGHT, JR. (eds.):** Scientific and Technical Translation. 1993. viii, 298 pp.

V **LARSON, Mildred L. (ed.):** Translation: Theory and Practice, Tension and Interdependence. (SUNY) Binghamton, 1991. viii, 270 pp.

IV **BOWEN, David and Margareta BOWEN (eds.):** Interpreting. Yesterday, Today, and Tomorrow. (SUNY) Binghamton, 1990. viii, 184 pp.

III **KRAWUTSCHKE, Peter W. (ed.):** Translator and Interpreter Training and Foreign Language Pedagogy. (SUNY) Binghamton, 1989. vi, 178 pp.

II **VASCONCELLOS, Muriel (ed.):** Technology as Translation Strategy. (SUNY) Binghamton, 1988. viii, 248 pp.

I **GADDIS ROSE, Marilyn (ed.):** Translation Excellence. Assessment, Achievement, Maintenance. (SUNY) Binghamton, 1987. vi, 151 pp.