Experience and Expertise in Conference Interpreting

An Investigation of Swedish Conference Interpreters

Elisabet Tiselius

Dissertation for the degree of philosophiae doctor (PhD)
at the University of Bergen

2013

Dissertation date: November 26, 2013
## Contents

SCIENTIFIC ENVIRONMENT ...........................................................................................................11

ACKNOWLEDGEMENTS ..................................................................................................................12

ABSTRACT .........................................................................................................................................15

THE PRESENT THESIS IS BASED ON THE FOLLOWING PAPERS .............................................17

1. INTRODUCTION ..............................................................................................................................18
   1.1 EXPERTISE AND INTERPRETING RESEARCH .................................................................18
   1.2 AIM ..............................................................................................................................................20

2. BACKGROUND ..................................................................................................................................22
   2.1 RESEARCH ON THE EFFECT OF EXPERIENCE IN INTERPRETING ..................................22
   2.2 INTERPRETING PRACTICE ......................................................................................................27
   2.3 EXPERTISE ...............................................................................................................................29
      2.3.1 Different concepts of expertise .........................................................................................29
      2.3.2 The concept of deliberate practice in the expertise approach .......................................35
      2.3.3 Research on expertise in interpreting ...............................................................................38
   2.4 ASSESSMENT .............................................................................................................................46
   2.5 RESEARCH QUESTIONS AND METHODOLOGICAL DEVELOPMENT ..................................50

3. DATA AND METHODS ......................................................................................................................53
   3.1 METHODS ....................................................................................................................................53
      3.1.1 Investigating expertise .......................................................................................................53
      3.1.2 Investigating the process ...................................................................................................55
      3.1.3 Investigating the product ...................................................................................................57
      3.1.4 Investigating the participants ............................................................................................58
   3.2 PARTICIPANTS .............................................................................................................................58
      3.2.1 The cross-sectional data – data set A ...............................................................................59
3.2.2  The long-term data – data set B .......................................................... 61

3.2.3  The raters .................................................................................................. 62

3.3  INTERPRETING DATA.................................................................................. 63

3.3.1  The EU speech.......................................................................................... 63

3.3.2  The NATO speech..................................................................................... 63

3.4  RETROSPECTIVE DATA AND ASSESSMENT FILES .................................. 64

3.5  METHODOLOGICAL DISCUSSION.......................................................... 64

3.5.1  Mixed-method design .............................................................................. 65

3.5.2  Re-test or not, and other challenges.......................................................... 67

3.5.3  Terminological inconsistencies .................................................................. 68

4.  SUMMARY OF ARTICLES AND GENERAL RESULTS ................................. 69

4.1  ARTICLE 1: “REVISITING CARROLL’S SCALES” (DATA SET A).................. 69

4.1.1  Background .............................................................................................. 70

4.1.2  Data and method ...................................................................................... 71

4.1.3  Major findings .......................................................................................... 71

4.1.4  Discussion .................................................................................................. 72

4.2  ARTICLE 2: “PROCESS AND PRODUCT IN SIMULTANEOUS INTERPRETING: WHAT THEY TELL US ABOUT EXPERIENCE AND EXPERTISE” (DATA SET A) ............... 73

4.2.1  Background .............................................................................................. 73

4.2.2  Data and method ...................................................................................... 74

4.2.3  Major findings .......................................................................................... 74

4.2.4  Discussion .................................................................................................. 76

4.3  ARTICLE 3 – “THE DEVELOPMENT OF EXPERTISE – OR NOT: THREE SIMULTANEOUS INTERPRETERS’ DEVELOPMENT OVER TIME” (DATA SET B) ................................................................. 77

4.3.1  Background .............................................................................................. 77
4.3.2 Material and method ...............................................................78
4.3.3 Major findings .........................................................................79
4.3.4 Discussion ...............................................................................80
4.4 **ARTICLE 4 – “EXPERTISE WITHOUT DELIBERATE PRACTICE? THE CASE OF SIMULTANEOUS INTERPRETERS” (DATA SET B)** .................................................................81
  4.4.1 Background ...........................................................................81
  4.4.2 Data and method ....................................................................82
  4.4.3 Major findings ........................................................................83
  4.4.4 Discussion .............................................................................84
4.5 **SUMMARY OF THE RESULTS FROM ALL FOUR STUDIES** .................................................................86
  4.5.1 Methodological results .............................................................86
  4.5.2 Research results .....................................................................86
5. **DISCUSSION** ............................................................................88
  5.1 **METHODOLOGICAL DISCUSSION** ...........................................88
  5.2 **DISCUSSION OF THE RESULTS** .............................................90
  5.3 **THE EXPERTISE THEORY AND SIMULTANEOUS INTERPRETING** .................................................................91
6. **CONCLUSION** ............................................................................95
**LIST OF REFERENCES** ....................................................................100
**ARTICLE 1** ..................................................................................111
**REVISITING CARROLL’S SCALES** ..................................................111
  1. **INTRODUCTION** .....................................................................111
    1.1 Purpose and research questions .............................................112
  2. **BACKGROUND** .......................................................................113
    2.1 Carroll’s scales .......................................................................113
  2.2 Applying grading scales to interpreting .....................................115
3. DATA AND METHOD ................................................................. 117

3.1 ADAPTATION OF THE SCALES .............................................. 117

3.2 Eliciting Material.................................................................. 120

3.2.1 The speech.................................................................... 120

3.2.2 The interpreters.............................................................. 121

3.2.3 Preparing the transcripts .................................................. 121

3.3 THE GRADING PROCEDURE ................................................... 123

3.3.1 The graders .................................................................. 123

3.3.2 Grader training............................................................... 123

3.3.3 Grading ...................................................................... 123

3.4 MEASURING SIGNIFICANT DIFFERENCE AND INTER-RATER RELIABILITY ................ 124

4. RESULTS ............................................................................ 125

4.1 INTER-RATER RELIABILITY .................................................. 125

4.2 INTELLIGIBILITY................................................................. 125

4.2.1 Intelligibility graded by non-interpreter graders vs. interpreter graders ................. 126

4.3 INFORMATIVENESS ............................................................. 127

4.3.1 Informativeness graded by non-interpreter graders vs. interpreter graders ............ 127

4.4 SPONTANEOUS COMMENTS FROM GRADERS .......................... 128

5. DISCUSSION ..................................................................... 128

5.1 LIMITATIONS .................................................................. 128

5.2 DISCUSSION OF THE RESULTS .......................................... 129

6. CONCLUSION ................................................................... 131

REFERENCES ........................................................................ 132

ARTICLE 2 ........................................................................ 136

PROCESS AND PRODUCT IN SIMULTANEOUS INTERPRETING ........... 136
1. Introduction .............................................................................................................. 137
  1.1 Aim and Scope ........................................................................................................ 138

2. Background ................................................................................................................ 138
  2.1 Expertise Research in Interpreting Studies ................................................................ 138
  2.2 Studying the Process ............................................................................................... 140
  2.3 Evaluating the product ............................................................................................ 140
  2.4 The Studies ............................................................................................................. 141

3. First Study: Interpreting Process .............................................................................. 142
  3.1 Material and Method ............................................................................................... 142
    3.1.1 Input Material .................................................................................................. 142
    3.1.2 The Interpreting Subjects .............................................................................. 142
    3.1.3 The Retrospection Procedure ....................................................................... 144
    3.1.4 Transcription and Analysis ........................................................................ 144
  3.2 Results ................................................................................................................... 145
    3.2.1 Processing Problems ..................................................................................... 145
    3.2.2 Instances of Monitoring ............................................................................... 148
    3.2.3 Strategies ....................................................................................................... 150

4. Second Study: Assessment of the Interpreting Product .......................................... 152
  4.1 Material and Method ............................................................................................... 152
    4.1.1 The Rating Files ............................................................................................. 153
    4.1.2 The Raters ...................................................................................................... 153
    4.1.3 The Rating Session ....................................................................................... 153
    4.1.4 The Scales and the Analysis ........................................................................ 153
  4.2 Results ................................................................................................................... 154
    4.2.1 Intelligibility and Informativeness ................................................................ 154
2. Background ........................................................................................................... 194

3. Methodology ........................................................................................................ 195
   3.1 Participants ....................................................................................................... 195
   3.2 Procedure for conducting the in-depth interviews ........................................... 196
      3.2.1 Identification of topics and core phenomena ......................................... 196
      3.2.2 The interview setting .............................................................................. 197
   3.3 Coding and analyses of the interview data ..................................................... 198

4. Qualitative Sides of Expertise ............................................................................. 198
   4.1 Language learning and language knowledge ................................................... 198
   4.2 General knowledge ......................................................................................... 199
   4.3 Communicative skills ...................................................................................... 199
   4.4 Focus ............................................................................................................... 200
   4.5 Coping with stress ......................................................................................... 200
   4.6 The interpreting skill ..................................................................................... 201

5. Deliberate practice, clear goals and openness to feedback ................................... 201

6. Discussion and Conclusions ............................................................................... 203

References ............................................................................................................... 205

List of tables in summary

Table 1. Age and experience of the cross-sectional interpreters .............................. 59

List of figures in summary

Figure 1. Novice-Expert continuum .................................................................... 43
Figure 2 Summary of participants, data, analyses and articles ............................. 69

List of tables in article 1

Table 1 Scale of intelligibility ............................................................................. 118
Table 2. Scale of informativeness ..................................................................... 119
Table 3. Scale of intelligibility on grading sheet................................. 120
Table 4. Scale of informativeness on grading sheet.................................. 120
Table 5. Age and experience of the interpreters.................................... 121
Table 6. Significance in gradings of intelligibility by non-interpreters (n=6)........ 125
Table 7. Significance in gradings of intelligibility by interpreters (n=6)............. 126
Table 8. Avg. scores of intelligibility graded by non-interpreters (n=6) and interpreters (n=6). 126
Table 9. Significance for grading of informativeness by non-interpreters (n=6).......... 127
Table 10. Significance for grading of informativeness by interpreters (n=6)........... 127
Table 11. Significance of grading of informativeness graded by non-int (n=6) and int (n=6) ... 127
Table 12. Processing problems. Classification according to Ivanova (1999). ................. 188
Table 13. Monitoring observations (Ivanova 1999). .................................. 189
Table 14. Strategies (Ivanova 1999). .................................................... 190

List of figures in article 1

Figure 1. Avg scores for intelligibility graded by int (n=6) and non-int (n=6)............. 126
Figure 2. Avg scores for informativeness graded by int (n=6) and non-int (n=6)........... 128

List of tables in article 2

Table 1. Age and experience of the interpreters....................................... 143
Table 2. Processing Problems. Definitions of the headings are given in appendix 1 .......... 145
Table 3. Processing problems long experience Ivanova vs. the present study................. 147
Table 4. Processing problems Ivanova’s novices vs. short and no experience interpreters .... 147
Table 5. Instances of monitoring .......................................................... 148
Table 6. Instances of monitoring Ivanova’s expert vs. the long experience interpreters ....... 150
Table 7. Instances of monitoring Ivanova’s novices vs. the short and no experience int. ....... 150
Table 8. Instances of reported strategies ................................................. 151
Table 9. Mean evaluation scores of intelligibility and informativeness....................... 154

List of figures in article 2

Figure 1. CA biplot of the association between interpreter experience and processing probl. .. 146
Figure 2. CA biplot of the association between interpreter experience and instances of monit. 149
Figure 3. CA biplot of the association between interpreter experience and strategies. ........ 152
Figure 4. Intelligibility: boxplot of between-group differences. ................................ 155
Figure 5. Informativeness: boxplot of between-group differences. ............................ 155

List of tables in article 3

Table 1. Intelligibility and informativeness: NATO speech, comparison.......................... 176
Table 2. Intelligibility and informativeness: EU speech, student rating only. ................... 177
Table 3. Intelligibility and informativeness: cross-sectional and longitudinal data set. .......... 177
Table 4. Length of interpreting, length of retrospection: NATO speech.......................... 177
Table 5. Length of interpreting, length of retrospection: EU speech .................................................. 178
Table 6. Processing problems and categories: NATO speech, longitudinal participants .......... 178
Table 7. Processing problems and categories: EU speech ................................................................. 179
Table 8. Reported monitoring categories: NATO speech ................................................................. 179
Table 9. Reported monitoring categories: EU speech .................................................................... 180
Table 10. Strategies: NATO speech, longitudinal experienced participants ............................... 180
Table 11. Strategies: EU speech, longitudinal and cross-sectional interpreters ......................... 180
Table 12. Macrostrategies: NATO speech, longitudinal experience participants ..................... 181
Table 13. Macrostrategies: EU speech, longitudinal and cross-sectional interpreters ............... 181
Table 14. Processing problems. Classification according to Ivanova (1999) ............................... 188
Table 15. Monitoring observations (Ivanova 1999) .................................................................... 189
Table 16. Strategies (Ivanova 1999) ............................................................................................. 190
Table 17. Macrostrategies (Ivanova 1999) .................................................................................. 191

List of figures in article 4

Figure 1. Thematic sketch of topics covered in the in-depth interviews .................................... 197
Scientific environment

This is a dissertation written at the Department of Foreign Languages at the University of Bergen, the Centre for Research on Bilingualism at Stockholm University and the Institute for Interpretation and Translation Studies at Stockholm University. The dissertation work has been supervised by associate professor Åse Johnsen, professor Birgitta Englund Dimitrova and professor Kenneth Hyltenstam.
Acknowledgements
Writing a dissertation means depending on people for support, encouragement and also critique. There are many people who have done that for me throughout this journey. The first persons to be mentioned, though, should be the ones who were the sine qua non of my work. I’m so grateful that Sarah Williams recorded so many interpreters with such fervency, as she gave me the starting point of my data. Thank you, Sarah, I hope I have given your material justice.

The other very important persons for this project are the interpreters – without their interest and willingness to participate in different research projects, this dissertation would not have gone far. The Swedish interpreting community is a small one, and I have pestered many conference interpreting colleagues with my research. I am forever grateful to you, dear colleagues, that you bore with me all the way from my Master’s thesis over ten years ago.

A PhD project does not take off without supervisors to encourage, question and guide the candidate, and put her right from time to time. I have been blessed with the best. You have all been crucial to me, Kenneth for believing in me from the very beginning, Birgitta for inviting me to do research with you, and Åse for helping me keeping the focus. Thank you!

Many other people have also been important for this dissertation: Maria Wingqvist, who sparked the fire when she asked the budding Master’s student, “Would you consider writing a PhD?”; Karin S. Lindelöf and Magnus Öhlander, who helped out when I was lost in ethnographic methodology; Lidun Hareide, who has been a wonderful colleague and sounding board on the Norwegian side; Gard Buen Jenset, who agreed to co-author a paper and put me out of my statistical misery; Stig Oppedal, without whose diligent proofreading and editing my prose would only seem like meaningless rambling; and Sandra Halverson, who brought me to Norway. Thank you so much for helping me in this work.
When it comes to my development in interpreting studies, I’m much obliged to the European Society for Translation Studies for generously giving me their summer school grant and making it possible for me to take the CETRA research summer school at KU Leuven. The CETRA course was a wonderful starting point and also gave me the fantastic opportunity to get help and support from Andrew Chesterman, Daniel Gile, Franz Pöchhacker and Miriam Shlesinger. The continued dedication and support that these stars of translation and interpreting studies have shown to a PhD student is truly wonderful. I will never be able to thank you enough, Andrew, Daniel, Franz and Miriam; Miriam, I hope that you continue to keep an eye on us from where you are now.

CETRA was also where I met my wonderful research buddies Šárka Timarová and Veerle Duflou. Thank you for the great discussions and for keeping a stern “quant” and “qual” eye on the work I’ve done. Through Šárka, I also finally got to meet Adelina Hild (Ivanova), who came into my research life very early when she kindly sent me the copy of her dissertation that I was so curious of. Adi! What can I say – thank you!

Another great place to hang out has been the Translation Process Research Workshop that Susanne Göpferich initiated and Ricardo Muñoz Martín continued. The workshop has an open-minded atmosphere, and presenting there has helped me develop the argument of my dissertation.

Emilia Iglesias Fernández deserves a particular note in this context. When I was looking for an opportunity to learn more about the work of Angela Collados Aís and the ECIS team, which Emilia belongs to, she organized an Erasmus teacher exchange in Granada and put me up in her home. It was a wonderful week filled with academic inspiration both for the dissertation and beyond. ¡Gracias, Emilia!

And in case you ever come up with the idea to edit another book during your PhD period, Cecilia Alvstad and Adelina Hild are the best friends to do that with.
I have belonged to a greater or lesser extent to three departments: the Department of Foreign Languages at Bergen University, the Centre for Research on Bilingualism at Stockholm University, and the Institute for Interpreting and Translation Studies at Stockholm University. I have been the odd one out (and possibly also the oddball) at all three departments, and thank all of them for including me and helping me in my dissertation work.

In the real world, outside the academic bubble, I’m so fortunate to have Johanna Järvinen, Johanna Broman-Åkesson, Pauline Dandois and Kristina Camitz as friends – thanks for putting up with me as I got increasingly self-centred, confused and busy. A particular thank you to Johanna Broman-Åkesson who also took the time to read all the interviews with me. Merci Pauline pour toujours me sauver. Tack för att ni har sett till att både jag, Carl, barnen, hunden och katten har överlevt, framförallt den här sista perioden.

It would be impossible to write a dissertation without support from my family. Carl, du är ljuset i mitt liv. Följ mig bortåt vägen så skiljs vi längre fram, vi fortsätter gå min väg, och din väg, huvudsaken är att inte var tar sin väg. Eva, Daniel och Anna, ni är bäst! Fortsätt att vara precis som ni är.

Brussels,
June 2013
**Abstract**

This dissertation investigates the process and product of interpreters with different levels of experience and explores the expertise approach (cf. Ericsson, Charness and Hoffman 2007) as applied to interpreters. The expertise approach claims that highly skilled performers, regardless of their chosen field, use the same type of strategies in order to reach the top levels of their profession. An important feature of the expertise approach is deliberate practice, a specific type of practice that highly skilled performers engage in so as to improve their performance.

The dissertation is based on four different studies featuring two different sets of participants. Two data sets – a *cross-sectional material* with nine participants on three different levels of interpreting experience (none, short and long), and a *long-term material* with three interpreters recorded at two different points in time – were analysed in terms of both processing and product data. The interpreting process was studied by retrospectively analysing and categorizing processing problems, monitoring and strategies, while the interpreting product was analysed by using holistic rating scales for intelligibility and level of information transfer of the interpreting product. In-depth interviews were also conducted with the long-term participants in order to investigate their perception of deliberate practice and their own view of their skill development. An important and integral part of the dissertation, apart from the results, was the development of the holistic rating scales (adapted from Carroll 1966), and the development of an in-depth interview study.

The conclusions of the dissertation are that there are measurable differences of interpreting skill between performers with little or no interpreting experience and performers with long interpreting experience, but this finding could not be supported by the long-term (intra-individual) study. Differences between the groups in the cross-sectional material could also be observed from the process data. Experienced interpreters
encountered fewer processing problems than less experienced interpreters and had more strategies at hand to solve problems. There were also clear differences in terms of instances of monitoring (i.e. controlling the interpreting process and output) between experienced interpreters and other subjects. Monitoring seemed to be a dividing line between experienced and inexperienced interpreters, and experienced interpreters had more processing capacity available to monitor themselves. This was also to a certain extent supported in the in-depth interviews, where the participants reported how they constantly evaluate themselves in terms of improving performance. A key assumption established in the beginning of the project – that experienced interpreters would claim, in the in-depth interviews, that they practise a great deal – was not supported, to our surprise. The interpreters recounted many practice-like activities but stated that they did not actually practise.

The dissertation concludes by calling for more studies on deliberate practice in interpreting, suggesting that the term “interpreter expert” should only be used with caution in scientific studies and that the particular features of expertise and deliberate practice in interpreting should be discussed.
The present thesis is based on the following papers


III. Tiselius, E. In preparation. “The development of expertise – or not: Three simultaneous interpreters’ development over time.”

1. Introduction

The conference interpreter community, just as any group, has its stars. There were the Kaminker brothers, George and André, who could translate ninety-minute consecutive speeches without notes (Satow 1979: 511). There was Wolfe Frank, who was considered the best interpreter at the Nuremberg trials (Gaiba 1998). Or Paul Mantoux, the historian turned interpreter during the war, and the only interpreter present at the peace negotiations in Versailles after the First World War (Mantoux 1955). More modern examples include Amanda Galsworthy, who has been the English interpreter for four French presidents (Lefort 2012), or Stalin’s interpreter Valentin Berezhkov (Berezhkov 1994). There are also local stars at different institutions or in different booths. Budding interpreters are told stories about older colleagues who are linguistic wizards or multi-sentence jugglers, always getting the message across. Every interpreter has a story of one particular day when the interpreting diploma was still fresh and s/he was assigned to work with one of these stellar colleagues. Stars are created in many different ways, but in order to be an interpreting star an absolute condition is to perform with excellence. In simultaneous interpreting there are few possibilities of cheating: since the cognitive load is extreme and the time is short, smooth-talking is not enough for an expert in interpreting. Interpreting excellence is not achieved overnight – on the contrary, many laborious hours lie behind a seemingly effortless performance, just as for any other profession. Researchers have labelled the type of excellence achieved through many hours of focused training “expertise”.

1.1 Expertise and interpreting research

Expertise theory was introduced to the interpreting research community by Barbara Moser-Mercer, who invited Karl-Anders Ericsson, a leading
proponent of the expertise approach within cognitive psychology, to the Ascona workshops in 1997 (Moser-Mercer 1997). Interpreting research had however focused on the skilled performers long before the expertise theory was introduced. Already early on the skills of experienced professional interpreters were investigated and compared to the performance of less experienced subjects, such as in Gerver’s (1971) research on source-language presentation rate and pauses and Goldman-Eisler’s (1972) early studies on segmentation and lag.

The expertise approach has turned out to be well-suited for interpreting research, and the study of expertise within interpreting has become a viable and well-established research area in interpreting studies. Moser-Mercer has looked at expertise from the learners’ perspective in different studies (e.g. Moser-Mercer 2000; Moser-Mercer, Fraunfelder, Casado & Künzli 2000). Several dissertations have focused on expertise in interpreting, such as Ivanova (1999), who looked at problem-solving strategies; Liu (2001), who investigated working memory; and Vik-Tuovinen (2006), who looked at expertise in a wider perspective by including preparation.

A challenge for expertise research in interpreting is the data collection. Studies tend to be cross-sectional, comparing students or novice interpreters with more experienced interpreters, as in the studies mentioned above. Studies with a more long-term aim often compare how interpreting students develop from the beginning to the end of their programme, as in Moser-Mercer’s studies. Longitudinal, or long-term, studies (for a terminological discussion on longitudinal versus long-term, see below in section 2.3.3) are likely to uncover other aspects of expertise than cross-sectional studies can reveal. It is therefore interesting to have both cross-sectional and longitudinal studies of expertise. Very few studies, if any, tend to be truly longitudinal, that is, stretch over many years and follow informants’ individual development over time. An obvious reason for this is the long time span needed for such a study.
combined with the level of falling off that such a sample would most likely suffer, a feature probably common to all longitudinal studies of expertise.

In this dissertation and in its literary review, different groups of interpreters will be addressed. The groups differ through their interpreting experience. Interpreting experience can be gained from an interpreting program, through professional practice or both. Terms used to describe these different groups are *highly experienced interpreters* for interpreters with long or very long professional experience, *little or short experience interpreters* for interpreters with short professional experience or interpreting students, and finally *no experience interpreters or subjects* for subjects participating in different studies and engaging in interpreting activities but with no prior experience of interpreting.

### 1.2 Aim

The starting point for this dissertation was the investigation of expertise in interpreting. At the beginning of the project I discovered a set of interpreting data that was recorded in the mid-1990s, and it seemed feasible to design a long-term study by making new recordings. With those early recordings in mind, a cross-sectional material was collected in order to complement and mirror the long-term material, and together the two data sets would be used to study the long-term development of expertise. By studying the informants’ interpreting process as well as their interpreting product, combining the results from the cross-sectional material and the long-term material, and examining their expertise through in-depth interviews, I aimed to answer the following questions:

- Is there a measurable difference in the interpreting skill from the student level to the highly experienced level?
- Is there a measurable difference in the interpreting skill both when it is measured cross-sectionally (i.e. inter-individually) and long-term (i.e. intra-individually)?
• If there is a measurable difference in the interpreting skill, what does this difference consist of?

• How do experienced interpreters perceive different factors in their long-term competence development?

In order to identify and study the development of the interpreting skill, it had to first be measured, and the measurement of the interpreting skill had to rely on some type of evaluation. Furthermore, the measurement of the product had to be complemented with an investigation of the process.

As the project developed, important methodological issues cropped up. A major part of the dissertation has also consisted of adapting, testing and developing different methods for investigating expertise. Various techniques for retrospection and the categorization of processing problems, monitoring and strategy use were tested and adapted in order to map the process. Scales for assessing product were also tested and developed. Finally, an interview guide for conducting in-depth interviews was developed.

Thus, this project compared the process and product of highly experienced interpreters, novice interpreters and non-interpreting subjects, both cross-sectionally and in the long term.
2. Background

The following section will give an overview of previous research on experienced interpreters from both a cognitive and a quality perspective, before discussing interpreting as a professional practice. The section concludes with an overview of the expertise theory in cognitive psychology, the concept of deliberate practice within the expertise theory, and the definition of an expert in interpreting research.

2.1 Research on the effect of experience in interpreting

Research into conference interpreting studied experienced interpreters long before the expertise theory entered the field. Researchers were interested in the cognitive effort in interpreting and the different skills needed to interpret. Major themes that have been studied within the cognitive framework include processing capacity and different cognitive efforts. Early researchers were, among other things, interested in whether an increased cognitive load affected interpretation and how interpreters handled the allocation of processing capacity.

Precursors in interpreting research found that response time (or lag in interpreting terminology) increases as the task complexity increases (Oléron & Nanpon 1965 in Pöchhacker & Shlesinger 2002: 42). Moreover, subjects with less interpreting experience deliver more fragmented versions than their more experienced peers, and interpreters make use of the speakers’ pauses for their interpretations (Barik 1973, 1975). Qualitatively significant differences in the output of highly experienced interpreters as compared with subjects with little or no interpreting experience, have been more difficult to establish (Barik 1975). However, a major difference between highly experienced and subjects without or with limited interpreting experience seems to be that experienced interpreters
segment the input more effectively (Barik 1975: 296). Speech rate, pauses and intonation affect interpreters’ output (Gerver 1971), and interpreters produce longer utterances with fewer pauses than the original speaker (ibid.). Interpreters’ segmentation of input differs from the speakers’, and source-language syntax affects both segmentation and ear-voice span, that is, the time from perception of an utterance in the source language until the production of an equivalent in the target language (Goldman-Eisler 1972).

These early findings could not establish conclusive significant differences between subjects with long interpreting experience and subjects with no or limited interpreting experience. But Goldman-Eisler (1972) for one found that cognitive load increased while interpreting compared with other speech production tasks and that highly experienced interpreters seemed to handle that increased cognitive load better than subjects without interpreting experience. She assumed that decoding the input would be the process that required the most attention and that monitoring and encoding are more automatized (Goldman-Eisler 1972: 139). Barik, however, assumed that this difference was due to language direction (Barik 1975: 296). Oléron and Nanpon suggested future studies on how much information can be grasped at one time, or of the interpreters’ ability to concentrate on several different things and perform simultaneously, in order to draw conclusions on the process (Oléron & Nanpon in Pöchhacker & Shlesinger 2002: 50).

Many studies in early interpreting research compare subjects with limited or no experience to interpreters with professional interpreting experience (cf. Gerver 1971; Barik 1973, 1975; Anderson 1979). Hoffman points out that it seemed natural for researchers in interpreting to examine the performance of professionals and contrast it with the one of trainees (Hoffman 1997: 190). From a cognitive perspective, the clearly perceived differences between subjects with and without interpreting experience could be used to gain information on the cognitive impact or change that develops with the subjects’ interpreting skills. As can be seen above,
however, it is not immediately evident how that difference could be measured in order to establish distinctive features between subjects without interpreting experience and interpreters with professional interpreting experience. In fact, some studies, such as Dillinger (1994), could not establish any distinctive difference between interpreters with professional interpreting experience and subjects without interpreting experience.

Sample size has been a methodological issue since the early days of interpreting research. Even today researchers struggle with collecting data large enough for purely experimental designs. The interpreting community, especially the simultaneous interpreting community, is small, and the highly experienced interpreters even fewer. Already Gerver commented that very few interpreters were available for experiments, and furthermore that since

not all of these are willing to take part in experiments, the design and execution of experiments on simultaneous interpretation becomes somewhat of a problem. In effect the choice had twice to be made between an incomplete experimental design or no experiment at all, and in the experiments described [here] the former decision was taken. (Gerver 1971: 26)

Traditional types of analysing tools in early research comprise error analysis (e.g. Gerver 1971; Gile 1985a); ear-voice span measurements (Gerver 1971; Goldman-Eisler 1972; Barik 1973); and assessment of interpreting (Seleskovitch 1975 as cited in Pöchhacker & Shlesinger 2002; Anderson 1979). Assessment is an important part both of interpreting practice and interpreting research and will be dealt with more in detail below in section 2.4.

Early studies of cognitive aspects of interpreting research also propose different models of the interpreting process: Gerver (1976) and Moser-Mercer (Moser 1978) propose general models of simultaneous
interpreting; Gile (1985b) focuses on the different efforts involved in simultaneous interpreting; and Darò and Fabbro (1994) map memory use. Whether focusing on the whole process, the invested effort or the different memory functions in use, all models mirror cognitive constraint of some sort – there is a limit to how much load different processes can handle. The most recent contribution to models of interpreting was proposed by Seeber (2011), whose cognitive load model aims to be a more flexible model and show how cognitive load increases and decreases depending on task.

More recent research into the cognitive aspects of interpreting has also looked at neurolinguistic aspects of interpreting. The plasticity of the brain makes it adapt to new conditions that simultaneous interpreting imposes on students of interpreting. Early neurological research in interpreting was done by Gran and Fabbro (1987). They have been followed by Rinne et al. (2000), who used PET (positron emission tomography) to study simultaneous interpreting, and more recently by Ahrens et al. (2010), who found significant differences in active brain areas between interpreting students’ interpreting and their normal speech production. Hervais-Adelman et al. (2011) also found indications of change in the bilingual brain of interpreters. It should be stressed that all four of the aforementioned studies focus on students of interpreting, and their findings also correspond to research in expertise. The brain’s single-domain general control network helps us learn new tasks and also plays a key role in controlling working memory in cognitive processing. However, the control network also limits the resources for working memory-dependent tasks and other novel tasks. As processing becomes automatized, the influence of the general control network decreases or disappears. Hill and Schneider (2007), in an overview of different studies on skills acquisition, show how the automatization of acquired skills changes brain plasticity (Hill & Schneider 2007: 675). The acquisition and mastery of new skills change the area activated in the brain, since the brain’s plasticity entails that it can
change area and amount of activity as skills are acquired and refined (ibid.).

The function of working memory in interpreting has been studied both from a novice/experienced dichotomy, and also in itself as one of the features involved in simultaneous interpreting processes. Liu (2001), who investigated working memory from an experience/novice perspective through a listening span test, found no significant differences, nor were there any significant differences in working memory span (Liu, Schallert and Carroll 2004); however, experienced interpreters were more accurate in their performance (Liu 2008). Other researchers (e.g. Bajo, Padilla and Padilla 2000) have found that memory span increases with experience. The reason for the contradictory findings on working memory is perhaps partly explained by Timarová (2012), who found that that interpreters’ working memory is related to their performance in simultaneous interpreting and that simultaneous interpreting is predominantly related to the central executive functions and not to memory functions. She concluded moreover that there was a link between interpreting experience and some working memory functions.

Research on the effect of experience in interpreting has an underlying assumption of basic translation ability (Englund Dimitrova 2005: 10). An individual who understands two languages also has a basic ability to transfer a message from one of the languages to the other. Englund Dimitrova points out that most research in the field of bilingualism takes for granted a basic translation ability. Englund Dimitrova posits that “basic translation ability is a necessary condition, but no guarantee, for further development of a (professional) competence as a translator, and possibly expertise in translation” (2005: 12). It is fair to assume that the underlying assumptions of studies using subjects with little or no interpreting experience are similar to Englund Dimitrova’s postulate. In the present dissertation it can furthermore be noted that for one group of subjects who
did not have any interpreting experience at all, the subjects still produced an interpreting product – it was undoubtedly a difficult task for these subjects, but they did not suffer a complete breakdown. Thus, at least these subjects seemed to possess some type of basic interpreting ability.

As seen above, although some research has shown that interpreters with professional interpreting experience have few omissions, deliver a complete message, segment effectively, handle cognitive load well and have great working memory capacity compared with subjects with little or no interpreting experience, other research has conversely concluded that there is scant difference in accurate delivery, handling of cognitive load or working memory capacity between interpreters with professional interpreting experience and subjects with little or no experience. From earlier research it can be concluded that investigating the effects of experience in interpreting is delicate and difficult. Intuitively, experience ought to improve interpreting performance, but research results are far from conclusive.

2.2 Interpreting practice

Today, interpreting training is common both for public service interpreting and conference interpreting. This section will only focus on conference interpreting training and practice, as the studies in this PhD thesis are restricted to simultaneous conference interpreters. Conference interpreters today are usually trained, although training facilities in this domain are a recent innovation as well. The first interpreting programme was founded in Geneva in 1941. Since then, a pedagogical tradition has evolved through groundbreaking work by Herbert (1952) and Rozan (1979 [1956]), via Seleskovitch and Lederer (1995) to modern classics such as Jones (1998) or Nolan (2005).

Western interpreter training follows more or less the same path. Interpreting pedagogy is firmly rooted in *A Systematic Approach to*
Teaching Interpretation (Seleskovitch and Lederer 1995). Following this and subsequent manuals, interpreting is introduced first through short memory exercises where students are encouraged to let go of the words and look for the meaning of the utterance, a meaning that they should then render in their target language. Memory exercises gradually become longer, and note-taking is introduced. When students master the basics of note-taking, they start to interpret in consecutive mode. Both memory exercises and consecutive interpreting are believed to be a basic preparation for simultaneous interpreting (Gile 2005b). After an extended period of consecutive interpreting (from a semester up to a year), students are typically introduced to the simultaneous mode.

Interpreting students are also taught to practise on their own, outside of teacher-led training. This has been an important characteristic of interpreter training since the early days, although Seleskovitch and Lederer do not provide guidelines for student-led practice but refer to how students should practise and how the teacher should guide that practice (e.g. 1995: 158). Students are expected to practise sub-skills such as language knowledge and general knowledge, often by reading newspapers, watching TV or listening to the radio, but they are also taught to practise interpreting and to record themselves in order to evaluate their performance (Gile 2005b: 135–136).

Interpreter training is guild-like in the sense that active interpreters teach their future colleagues. Furthermore, interpreter training has been developed from a pragmatic rather than from a theoretical perspective. Interpreter training has since the early days been based on active interpreters’ perceptions of what needs to be taught in order to succeed as an interpreter. Their views are confirmed as their students graduate and practise successfully. Sawyer (2004) and Iglesias Fernández (2003) have made very comprehensive overviews on interpreter training. Sawyer found that the extensive research and debate on assessment in interpreting has
been poorly reflected in actual interpreter training (2004: 211). He also stresses the need for test validation (2004: 231). It could probably be claimed that interpreter training is more practice-driven than research-driven. This does not necessarily mean that interpreter training is inadequate, but although much has been studied and written on the matter, interpreter training still frequently takes its starting point in tradition, and there are no major empirical studies on interpreting methodology or didactics (cf. Pöchhacker 2004: 183).

As mentioned above, students are taught to practise and assess their interpreting skills at the interpreting programme. Whether interpreters continue to do so in their professional careers has not been mapped in any larger studies, although a study by Leis (2003) suggests that interpreters do assess themselves. A strong norm in conference interpreting is preparation (e.g. Bühler 1986), that is, students are taught to prepare by improving their background knowledge and enhancing their terminology within a certain topic. Professional interpreters are expected to prepare, and professional experience depends not only on hours in the booth but also on practice and preparation.

2.3 Expertise

The following section introduces different approaches to expertise and the concept of deliberate practice. Expertise has been studied from many different perspectives, ranging from theories where talent is the only condition for expertise to those where focused training is deemed more critical for reaching an expert level.

2.3.1 Different concepts of expertise

Already Plato was interested in the expert mentality, contending that humans could be divided into three different types according to their innate aptitude: soldiers, workers and leaders (Ericsson 2009). Over two
millennia later, the British anthropologist and psychologist Francis Galton (Ericsson 2007a: 684) observed that distinguished contributors to society all came from more or less the same background, leading him to assume that talent and excellence were due to an inherited difference in mental capacities. The idea that expertise is developed through training and practice, which is a cornerstone in many current definitions of experts (cf. Ericsson 2007b: 10–12), was a reaction to the prevalent notion that talent was an absolute condition for success in different fields. Therefore, rather than possessing and relying on a unique talent, the aspiring expert must be prepared to spend many hours of focused practice, often from a very early age. A famous contribution to prove this claim is the Polgár couple, who trained their daughters very early on to become elite chess players and thereby demonstrated that pure talent and the supposed male advantage in chess are pure fantasy (Ericsson, Prietula & Cokely 2007). In this view, expertise as opposed to talent means that the expert has spent many years of specific focused training, so-called deliberate practice (Ericsson 2004: see below, section 2.3.2). Certain other experts, for instance athletes and musicians, also start at a very young age.

The theory of expertise, which has become influential in both Interpreting and Translation Studies, has been developed by researchers in cognitive psychology such as Ericsson, Charness, Feltovich and Hoffman (2007). From a cognitive perspective, an expert is an individual who has acquired great knowledge in a given field and who can make use of this knowledge to outperform other performers. To continue in Ericsson’s words: “expertise then refers to the characteristics, skills and knowledge that distinguish experts from novices and less experienced people” (2007b: 3). 

**Expert performances** are reproducible superior performances of tasks that capture the essence of the respective domains (Ericsson et al. 2007: 3–4). Furthermore, two types of expertise can be singled out, namely routine expertise and adaptive expertise, where routine experts excel in well-known routinized tasks, whereas adaptive experts are able to handle new
tasks and can apply previous knowledge to new situations within their area of expertise (Sonnentag, Niessen & Volmer 2007: 377–378).

Ericsson and Smith (1991) pointed out that studies of expertise often only looked at differences between experts and novices or less experienced individuals, instead of investigating expert performance characteristics within a particular domain. They believed that an expert may be socially recognized as an expert without necessarily showing superior performance in absolute terms compared with other performers (Ericsson & Smith 1991). In an earlier article (Tiselius 2010), I argued that being an expert is impossible in a social vacuum and that, in line with Ericsson and Smith’s argument, expertise requires both social appreciation and superior performance.

Another concept of expertise, from a more heuristic perspective, is interactional expertise, proposed by Collins and Evans (2007: 2). Interactional expertise is a type of expertise that is co-created between parties in a particular field. Collins and Evans point out that expertise can be defined on several different levels, with the most advanced level being contributory expertise, which they define as the stage when an individual has gained specialist knowledge and can help disseminate and increase such knowledge (Collins & Evans 2007: 2). According to the theory of interactional expertise, an expert is not always the best person to decide how to put his or her expert knowledge into practice, and it is through mutual discussions between experts and lay people that the best solution can be found.

The expertise approach proposed by Ericsson and Smith (1991) focuses on the individual performer. According to Ericsson and Smith, expertise in a field is achieved through a combination of various characteristics (1991: 7, 20–21, 27–28). These characteristics, which are also the ones applied in the research project reported here, consist of at least the following:
(1) *Experts have regular outstanding performances in their field of expertise.* The expert has to show regular proof of expertise; a single top performance is not equal to expertise.

(2) *Experts have access to expert knowledge when needed.* Experts do not necessarily outperform other participants on routine tasks, but excel over novices in difficult situations encountered within their area of expertise.

(3) *Experts have long experience in their field of expertise.* Experts have spent at least ten years or 10 000 hours on task and in practice. It should be stressed that this is the weakest predictor of expertise. In many contexts, in particular in popularized accounts of the theory (e.g. Gladwell 2008), ten years of experience has been put forward as a sole or at least dominant factor to determine or achieve expertise. Clearly, non-expert performers may have spent an equal amount of time on task, without achieving expert levels of performance. However, expertise is hardly possible without extensive experience.

(4) *Experts engage in deliberate practice.* Deliberate practice is a highly focused and regular practice, completed at a time set aside only for practice and solely aimed at improving the given skill. It is also characterized by specific exercises and is often coached.

(5) *Experts have clear goals.* The notion of having clear goals is partly connected with the concept of deliberate practice, as the practice is goal-defined. Furthermore, final goals are usually divided into reachable part-time goals on both the micro and macro levels.

(6) *Experts are open to feedback.* Experts have a positive view of receiving feedback and are good at integrating it, both from superiors and peers.
A challenge for anyone wishing to investigate expertise in domains where there is no external ranking is to define outstanding performance. Ericsson and Smith (1991) give a three-step method for doing this: (1) scientifically analyse the domain, its particular expert skills and performance within the framework of general cognitive theory; (2) identify the task’s process and structure and the performers’ behaviour; and (3) show how superior performance in that field is built up through the given cognitive processes and how they were acquired. The three-step model is ambitious and far-reaching, and it would entail a very large project to cover these three steps in order to investigate, for instance, simultaneous interpreting. It can be claimed, however, that the growing number of studies on expertise in interpreting helps to build this three-step model. Ericsson (1996) argued that perhaps not all domains are possible for understanding and measuring expertise, as it requires expert performances of objective superiority that can be reproduced. It is indeed a challenge for interpreting research to show that these requirements are achievable.

Ericsson’s expert approach has had a great impact on expertise research both in psychology and translation and interpreting studies. There are other proposals and models of how to interpret the notion of expertise from a cognitive perspective. Shanteau’s (1992) theory of expert competence aims to reconcile two views existing at that time, namely the cognitive perspective that claimed that experts were cognitively different in every aspect compared to other performers, and research into judgment and decision where experts had made flawed decisions despite their expertise. Shanteau suggests they are both right, but the analysis is incomplete. Instead, in his theory he claims that expertise is built up of five components, namely (1) a sufficient knowledge of the domain, (2) the psychological traits associated with experts, (3) the cognitive skills necessary to make tough decisions, (4) the ability to use appropriate decision strategies, and (5) a task with suitable characteristics. Shanteau may prove useful for this PhD project and for expertise in interpreting, as
measuring expertise in interpreting seems to be a notoriously challenging activity.

If Ericsson’s expert characteristics are contrasted with Shanteau’s expert components, it is clear that they are neither completely opposite nor completely parallel. Shanteau’s second component, “psychological traits associated with experts”, could encompass Ericsson’s “regular outstanding performance” and “access to expert knowledge when needed”. On the other hand, the “access to expert knowledge” would also encompass both Shanteau’s “sufficient knowledge of the domain”, “cognitive skills necessary to make tough decisions” and “the ability to use appropriate decision strategies”. Ericsson also adds “deliberate practice”, “clear goals” and “openness to feedback”. These three cannot easily be put into any of Shanteau’s components. They contribute to for example “psychological traits…” or “cognitive skills…”, but they are not an uncontroversial part of them. An important difference between Shanteau’s components and Ericsson’s characteristics is that Ericsson’s characteristics have a developmental part. They encompass the learning perspective by stressing the importance of a subject’s deliberate practice and openness to feedback. The sociological part of Ericsson’s expertise approach lies in the notion of the subject who engages in deliberate practice and receives feedback and coaching from peers or coaches.

In more recent works, Weiss and Shanteau (2003) have developed an index to empirically assess professional expertise. In their presentation of the index they mention precisely the problem of measuring fields without ranking. They say:

For many tasks at which experts make a living, no measurable outcome exists. How is one to know if the wine taster has judged accurately or if the professor has graded the essays well? Adherents of the expert performance approach would question the merits of studying such domains. Although there is no hint of an objective
external criterion, we believe that some people do these tasks better than others and that people improve their performance. (Weiss & Shanteau 2003: 105)

Unfortunately, when looking closer at their index, it pertains to expert judgments and evaluations and not at the type of expert performance present in interpreting.

Recently, Muñoz Martín (forthcoming) suggested an adaptation of Shanteau’s five components to translation expertise. Muñoz Martín suggests five dimensions that consist of (1) knowledge, (2) adaptive psycho-physiological traits, (3) problem-solving skills, (4) regulatory skills, and (5) the self-concept. These five components, although still lacking the concept of deliberate practice as a dimension of its own, are more appealing for studying expertise in interpreting. The five dimensions were put forward very recently and have not yet been empirically tested. For the present thesis the notion of deliberate practice is very much the crux of the matter. So for the purpose of the PhD project reported here, Ericsson’s expert approach and characteristics will be followed in the strictest sense possible.

### 2.3.2 The concept of deliberate practice in the expertise approach

As seen above, one aspect of Ericsson’s expert theory is the performers’ deliberate practice, a developmental feature of expertise. Ericsson divides the performer’s activity into three types (Ericsson, Krampe & Tesch-Römer 1993: 368):

1. **work** – an activity is defined as work when it is publicly performed and most often performed for remuneration.
2. **play** – an activity is defined as play when it is performed without remuneration, and without a particular goal for the activity, the performer’s pleasure during activity is an important part of play.
deliberate practice – an activity is defined as deliberate if it is performed at a clearly delimited occasion, with specific exercises (often decided beforehand). The activity is also performed with a clear goal to improve or refine the activity and with an evaluation of the performance.

Experts can be said to be constantly challenging the status quo of their performance, and the expert’s deliberate practice is the instrument for that challenge. Highly skilled performers can either stagnate in an automatized mode, or they can excel in expertise by engaging in deliberate practice (Ericsson 2007a: 685). Neither arrested development nor an automatized mode should be confounded with routine expertise (see section 2.3.1). Routine experts have not necessarily stagnated in automatized mode, but nor do they necessarily adapt their expertise to new challenges as adaptive experts do. Deliberate practice is the counteraction to stagnation. The performer who engages in deliberate practice does so over longer periods of time, and the occasions of deliberate practice are focused and well-planned. The performers’ practice is also analysed either by the performers themselves or by their peers or coaches according to the set goals or expected levels of achievement (Horn & Masunaga 2007). The planning and evaluation of the practice is thus what contributes to the development of expertise. Horn and Masunaga also define deliberate practice as

focused, programmatic, carried out over extended periods of time, guided by conscious performance monitoring, evaluated by analyses of level of expertise reached, identification of errors, and procedures directed at eliminating errors. (2007a: 601, my italics)

Ericsson adds that

the core assumption of deliberate practice is that expert performance is acquired gradually and that effective improvement of performance requires the opportunity to find suitable training tasks that the performer can master sequentially. (2007a: 692)
The individual’s deliberate practice is also guided by clear goals and openness to feedback. It is important for experts to be able to break down their activities into reachable part-time goals that can be achieved over shorter periods of time. Performers acquire expertise in their field gradually. Feedback and learning from peers are also important activities in the development of expertise. Experts develop through feedback from coaches, and by observing their peers (Ericsson 2007a: 692).

Although it may be challenging, as has been laid out above, to map and measure superior performance in simultaneous interpreting, the concept of how deliberate practice is executed over longer periods of time seems even more challenging to observe and investigate. It cannot be measured through experiments, but must rather be studied through interviews or journals. Studies that look at subjects’ deliberate practice over time in other fields include Sosniak (2007), who used retrospective interviews, and Deakin et al. (2007), who used journal studies. Deliberate practice can also be studied on a micro level, where the use of practice techniques at one particular (often experimental) occasion is studied. The techniques used at this particular session can then be compared between highly skilled performers and less skilled performers (cf. Zimmerman 2007). The fourth article in this thesis is devoted to an in-depth interview study of skilled interpreters’ deliberate practice. Prior to the actual interview study, a pilot focus-group study was made (Tiselius 2010). Two un-moderated focus group discussions were carried out over different themes in interpreting. The participants were conference interpreters of the Swedish booth at the European Parliament, both male and female and with a wide age and experience range. The aim of the study was to explore the sociological aspect of expertise, that is, how interpreters viewed their colleagues, work, customers and so forth. The focus group study showed a terminological challenge connected with the concept of deliberate practice. Participants did not intuitively understand the concept of deliberate practice. As a consequence, the analysis of the in-depth interviews in article 4 required
the researchers to interpret participants’ responses to different trigger questions in order to study deliberate practice. Deliberate practice is a scientific research concept, a theoretical construct. It is not necessarily the pedagogical or professional term used by professionals themselves. Any interview or questionnaire on deliberate practice will have to address how professionals in a particular field label that particular type of practice and how to make them talk about the different parts of that construct. Participating in an interview could possibly also qualify as deliberate practice, a type of learning through introspection. A full account of the methodology is given in article 4.

As has been shown, deliberate practice is a crucial part of Ericsson’s expertise approach. Practice and preparation are also important features of interpreting, as discussed above under section 2.2. When studying expertise in interpreting, it therefore seems inevitable that we must investigate interpreters’ possible engagement in deliberate practice as well as their practice habits.

2.3.3 Research on expertise in interpreting

Expertise research in interpreting studies was briefly introduced in section 1 of the introduction. The first part of this section gives an overview of the subjects in earlier studies on interpreting expertise, their profiles, and how are they labelled. The definition of an experienced interpreter or even an interpreting expert differs a lot as shall be seen. The second part of the section will look at results of research that contrasts experienced interpreters with subjects with little or no interpreting experience.

In one of the first articles in interpreting studies on expertise, Hoffman (1997: 192–193) wrote that “both psychological research on expertise and expert system development efforts have actually tended to define expertise rather loosely, or variously […]. A general challenge to scientific psychology is to generate a definition of expertise that focuses on
cognitive functionality and yet can be used operationally to identify experts”. Moser-Mercer et al. (2000: 108) wrote that “in translation and interpreting it is often acknowledged that the student obtaining his final diploma can call himself an expert with some degree of justification, but that years of experience in the field are still required for him to become a full-fledged professional”. This statement may hold true for the profession, but it is a rather unfortunate wording for interpreting expertise research. In a relative view of expertise, a recent graduate from an interpreting programme will most likely have more expertise than a student, but following Ericsson’s expert characteristics the recent graduate is probably far from being an expert. In his review of interpreting research, Hoffman (1997: 199) divides performers into different categories depending on their former experience of interpreting. He calls subjects without any interpreter training or professional interpreting experience a naive, pointing out that “novice” is actually a misnomer of “naive” in many studies, as the term is used for subjects without any previous knowledge of the field in question. Students are called novices when they start out in their interpreting programme, initiates when they have been initiated to a new skill (e.g. simultaneous interpreting), and apprentices when they are in their final stages or are recent graduates. Interpreters with five years of professional experience after graduation are labelled journeymen. This categorization loosely follows Dreyfus and Dreyfus’ (1980) proposal of a skill acquisition model comprising the levels of novice, competence, proficiency, expertise and mastery; this system has its roots in the terminology of crafts guilds. Following Hoffman’s categories, the subject with a recent interpreting diploma has reached the level of final-stage apprentice and can most likely not call himself an expert in absolute terms. In translation studies, Jääskeläinen (2010) asked the pertinent question of whether all professionals are experts. Jääskeläinen points out that early studies on translation process contrasted non-professionals and professionals, and it turned out that many of the professionals did not
produce any outstanding performances. Discussing definitions of expertise, she observes that although expertise research can be divided into two research approaches, *absolute expertise* (study of outstanding performers) and *relative expertise* (comparison between novices and experienced performers), the focus in translation studies has been on relative expertise (2010: 217). She furthermore points out that the professional participants in different studies may not have been completely screened following all expertise criteria (for example, several of the studies comprise participants with less than ten years of experience), and there is no investigation of the type of experience the participants have had.

Going back to interpreting studies and the studies published on expertise, there are several different definitions of experts or professionals in this domain as well. Vik-Tuovinen (2006: 129) points out that sorting the participants in interpreting studies into different experience categories as the ones mentioned above is not a very straightforward task. Reviewing a few studies of expertise in interpreting supports Vik-Tuovinen’s observation. In Chincotta and Underwood’s (1998: 8) study on bilingual digit span, the professional group (n=12) comprised simultaneous interpreters with at least 100 hours of interpreting practice, while the non-expert control group (n=12) consisted of students of English with no interpreting experience. Ivanova (1999) studied professional interpreters with an average of nine years of experience (n=8) in an expertise study on discourse processing; the novices in her study were interpreting students with three months of experience (n=8). Moser-Mercer et al. (2000: 126) compared the performance of novice and expert interpreters in three different experiments: in the first experiment the professional interpreters (n=5) had between five and ten years of experience, while the students (n=5) were recruited from the first semester of the interpreting programme; in the other two experiments, they labelled their participants professionals interpreters and students without specifying their background
(the number of participants was n=6+6 and n=5+5 respectively, and five of the participants may possibly be the same as in the first experiment). In a PhD dissertation on working memory (Liu 2001: 32) and a subsequent article (Liu et al. 2004: 24), Liu studied professional interpreters (n=11) with at least one year of full-time interpreter training and at least two years of professional experience with at least 40 days of interpreting per year. She had two groups of novices in her study, one were students at the end of their first year of the interpreting programme (n=11), the other group was at the end of their second year at the interpreting programme (n=11). Vik-Tuovinen (2006: 127) included preparation and debriefing in her study of interpreting at different experience levels. Her expert interpreters (n=7) had professional experience of between one to fifteen years, with 10 to 100 days of interpreting. There were two groups of novices in Vik-Tuovinen’s material: the first one was recruited at the first term of the interpreting programme, and at the time of the recordings they had 10 hours of interpreting practice (n=6); and the second one was recruited at a later stage in the training where the students had well over 180 hours of interpreting practice (n=8). Köpke and Nespoulous (2006: 6) studied the differences of working memory between novices and experts. They recruited 21 professional interpreters (12 staff and 9 freelancers) who had between 4 and 35 years of experience. The interpreting students (n=18) who participated in their study were recruited in their second and final year of interpreting studies, and they had just started simultaneous interpreting practice. Köpke and Nespoulous also had two control groups, one consisted of bilinguals (n=20) and served as the control group for the professional interpreters, the other one consisted of students (n=20) and was the control group for the interpreting students.

Though this overview does not claim to cover all studies that have been done with an expert approach or within the expert theory in interpreting, it is quite clear that no consistent profiling exists of either the highly experienced or the less experienced participants. It shows how persuasive
the expert concept is in interpreting studies. In most studies (including my own article 1, because of an unfortunate misunderstanding in the editing process), experienced and professional interpreters are indiscriminately labelled experts regardless of length of experience or performance. In her literature review on experts and interpreting, Liu (2008: 160) points out that since studies on expertise in interpreting use the relative concept of expertise, this entails that when more experienced interpreters are compared with less experienced ones, then any more skilled group can be considered experts and any less skilled group novices. The terminological issues of expertise are the same as Jääskeläinen (2010) emphasized above. Nevertheless, although there may be many indications of expertise, it is impossible to conclude without prior screening that the professional interpreter actually is an expert in the strictest sense of the expertise theory definition. Moreover, the description of the participants or subjects needs to be minute in order to enable literature reviews, study comparisons and replication.

It is clear from the different suggestions of developmental categories by Dreyfuss and Dreyfuss (1980) and Hoffman (1997) described above, as well as from the studies reviewed in this section, that the novice-expertise dichotomy is in fact a continuum. There are other dichotomies at play as well: for example, a student of interpreting acquires interpreting experience in class but no professional experience, while a professional interpreter can have professional experience without having undergone a training programme. The continuum and the different dichotomies in play are described in figure 1. Interpreters who have graduated from an interpreting programme and started to receive remunerated work are professionals, as are experienced interpreters and also expert interpreters. Although novice interpreters can develop into experienced interpreters and also become experts, experience in itself will not make them experts, and far from all professional interpreters can be labelled experts in Ericsson’s terms. In this dissertation, the participants of the different studies are
labelled laypersons (meaning they have no experience of interpreting neither from training, nor professionally), interpreters-in-training, short professional experience interpreters and long professional experience interpreters (two groups with fifteen and twenty-five years of experience, respectively). The laypersons have no previous experience of interpreting, but participated in the same interpreting experiments as the other participants. In the case of the no-experience subject, experience refer to interpreting experience, they may have many other areas of experience. These labels are added in the last line of figure 1

<table>
<thead>
<tr>
<th></th>
<th>Laypersons</th>
<th>Students of interpr.</th>
<th>Recent graduates</th>
<th>Interpreters w/exp short</th>
<th>Interpreters w/exp long</th>
<th>Expert Interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof. Interpreting Exp.</td>
<td>No</td>
<td>No</td>
<td>Possible</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Interpreting Exp.</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*Figure 1. Novice-Expert continuum.*

Findings from studies of expertise in interpreting have been thoroughly reported in Liu (2008). She reviews an important part of the cognitive, empirical studies made on interpreters with professional experience, and cognitive, empirical studies where interpreters with professional experience are compared with students of interpreting or subjects without interpreting experience. Liu approaches the studies from the perspective of interpreting skills. She divides the interpreting skill in three parts, as she suggests that there are three obvious processes in interpreting comprehension, translation and production (2008: 161); these three main skills are then divided into sub-skills and cognitive abilities (concurrent articulation, articulatory suppression, working memory and attention
Through findings from different studies (e.g. Goldman-Eisler 1972; Barik 1975; Chernov 1979; Dillinger 1989; Isham 1994; Bajo, Padilla & Padilla 2000; Liu 2001) that have compared interpreters with different levels of interpreting experience and their performances in these cognitive areas, Liu distils some common features of the sub-skills among experienced interpreters that may serve as an indication of expertise in interpreting. Results from the different studies show that experienced interpreters are better at semantic processing than subjects without interpreting experience, and also that experienced interpreters are better at selecting the most important meaning units when circumstances called for that. These results also echo the findings by Vik-Tuovinen (2006), who found that the interpreters with short interpreting experience in her study focused more on the meaning of single words, whereas the interpreter with professional interpreting experience focused on understanding the content. Liu (2008: 164) goes on to show that studies have found that experienced interpreters process longer chunks than subjects with shorter or no interpreting experience, and that experienced interpreters from English into Russian produced fewer syllables (2008: 165). She also mentions that few studies have investigated how experience affects the interpreter’s delivery. Vik-Tuovinen (2006: 305) found that the experienced interpreters her material were much more conscientious about delivering the best possible product to their clients. In terms of monitoring output, Liu (2008: 167) points to several studies on delayed auditory feedback that have shown that interpreters with professional interpreting experience are less disturbed by delayed auditory feedback. Working memory studies, including Liu’s own (2001), have shown that interpreters with longer professional experience have a larger digit span than their less experienced counterparts. When it comes to attention, Liu (2008: 173) mentions Cowan (2000/2001), who suggested two explanations for attention function in simultaneous interpreting: (1) rapid attention switching between listening and speaking, and (2) well-practised listening and speaking skills. Liu
argues that the studies she reviews support Cowan’s two assumptions, for instance, the fact that pauses and pause length increase as cognitive load increases, or the fact that subjects with little or no experience produce more fragmented output. Liu suggests that this is due to the experienced interpreters’ ability to have an overall perspective on the interpreting situation, a conclusion that also Vik-Tuovinen puts forward (2006: 308–309). It should be pointed out though, as Liu in fact does (2008: 160), that several of the studies Liu refers to compare experienced and inexperienced interpreters, rather than experts and novices in Ericsson’s strictest sense. Liu concludes from the literature review that expert interpreters produce fewer errors and faster responses and use less effort. She goes on to say that there is more to expert interpreting than speed and effort, such as qualitative differences of process and output. She observes that expert interpreters seem to have developed well-practiced strategies in each of the comprehension, translation, and production processes. […] These strategies are developed and practiced as a result of the interaction among the comprehension, translation, and production processes that are specific to the needs of simultaneous interpreting. […] It seems that expert interpreters have developed an ability to efficiently manage their attention so that it can be switched between different processes. (2009: 174)

Liu’s conclusions are by no means controversial or questioned here. But from the background of the participants in the material she reviewed, the results are possibly true for experts, but they are first and foremost true for the experienced interpreters who participated in the different studies. These interpreters are most likely both professional and experienced, but can they be called experts in absolute terms? Liu concludes that interpreting studies is only beginning to piece together the evidence to create a more coherent picture of expertise in interpreting (2008: 174). This is quite true, and in order to create this more coherent picture of
expertise we need to clarify definitions, as common definitions are a pragmatic way to determine what is comparable and relevant.

Finally, it should be pointed out that most, if not all, studies comparing different subjects with different types of interpreting experience are cross-sectional. One of the sets of data in this dissertation consists of participants that were recorded at two different points in time: first when attending an interpreting programme in the mid-1990s, and then in the present day for the purpose of this project. As the project developed, different ways of labelling these data was discussed. The term “long-term” was chosen over “longitudinal” for several reasons. First, a typical longitudinal study comprises several points of study or contact over a longer period of time, whereas this study only had two points of study (during the interpreting programme and fifteen years later). Second, as Hansen (in press) points out, if the first experiment is repeated and long periods of time go by between the first data collection and the second time, it can be questioned whether the exact same experiment can be performed, with the same methods and under the same conditions. In analogy to the German difference between Langzeitstudien and Längsschnittstudien, she decides to use “long-term study” as equivalent to Langzeitstudien, and in this text I will follow Hansen’s lead. Similar data sets as in this dissertation project have been used in translation studies by for example Hansen (2008), Azbel-Schmidt (2005) and Göpferich (2013).

2.4 Assessment

Up until this section, the discussion of the theoretical background has dealt with the processes of interpreting and how to study them. This section deals with the product of interpreting. Assessment and evaluation of interpreting assume some type of approach to quality. But the purpose of this PhD project has not been to determine or define quality in interpreting, and looking at their interpreting product is but one possibility for
investigating the difference between groups of subjects. Presumably, interpreters with long professional interpreting experience deliver a product of higher quality than subjects with little or no interpreting experience. In order to find out whether there is a quality difference in the product between these different groups, the product has to be assessed. It should be stressed that the aim of the assessment (and the instrument used) was to test the main skill of interpreters (the interpreting skill) rather than different sub-skills (e.g. language, memory or coordination).

How to assess interpreting is another crucial topic in interpreting studies. The first article of this thesis is devoted to developing an assessment tool, and three of the four articles discuss the evaluation of interpreters. As Angelelli and Jacobson (2009: 3) point out, few assessment instruments or methods in interpreting are based on valid and reliable measures stemming from empirical research. In order to remedy this, they suggests a holistic, rubric-based system that can be tested for validity and reliability (2009: 38–39). They note that there is a tension between theory and practice when it comes to assessment, and that “practitioners believe that expertise in testing is obtained by practical experience” (2009: 45). The trust in practical experience may be a reason for the relative lack of documented testing instruments for interpreting performance. This in turn may also underlie the manifold flora of testing instruments in the research literature. Many researchers develop their own instruments, or rely on traditional, intuitive grading. Kalina (2005) proposed several instruments in order to assure quality by assessment, establishing a model where interpreting should be assessed not only from the output on task, but also from all the different features involved in creating high quality before and after the interpreting assignment (2005: 780). In a more recent contribution (Kalina 2011: 169), she proposes a protocol for assessing students in exams. Both of these proposals are of a componential type. In light of the many questionnaires that have been used with the aim of establishing how quality is perceived by both users of interpreting and interpreters, Moser-
Mercer (2009) stresses the importance of defining the construct of quality “clearly, precisely and unambiguously” (2009: 146), stating that the construct has to be operationalized in order to design valid and reliable measures.

Validity, whether the instrument is actually measuring what it was designed to measure, is crucial in this context. When it comes to validity, Collados-Aís (2011) and her team have made important contributions. In a series of studies they have broken down the components that are typically present in interpreting assessment, for instance, accuracy, accent and speed, and have shown that the assessment of these components are not necessarily valid as other components affect the evaluation too. For instance, the exact same speech received lower scores for accuracy if delivered with a foreign accent, despite raters’ pre-assessment claim that accent was unimportant.

Clifford (e.g., 2001, 2004, 2005) has also contributed to the field of assessment. He points out that assessment has often been linked to a text-semantic similarity or exact reproduction. This stems from the view of the interpreter as a neutral conduit rather than a participant in the event. As explained by Clifford,

\[ \text{the conduit portrays interpreting as an exercise carried out on linguistic forms, one in which even the smallest changes in perspective are not permitted. As noted in the literature, the conduit has at times been called the traditional perception in interpreting [...] its central perspective [...] and even its ideal [...]. (2004: 92)} \]

In an earlier article, however, Clifford (2001: 366) argued that assessment of interpreters should determine whether they have the competencies required for professionals, and he asked the pertinent question of which competencies need to be assessed in professional interpreting. Clifford also refers Berger and Simon’s (1995, cited in Clifford 2001: 373–374) four-step assessment cycle: (1) intention (what is being assessed and why);
(2) measurement (data collection and marking); (3) judgment (when judging, a common system, understood by all assessors, must be used); and (4) decision (a decision is fair and equal if previous steps are followed rigorously). Clifford also cites Berger and Simon’s principles of quality assurance in assessment (1995, cited in Clifford 2001: 375): (1) validity (the instrument measures what it was intended to measure); (2) reliability (it gives the same results in tests and re-tests); (3) equity (assessors are aware of possible gaps in performance between different groups; (4) utility (an instrument is practical to use in any given situation, i.e. not too expensive, complicated or bulky); and (5) comparability (the test can be compared although different conditions apply, e.g. different language combinations).

Although not irrelevant for an assessment in a research study, Clifford’s assessment cycle may be less applicable, but several of the quality assurance principles should be as important for research as for practical applications. When the instrument used in this PhD project was developed and adapted to the studies, great care was taken to ensure validity, reliability and utility, as can be seen from the presentation in Article 1. In the case of this thesis, an instrument used earlier used by Anderson (1979) – namely Carroll’s two scales (1966), one for intelligibility (whether the interpretation in this case is understandable in the target language) and one for informativeness (how much of the information from the source language message that is kept in the target language message) – was chosen for further investigation and development. The adaptation of the scales to this thesis will not be discussed here, as it is done in depth in Article 1 and below in section 3.1.2., suffice to say that they were chosen for their holistic and non-componential character. As for Clifford’s principle of equity in assessing quality assurance, the conditions for determining equity do not seem quite applicable to this type of research. Clifford’s last principle was tested to some extent, as the comparability was checked for assessment from audio files as compared with transcripts.
The comparability of the scales will have to be determined in the future, however, when they are hopefully used in other studies. One thing that springs to mind when re-reading Clifford’s 2004 article is that an exact match between interpreting and the original is not desirable, as that would reduce the interpreting into an incomprehensible word-for-word rendition. Hence, full score for informativeness in Carroll’s scales may not represent an ideal interpreting.

It is also important to discuss who should rate the interpreting product. Interpreters have knowledge of the interpreting process and of what is required of the interpreting product. They are also trained to assess themselves and their peers from the interpreting programme. On the other hand, they may be familiar with, or even friends with, the interpreter who is being rated, which in turn may bias the rating. Furthermore, interpreters are not the end users of the interpreting product and might not share their perspective on what is important. In a scientific study, it may seem natural that the researcher assesses the interpreting. But the researcher may also be biased, whether by knowing the subjects or by meta-knowledge of the interpreting process. In the case of the present PhD project, both interpreters and laypeople (i.e. non-interpreters) were used as raters.

Assessment is an important part of investigating the interpreters’ product. Even though quality is, and has been, a hot topic in interpreting studies since its beginnings, thoroughly researched assessment instruments are still lacking.

2.5 Research questions and methodological development

The research aims of this PhD project were twofold. As described in section 1.2, the project had both methodological goals and research questions. The research was guided by the following questions:
1) Is there a measurable difference in the interpreting skill from the student level to the highly experienced level?
   a. It was assumed that there would be a measurable difference in the interpreting skill.
2) Is there a measurable difference in the interpreting skill both when it is measured cross-sectionally (i.e. inter-individually) and long-term (i.e. intra-individually)?
   a. It was assumed that there would be a measurable difference in the interpreting skill regardless of data.
3) If there is a measurable difference, what does this difference consist of?
   a. It was assumed that there would be a difference in the assessment.
   b. It was assumed that there would be a difference in the interpreting process.
4) How do experienced interpreters perceive different factors in their long-term competence development?
   a. It was assumed that experienced interpreters would claim that they practise a lot.
   b. It was assumed that experienced interpreters would claim that they constantly strive to improve their interpreting performance.
   c. It was assumed that experienced interpreters would be able to talk about their goals, on both the micro and macro levels.
   d. It was assumed that experienced interpreters would claim that they made use of their colleagues for feedback and help.
   e. It was assumed that experienced interpreters would be able to describe how they work under pressure.

The methodological development comprised the following questions:

1) Can holistic scales for measuring intelligibility and informativeness be developed into a valid and reliable measuring instrument for quality in interpreting?
   a. It was assumed that the scales could be developed and tested so as to form a valid and reliable measuring instrument for quality in interpreting.
2) Will holistic scales work equally well as a measuring instrument whether used by laypersons or experienced interpreter raters.
a. It was assumed that the scales could be used by different raters and still generate valid results.

3) How should an in depth-interview be carried out in order to yield results on the concept of deliberate practice?

a. It was assumed that an interview guide had to be created where participants would be prompted to discuss issues of deliberate practice without being familiar with the research concept of deliberate practice.
3. Data and methods

This section will provide the methodological background for the four articles, discuss the rationale for choosing these particular methods and instruments, and present the participants.

It should be stressed at this point that I am an active conference interpreter myself. I have strived to study my material and conduct my project from an etic perspective, using a scientist-oriented approach. It is, nevertheless, impossible to completely shed my emic knowledge and bias. I hope, however, that I have been sufficiently accurate in my research design and the presentation of both method and results so as not to bias my findings.

Several of the research questions deal with methodological issues, and it has already been pointed out that one of the aims of the PhD project was to develop and test different methodologies. This section will therefore also include a methodological discussion. The instruments used will only be briefly introduced, however, as they are thoroughly discussed in the articles.

3.1 Methods

The first part of this section will deal with different methods for collecting and analysing data.

3.1.1 Investigating expertise

The data in this project have been taken from two groups: a long-term group and a cross-sectional group. The long-term group is unique: the participants in that group were recorded for research purposes for the first time when they attended an interpreting programme in the mid-1990s. Much of the work on this thesis has revolved around which methodological approach would be the most suitable to make use of this
unique material, and the methods have also been developed and refined with this in mind. As will be shown below (3.2.1 and 3.2.2), both the experienced interpreters of the cross-sectional material and the participants in the long-term material, after having gained more than fifteen years of experience, all showed superficial signs of expertise. They had reached the levels of contributory expertise labelled by Collins and Evans (2007: 14). The participants had credentials, experience and a track record, and have acted as examiners, peer-reviewers and so forth. Though they are experts according to Collins and Evans’ terminology, they have not been tested according to Ericsson’s criteria, the more cognitive side of expertise. Though it was not the main aim of this thesis to determine whether or not these participants are experts in Ericsson’s terms, the project’s various measurement will presumably provide some insight in that regard.

As pointed out above, the expert performs consistently at a superior level compared with other performers. The investigation of performance is therefore central when studying expertise. In her literature review, Liu (2009) observed that the experienced interpreters in the studies she reviewed showed qualitative differences of both process and output compared with less experienced subjects. Ericsson and Smith (1991: 8) want research within the expertise approach to describe the critical performance under standard conditions. The performance should be analysed, and the components that make it superior should be identified. Since interpreters with long professional interpreting experience and subjects with little or no interpreting experience seem to differ in aspects both product and process, it was deemed important to study both aspects of the performance. Social implications of expertise, although important, have not been investigated in this work, simply for delimitation reasons. A pilot study on sociological aspects of expertise in interpreting was however done as part of a research training course (Tiselius 2010). This study is only referred to from a methodological perspective, however, and not as part of the thesis.
3.1.2 Investigating the process

Englund Dimitrova (2005) convincingly demonstrates the prudence of not drawing too many conclusions on the process simply on the basis of the product. In her study she showed that a hypothesis based on the textual evidence was refuted by process data (2005: 36). It is therefore wise to study both process and product. For investigating the process in interpreting in a non-invasive way, however, methods and instruments of data collection are in sore need. Introspection is a way to make both tacit knowledge (Collins 2010: 4) and invisible processes explicit and accessible. When studying interpreting, concurrent introspection is not available – the interpreter cannot verbalize at the same time as s/he is interpreting. Immediate retrospection is therefore one of the few introspective data collection methods at hand for tapping into the process. Retrospection has a few additional challenges compared with introspection. For example, it needs a cue in order to be appropriate, particularly for longer retrospections (in the case of this thesis, the task was between roughly nine and eleven minutes). It also needs to be immediate: since participants can only be expected to completely remember and verbalize a task of 2–10 seconds (Ericsson & Simon 1993: xvi), the longer time that elapses between task and retrospection, the more of the process is likely to be forgotten. The role of the researcher and the instructions given to participants are also important. The researcher is not a participant in the intro- or retrospective interviews. It is therefore key that the instructions are clear and that the researcher is positioned so as not to invite interaction (preferably obliquely behind). After introductory instructions have been given, the participant is told to keep talking. A potential pitfall is that the participant might start to explain and describe the process rather than just verbalize it. Ericsson also mentions objections raised in psychological research against using the subjects’ own verbalizations as scientific data (1993: 1). He argues that information processing models of the cognitive process make it possible to create an
explicit and objective encoding process, which in turn enables the data to be examined objectively (1993: 4). Finally, when analysing the coded protocols, one must keep in mind that subjects might forget, recall something different than the actual process or slip into explaining or describing the process. The use of retrospective protocols and their challenges for both translation and interpreting have been explored in two articles by Englund Dimitrova and Tiselius (Englund Dimitrova & Tiselius 2009; Englund Dimitrova & Tiselius, submitted). These articles are not part of the dissertation but have been essential in understanding and developing retrospection as one of the data collection methods in the thesis. Rather than choosing a single component of the process for further study, I chose retrospection despite its potential limitations, since it was considered to be the least invasive and to possibly provide broader insight into the whole process. Another advantage of choosing retrospection was that it allowed comparison with Ivanova’s (1999) study. Ivanova used immediate retrospection with a transcript of the source language speech as cue. She sides with Ericsson and advocates protocol studies “as the most suitable of all currently available methods for the study of skilled and expert performance” (1999: 164). Ivanova also stresses the importance of coding the protocol with an open mind, and that coding for open-ended tasks such as simultaneous interpreting can be done on a more global level with for instance strategy use during the task in mind (1999: 165). There are very few studies within the same theoretical paradigm of expertise in interpreting studies. Hence, it is a strength for this PhD project to be able to compare the results of other studies. It should be stressed however that the data in the present PhD project are only comparable with a small part of Ivanova’s data. Retesting will nonetheless give reliability and validity to the methodology. The categories with explanatory notes can be found in article 2, appendix 1 and tables A.1, A.2 and A.3.
3.1.3 Investigating the product

As discussed above under section 1.4, the decision to look at the product and not only the process started a lengthy process of choosing and testing the instrument for the study. At first, I believed that the starting point should be to define how high-quality interpreting should be understood. An example of such a definition could be the European Parliament’s description of interpreting, cited in Vuorikoski (2004: 19), where it is stated that interpreting is not a word-for-word translation, but a faithful transmission of the source language message, rendered accurately in the target language. But as Vuorikoski noted in her study (2004: 22), the present dissertation also found that there are too many variables in interpreting, and too many types of interpreting, to identify a static, all-inclusive definition of high-quality interpreting. Or in Pöchhacker’s (2004: 153) words, “quality appears not as a self-contained topic but as a complex, overarching theme in which all aspects of the interpreter’s product and performance – textuality, source-target correspondence, communicative effect, and role performance – play an integral part”. Furthermore, as mentioned above, the works of Collados Aís (2011) point out the difficulties of componential assessment that seem to stem from an idea of absolute quality.

Then I discovered Caroll’s scales through the work of Gerver (1971) and Anderson (1979; 1994). These holistic scales were developed for machine translation and used by Anderson and Gerver, but were not used in later studies in interpreting studies. Though the authors of these later studies did not seem to be disappointed with the use of the scales, nor dissuade readers from using them, the research community nonetheless seemed to prefer more componential assessment methods. But since the scales seemed appealing from a holistic perspective, I decided, as described in section 1.4 above, to try them out for this project.
3.1.4 Investigating the participants

The long-term participants (see below) in this study had agreed to participate in in-depth interviews about their interpreting background and their views on interpreting. The interviews in this study were modelled on Kaijser and Öhlander (1999). The interview method chosen was quite different from the structured retrospections described above in section 2.1.2. This is a type of interview similar to the one Koskinen (2008) used in her study of translation in the European Union. The interview is structured inasmuch as the parties involved understand that it is an interview, time is set aside for the interview and the researcher has a clear objective for the interview; moreover, an un-structured interview does not use predetermined questions, but rather a mind map of topics (or something similar) that s/he would like to discuss. The discussions flow without constraint, and it is crucial to allow follow-up questions in all relevant directions. The interviewer also actively participates in the interview, and the interviewer’s identity, for instance as an expert in another field (researcher) or in the same field (colleague), is important. A drawback of this method is that the subjects do not necessarily answer the same questions. Article 4 discusses the methods and results of these in-depth interviews.

3.2 Participants

All the participants in the different studies of this project had Swedish as their mother tongue. All groups of interpreter-subjects with professional interpreting experience (n=9) had English as a passive working language (C-language in AIIC terminology, that is, a language the interpreter has full understanding of and works from but not into). The subjects in interpreter training (n=3) had English as one of their passive working languages in the training programme. The no-experience subjects (n=3) had English as a strong foreign language; they were not screened,
however, and merely self-rated their proficiency. The professional interpreter participants (n=9) constitute an undeniably small group, but it should be stressed that the entire population of English-Swedish conference interpreters is also very small. AIIC lists 34 interpreters with that combination; even if that figure is boldly doubled (given that not all conference interpreters are members of AIIC), it still makes a population of N=68. Using that estimate it would mean that the participants make up approximately 13 % of the entire population; in that perspective, the group may be small, but would still be representative. Furthermore, the fact that the data contain both a cross-sectional and a long-term group make both inter-individual and intra-individual comparisons possible.

3.2.1 The cross-sectional data – data set A

The cross-sectional data include nine participants, divided into three groups according to the interpreting experience of the participant (see table 1, also reproduced in Article 2). None of the participants received any economical remuneration.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age span</th>
<th>Years at university</th>
<th>Int. training diploma</th>
<th>Years of int. experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>No experience</td>
<td>20–29</td>
<td>4</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Short experience</td>
<td>30–49</td>
<td>4</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Long experience</td>
<td>50–60</td>
<td>4</td>
<td>Yes</td>
<td>25+</td>
</tr>
</tbody>
</table>

The no-experience group (from now on the NE group, cf. figure 1 above) consists of students recruited at the Institute for Interpretation and Translation Studies at Stockholm University. The NE group comprises three females. They were recruited from the first semester introductory course to translation and interpreting: students there were sent an e-mail asking them to participate in a study; several students volunteered, and those with Swedish as their mother tongue and who claimed to have a high proficiency of English were chosen. Although they had no prior experience of interpreting, the NE group had been in an interpreting booth
during their introductory course and were thus familiar with the concept of interpreting. The NE group did the interpreting task in the training booths at the Institute for Interpretation and Translation Studies at Stockholm University.

The short interpreting experience group (from now on the SE group, cf. figure 1 above) consists of interpreters who had graduated from an interpreting programme two years previous to the data collection event. The SE group comprises one male and two females. They were all accredited to the European institutions, but were not yet AIIC members.¹ AIIC membership is subject to application and screening, and it is fair to assume that the subjects were not AIIC members because of their fairly limited professional experience (they would be eligible for AIIC membership first after 150 working days). Since the SE group consisted of recent graduates, they had not had any experience with teaching or assessing interpreting, though they did have experience with self and peer assessment during the interpreting programme. Their recruitment to the study was based on the number of years of experience after graduation, but they also constituted a convenience sample in the sense that they were directly approached through e-mail and asked if they were willing to participate in a study. The SE group performed their task at the researcher’s workplace and not in a booth.

Finally, the long interpreting experience group (from now on the LE 25 group, cf. figure 1 above) consists of highly experienced interpreters. It comprises one male and two females. They had all the superficial signs of expertise. The LE 25 interpreters all had a diploma from an interpreting programme. They had been working actively for at least twenty-five years, with an average of at least 100 days per year. They all had both teaching

¹ AIIC membership is gained with at least 150 days of working experience and the signature of three AIIC members guaranteeing the applicant’s quality and work ethics; see http://aiic.net/node/2395/joining-aiic/lang/1 (accessed 12 April 2013).
and assessment experience. They were accredited freelance interpreters at the European institutions and were AIIC members. All interpreters of the LE 25 group were furthermore considered “good interpreters” in the interpreting community. As was also shown under 2.1.1, the LE 25 group could be defined as having contributory expertise (Collins 2007: 14). It was a convenience sample, in that I asked fellow interpreters with the right profile (i.e. accreditation to EU institutions, AIIC membership, experience as interpreting teachers and examiners, long professional interpreting experience, and, finally, a reputation for being “good interpreters”) whether they would like to participate in a study. The LE 25 group participated in the experiment in the booth at their workplace outside of working hours.

### 3.2.2 The long-term data – data set B

The participants in the long-term group were recruited the first time when studying at an interpreting programme in the mid-1990s. Williams (1995) was designing a major study on processes in simultaneous interpreting that aimed to study factors such as anomalous stress, prosody and pitch in interpreting. For the purpose of the project, she recruited both interpreting students and professional interpreters. The students recruited in her study had several different combinations of working languages. They were recorded interpreting in both simultaneous and consecutive mode and when talking freely (both in their mother tongue and their foreign languages). For this dissertation, the tapes from Williams’ project were generously made available from the Centre for Research on Bilingualism in Stockholm. The tapes were studied and four possible subjects were identified. The subjects were identified on the following criteria: (a) having Swedish as their mother tongue; (b) having English as a C-working language; (c) remaining active interpreters; and, (d) most importantly, willingness to participate in a new study. Moreover, the interpreting on the
tapes needed to be from the same source speech in order to make comparisons between the subjects possible.

Of the four possible participants, three were available at the time of the new recordings. They were all staff interpreters at the European institutions and had been so for some fifteen odd years. They had experience both as interpreter trainers and as assessors, and they were all considered “good interpreters” by their peers. One was an AIIC member. At the time of the new recordings, this group had also reached levels of contributory expertise (Collins & Evans 2007). All three were female, and they kindly agreed to participate in new recordings and in-depth interviews. As described in figure 1 above, the subjects in data set B will be referred to as subjects in training (the IT group) when their student data are referred to and as long experience interpreters (the LE 15 group) when their professional data are referred to.

### 3.2.3 The raters

Though the raters are not the main participants per se, they are nevertheless important to the studies. The raters were interpreters and non-interpreters who rated the quality of the interpretings using the holistic scales. They all had Swedish as their mother tongue. There were several different groups of raters, two for data set A and two for data set B. For data set A (cross-sectional), the raters were (a) university students without previous specific knowledge of interpreting (n=6), and (b) interpreters with professional interpreting experience and experience with examination and peer assessment (n=6). They are described in more detail in article 1 of this dissertation. For data set B (long-term), the raters consisted of (a) university students without previous specific knowledge of interpreting (n=12); (b) interpreters with professional interpreting experience and experience from examination and peer assessment (n=12); and (c) another set of university students without previous specific knowledge of interpreting (n=9). Groups (a) and (b) rated the NATO speech (see section
3.3.2 below) and group (c) rated the EU speech (see section 3.3.1 below).
The raters for data set B, and the rationale for the different raters of the
different speeches, are described in more detail in article 3 below.

3.3 Interpreting Data

This section presents the two speeches that were used to elicit interpreting
data from the subjects. All subjects interpreted the EU speech, and the
long-term group also interpreted the NATO speech twice with fifteen
years in between.

3.3.1 The EU speech

All subjects (n=12) interpreted a speech from the European Parliament,
given originally by Commissioner Byrne in 2001. It was a fairly general
speech, but very fast (141 words per minute on average) and also
pronounced with a heavy Irish accent. For this project, the speech was
transcribed and tweaked to add some additional difficulties (names and
figures). The speech was then re-recorded by a native English speaker with
Received Pronunciation. In its re-recorded version, it was also controlled
for speed (119 words per minute on average). The EU speech can be found
in article 2, appendix 2.

3.3.2 The NATO speech

The three long-term participants also interpreted a NATO speech that was
used as teaching material during their training. The speech (and the
interpreting from the IT group) was chosen for the following reasons: (a)
the point in time in the training programme (the chosen speech was
delivered when the students had been practising in the simultaneous mode
for a couple of months, and the speech would then supposedly also present
some difficulties for the LE 15 group); (b) it was not an exam-level speech
for the students (as exam-level speeches have their own genre); (c) relative
difficulty (speed, terminology, themes; as said above, it was important that the speech should present some challenges for the LE 15 group and yet not be too difficult for the IT group); and (d) sound quality of both speech and interpretings (these were old tapes, some of them recorded on small recording devices, and it was important to hear clearly, both for interpreting and transcription). The NATO speech can be found in article 3, appendix 1.

3.4 Retrospective data and assessment files

As described above in section 3.1.1, all participants, except the IT group, performed retrospection immediately after the interpreting task. The retrospection was cued with a transcription of the source speech and then recorded and transcribed. The transcripts served as protocols in the categorization process following Ivanova’s categories (1999). The categories are found in article 2, appendix 1.

The interpretings in both the cross-sectional and the long-term data were transcribed and transformed into assessment files. The transformation consisted of dividing the interpretings into smaller units according to idea and then mixing them randomly. This is described in detail in article 1, and an example of the assessment files can be found in article 1, appendix 2.

3.5 Methodological discussion

This section will reflect on some of the methodological issues and challenges that have not been touched upon earlier in this account, such as the use of mixed-method design and the choice to pursue the investigation despite a very small material in the long-term study.
3.5.1 Mixed-method design

Just as in other disciplines, researchers have highlighted the benefits of triangulation in interpreting studies too (see for instance Gile 2005a or Hild 2007). In trigonometry and geometry, triangulation means finding the unknown third point by using two known points; in social studies, triangulation is the use of at least three (but preferably more) different studies, theoretical perspectives, investigators and data sets to examine a certain topic (see for instance Denzin 2006 or Scott and Marshall 2009). Presumably, researchers would obtain more robust results by using a variety of means, such as different researchers or data sets, to investigate a certain concept or construct in interpreting. Not many studies use triangulation in interpreting studies, however, and the few researchers who do in fact triangulate use mostly the within-method, that is, they triangulate with different varieties of similar methods (Denzin 2006: 472).

Quantitative method designs dominate in conference interpreting research. And many research objects, such as cognitive load, working memory or the effects of interpreting under pressure, are easily and more appropriately researched quantitatively. But there are other topics, such as the perception of role or identity, that are not so straightforwardly refuted or supported by a yes/no hypothesis, traditionally used in studies with a quantitative approach. Diriker (2004) and Monacelli (2009) are among the few who use qualitative methods to investigate conference interpreting. Monacelli investigated voice (not the physical voice in this case, but the voice as a representation of the interpreter’s persona) in simultaneous interpreting from a constructivist epistemology in order to study the speaker’s discourse and the interpreter’s rendering of that discourse. Diriker used critical discourse analysis and semi-structured interviews in her study, which aimed at distinguishing between different discourses on simultaneous interpreting and how the interpreters put that discourse into practice as they interpret.
Though both quantitative and qualitative research methods are in use in Interpreting Studies, mixed-method designs do not seem to be widely employed. Such designs can for example be used to enrich the understanding of quantitative results by providing certain insights into the subjects in question: for example, a questionnaire with multiple choice or Likert Scale answers can be illustrated by quotes from the open-ended questions at the end of the questionnaire (Patton 2002: 5). In short, mixed methods help the researcher to approach an object of study from different angles. When studying a particular area, such as expertise in interpreting (as will be described below), mixed methods may be used to enlighten parts of the issues raised by the expertise approach that cannot be reached through an experimental design.

The reason in this particular case for combining qualitative and quantitative methods lay in the struggle to find the best way to analyse an existing and unique material, namely that of the late Sarah Williams (1995). Her article from 1995 gives an overview of her intentions with the material. This material consisted of recordings of interpreting students in the 1990s, as described above in section 2.2.2. By deciding to use that material, it was also necessary to find out what made sense to the material, to paraphrase Quinn Patton (2002: 72). It also meant that instead of starting to work within a specific theory or with a clear hypothesis or research question in mind, the project began with a material, and research questions and hypotheses needed to be adapted accordingly. In addition, William’s material was recorded for other aims than this project. When the interpreters who were recorded as students had been identified and agreed to participate in new recordings, the challenge was to design a study that would yield interesting results from this unique data. Since the starting point was the possibility of obtaining a material that could pave the way for a long-term study, the theory chosen was the expertise approach. Furthermore, a mixed-method design was developed in order to analyse both interviews and experimentally yielded material. In order to obtain a
broader perspective, the long-term material was supplemented (as described above) with recordings of the same subjects fifteen years later and with the cross-sectional material that featured nine subjects. It is hoped that the mixed-method design in this thesis, although presented in different articles, will contribute to a fuller picture of the development of experience and maybe also expertise.

3.5.2 Re-test or not, and other challenges

As discussed above in section 3.1.2, an advantage of using the same methodology as part of Ivanova’s (1999) study was to be able to relate to her results. It would also provide stronger support for the results here. There are two challenges to this claim though. First, none of the studies reported here is an exact replication, since Ivanova’s two groups and the three plus two groups in this thesis are not identical; and although the source language is English in both cases, the mother tongue differs (Bulgarian in Ivanova’s case and Swedish in this case). Second, Ivanova’s novices are interpreting students in their second and final year at the interpreting programme. If the different groups are labelled according to Hoffman’s terminology (cf. section 2.3.3 p. 31), the novices in Ivanova’s study are apprentices, whereas the novices in this dissertation can be divided in three groups: novice bilinguals (NE group), apprentices (IT group) and journeymen (SE group). According to a strict scientific definition, where replication is to reproduce an experiment with the exact same conditions and with the aim to obtain the same results (Scott & Marshall 2009: 646), this is not a replication. It will however qualify as a re-test, or a re-study (Scott & Marshall 2009: 647).

The differences between the different groups pointed out above could also be of importance when the groups are compared using the Carroll scales.

An important drawback was the fact that one retrospective interview from the long-term material was lost because of a technical mishap. Since the
long-term material was scarce to begin with, the loss of that file was clearly a blow. I decided to continue analysing the material and carry on with the study, as the material was truly unique, and deserved every effort to be shared and analysed.

Despite the challenges and drawbacks of the material, it is my hope that the results will still be considered solid and interesting.

3.5.3 Terminological inconsistencies

Finally, it should be mentioned that because this is an accumulative thesis with some articles published before the project’s end, the terminology will vary slightly in the various articles. In particular, the subjects evaluating the interpreting will be variously labelled graders and raters. The terms subject and participant will be used interchangeably throughout the dissertation to refer to the members of the various groups. In social sciences, “participant” is often preferred over “subject” as researchers stress the active involvement of the individuals participating in a study. For the in-depth interviews, “participant” was thus more natural.
4. Summary of articles and general results

This section summarizes the results of the different parts of the study reported in four different articles. The material was collected as one whole unit that was studied and analysed from different perspectives, each reported in a different article. Article 1 is methodological, article 2 reports the results from the cross-sectional data and article 3 reports the results from the long-term study; finally, article 4 is a report of the in-depth interviews of the long-term interpreters. The mind map in figure 2 illustrates the different data sets, analysis methods and articles. The focus of this summary is mainly on results and somewhat on methodology, but as issues have already been discussed in section 3, the methodological sections have been kept at a minimum.

<table>
<thead>
<tr>
<th>Article 1</th>
<th>Eliciting material</th>
<th>Primary subjects</th>
<th>Raters</th>
<th>Retrospection</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article 1</td>
<td>EU-speech</td>
<td>Cross-sectional NE, SE, LE25</td>
<td>Interpreters</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Article 2</td>
<td>EU-speech</td>
<td>Cross-sectional NE, SE, LE25</td>
<td>Interpreters</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Article 3</td>
<td>EU-speech, Nato-speech</td>
<td>Long-term IT, LE15</td>
<td>Interpreters</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Article 4</td>
<td>Interview mind map</td>
<td>LE15</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Figure 2. Summary of participants, data, analyses and articles.

4.1 Article 1: “Revisiting Carroll’s scales” (data set A)

Article 1 was the first article published in this PhD project. The article used data set A, that is, material from the cross-sectional group of interpreters.
4.1.1 **Background**

The study tests the scales instrument used to assess the interpretations in the PhD project. Focusing mainly on methodology, the study provides an overview of Carroll’s scales, the notion of quality in interpreting and the rationale for using Likert-type, holistic scales to rate interpreting. Since the assessment of interpreters’ end product was central to the whole project, Carroll’s scales were chosen to be adapted and evaluated further since they appeared to be valid and easy to use.

The purpose of the study was first to investigate whether holistic scales give valid results when assessing simultaneous conference interpreting products, and second whether there was any difference between laypeople and professional interpreters when rating with holistic scales.

Carroll’s research areas were assessment tools for language testing (Stansfield & Reed 2004) and machine translation. He developed two scales for evaluating machine-translated texts (Carroll 1966). Carroll favoured an integrative testing design. He established the need for two evaluative scales based on the two constructs of *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 (Carroll 1966: 57).

Gerver (1971) and Anderson (1979) used Carroll’s scales to assess interpreting. Gerver did not provide any critical analysis of the application of the scales, but Anderson questioned whether the scales as instruments were sufficiently fine-tuned for measuring the output of interpreting.

Carroll’s scales can account for central aspects of the interpreted event, but not for its entirety as a communicative event.
4.1.2 Data and method

In the study, the scales were adapted to interpreting before being tested. They were adapted to spoken language and Swedish, and highly similar steps were merged so as to reduce the number of steps from nine to six. The adapted scales and are found in tables 1–4 in article 1.

The nine interprettings of the cross-sectional material were used (see section 2.3.1 above) as eliciting material and transformed into assessment files. The three groups were labelled “long-experience (LE) interpreters”, “short-experience (SE) interpreters” and “no-experience (NE) interpreters”, respectively. Their interprettings were carefully transcribed and then transformed into a written text, adding punctuation according to intonation. The text version of each rendition was then divided into 18 interpreting units according to Lederer’s (1978: 330) units of meaning or translation units (Gile 2009: 101). Two criteria were taken into consideration: intonation and idea. The assessment files were built up from the divided interprettings. Each rating file comprised excerpts from all interpreters, randomly mixed.

The raters were six university students (non-interpreter raters), and six professional interpreters (interpreter raters) who had both trained and evaluated interpreters. The raters were instructed at the beginning of each rating session. After rating, the protocols were checked for significant difference and inter-rater reliability.

4.1.3 Major findings

The inter-rater reliability was acceptable for all groups and raters, although slightly higher for interpreters. For intelligibility, as rated by both non-interpreter and interpreter raters, a t-test showed significant differences for all groups except between the SE and LE 25 groups. Informativeness, as rated by both non-interpreter and interpreter raters, showed significant
difference between all groups. Rating scores and $p$-values can be found in tables 6–11 of article 1.

### 4.1.4 Discussion

The two questions posed in the study (i.e. whether holistic scales give valid results when assessing simultaneous conference interpreting products, and whether there was any difference between laypeople and professional interpreters when rating with holistic scales) were both answered in the affirmative. Since three groups of interpreting subjects with clearly different profiles were tested, it could be expected that both their intelligibility and informativeness ratings would significantly differ. Furthermore, in all but one case both rater groups found significant differences between the interpretations; the case where there was no significant difference was intelligibility (defined as understandable, spoken Swedish) between the SE and LE 25 interpreters’ product. This near-equal intelligibility is perhaps not surprising: since the SE interpreters had graduated from an interpreting programme, passed a freelance test for the EU institutions and worked for two years, they had thereby most likely gotten experience in delivering understandable, spoken Swedish.

There are some limitations to this study of the scales’ applicability, though. The sample was small, both in terms of the raters and the interprettings rated, and it was limited to English-Swedish, which may in a wider perspective be a potential limitation (though not for this dissertation project). Despite these limitations, the study supported the choice of holistic scales as a grading instrument for the rest of the PhD project, something that also opens up for testing them in a broader application. The first step of such a broader application could be to test them in a live interpreting context, for example an entrance test or an exam.
4.2 Article 2: “Process and product in simultaneous interpreting: What they tell us about experience and expertise” (data set A)

Article 2 was co-authored with Gard B. Jenset. I was responsible for collecting the data and choosing the instrument for analysis, and was the main writer of sections 1, 2, 3.1 (except second part of 3.1.4), 4.1 and 5. Jenset was responsible for choosing the statistical analysis instrument and for the statistical analysis, and he was the main writer of sections 3.1.4 (second part), 3.2 and 4.2. Both authors participated in the development and editing of the whole article.

4.2.1 Background

The results from the analysis of the cross-sectional material are presented in this article, with the expertise approach as the theoretical background. The study of the cross-sectional material was divided into two parts, one studying the process through retrospection and the other studying the product through assessment. The overall aim of the article was to answer the following question: Can performance differences be established between three groups of interpreters with different levels of experience (NE, SE and LE 25)? The first part investigated whether the three groups differed in their interpreting process, as manifested through reported processing problems, instances of monitoring and strategies (see appendix 1 of article 2). The second part examined whether the three groups differed in their interpreting product, as assessed by raters using Carroll scales; this part is an enhanced analysis of the material in article 1. Ericsson and Smith (1991: 15) relate quality to the investigation of expertise by stating that “although judges can reliably assess the superior quality of the product, it is difficult to analyse such products in order to identify the measurable aspects capturing the superior quality of the product.” It cannot be stressed enough the importance of combining the assessment of quality with the
investigation of the cognitive processes underlying the production of that quality.

4.2.2 Data and method

Both parts of article 2 analysed the cross-sectional material in data set A (described above in section 3.2.1). The first part is based on retrospective protocols coded according to Ivanova (1999), and the occurrences of the different categories were further analysed by using correspondence analysis (CA). Correspondence analysis is a type of multivariate statistical analysis where one variable (experience in this case) has a more or less explanatory value, and different responses can be studied from this variable (in this case how much e.g. omission is related to experience).

The interpreters in this experiment interpreted an EU speech, and immediately afterwards they performed a retrospection from a transcription of the source speech as cue. The interpreting and the retrospection were then transcribed for the analysis.

In the second part of article 2, the ratings made in article 1 were analysed further in order to support the results more solidly, as article 1 had focused more on methodology than actual results. This time the ratings were run through two Friedman rank sum tests as well as a Nemenyi-Damico-Wolfe-Dunn post-hoc test for pairwise comparisons.² A final test was performed to establish whether ratings differences between the two rater groups (non-interpreters and interpreters) could be a distorting factor.

4.2.3 Major findings

The analysis in part one showed, first, that the main difference in processing problems lay between NE interpreters on the one hand and SE and LE 25 interpreters on the other. When the different target language

² The R code for this test was provided by T. Galili from http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code/.
processing problems were studied more in detail, it was clear that the NE group struggled more with comprehension and simultaneity issues, while the SE and LE 25 groups were typified by problems such as perception or finding an equivalent. The LE 25 group was positively associated with difficulties stemming from source language input rate and syntactic processing, while problems finding linguistic equivalents were positively associated with the SE group. The processing problems encountered by the subjects in this study were compared with Ivanova’s (1999) two groups using Spearman’s rank test, and no significant difference was found between the groups in this experiment and Ivanova’s.

Second, in the case of monitoring, a difference was found between the NE and SE groups on the one hand and the LE 25 group on the other. The LE 25 group was positively associated with controlling the accuracy of the translation before utterance. There is a difference between the NE and SE groups in time management issues and internal comments on the speaker, both of which are positively associated with SE interpreters. A qualitative analysis of the monitoring instances of translation showed that despite the similarities in raw data for LE 25 and NE interpreters, the LE 25 interpreters reflect on better ways to interpret a certain utterance, whereas the NE interpreters try to find a general coherence in the output. When compared with Ivanova’s groups (1999), the raw figures suggested differences between the experienced interpreters, but the Spearman rank test did not confirm those differences. The novice groups showed no significant difference in the Spearman rank test.

Third, in the case of strategies, the greatest difference lay between the NE and LE 25 groups. Considering the task was simultaneous interpreting, deletion was unsurprisingly the most common strategy for all groups. Overgeneralization was strongly associated with LE 25 interpreters, while creative interpreting was strongly associated with NE interpreters. The strategy figures could not be compared with Ivanova’s groups, as the analysis in the two studies differed here: Ivanova counted strategies related
to processing problems only, whereas this study counted all strategies. A separate count in order to compare the two studies could of course have been done, but since participants in this study often did not connect their reported strategy to a particular processing problem, such a comparison seemed deficient.

In the second part of article 2, the statistical analysis showed that the processes differed significantly for all three levels of interpreting experience. Moreover, the second round of statistical testing confirmed the results from article 1, there were significant differences for all the groups except for intelligibility between SE and LE 25 interpreters. Finally, it was also found that there were no significant differences between the two groups of raters.

4.2.4 Discussion

The data in article 2 are too small to draw any major conclusions on the differences between the NE, SE and LE 25 groups. Significantly, however, the results support Ivanova’s (1999) findings where the groups and data collection are comparable. The results also support other researchers’ results: For instance, Vik-Tuovinen’s (2006) conclusion that beginners focus on source text and linguistic expression is supported by the NE subjects, who struggle with problems of lexical access (a typical source text and linguistic problem). For the LE 25 group, the monitoring category translation was prevalent, a fact that may support Vik-Tuovinen’s (2006) findings that experienced interpreters focus more on situational factors than less experienced interpreters. The results also agree with Liu’s (2001) result that experienced interpreters monitor output better. The findings of this study show that LE 25 interpreters make more use of monitoring strategies, in particular to check the appropriateness of the utterance and reflect on the speech or the speaker. Certain processing problems, instances of monitoring and strategies may indeed tell us something about expertise, and they may indicate which components of the performance
should be studied in detail to find the superiority Ericsson and Smith (1991) encourage us to look for. The present study suggests that these components might be the ability to monitor and the skill to deliver interpreting with very little information loss compared with the original.

It was significant for this dissertation that the three groups of interpreters showed measurable differences, and that two reliable instruments and a consistent control group were available for investigating the long-term subjects in articles 3 and 4.

4.3 Article 3 – “The development of expertise – or not: Three simultaneous interpreters’ development over time” (data set B)

4.3.1 Background

Article 3 reports the result of the investigation of the quantitative process and product data from the three long-term interpreters. The aim of this study is to investigate whether and how interpreting performance improves over time. The starting hypothesis is that there will be improvements over time in the interpreters’ performance, and that their performance the second time around will correspond to that of experienced peers.

As discussed above in section 3, it is important to study both process and product when studying an interpreter’s development. The results in article 2 showed that there were many differences in terms of both process and product between the NE interpreters and SE and LE 25 interpreters. Following Englund Dimitrova (2010), the process is defined as the cognitive activity of producing a target speech in one language from a source speech in another language. The product is defined as the target speech.
The study compared long-term interpretatings from both a process perspective, using retrospection, and a product perspective, using holistic scales.

4.3.2 Material and method

The participants in this study are the three interpreters featured in the long-term data (data set B), and are described in depth above in section 3.2.2. They are variously labelled the IT group (when recorded at the interpreting programme) and the LE 15 group (when recorded fifteen years later). In addition, the cross-sectional participants, described in section 3.2.1, were used as a control group.

The LE 15 interpreters interpreted two speeches: the EU speech (see section 3.3.1) and the NATO speech (see section 3.3.2). The NATO speech was interpreted on two occasions, the first time during the interpreting programme and the second time fifteen years later. After this latter interpreting, the subjects carried out retrospections of their interpreting. The retrospection was cued with a transcript of the original speech, with normalized orthography and punctuation. Unfortunately, one of the retrospection files was lost due to a technical mishap. The five remaining retrospections were analysed by the author together with a research colleague and coded for processing problems, instances of monitoring and strategy use. The interpretatings of the EU speech were divided into smaller units and randomly mixed into six rating files. The interpretatings of the NATO speech by the LE 15 group were mixed into smaller units together with the interpretatings of the NATO speech by the IT group, and assembled randomly into six rating files with examples from all three interpreters both as an IT (student) and as a LE 15 (professional and experienced). The NATO files were assessed by both interpreter raters and non-interpreter raters, whereas the EU files were only assessed by non-interpreter raters. The raters used the holistic scales tested and adapted in article 1.
4.3.3 Major findings

The main reason for the NATO speech being assessed by both interpreter raters and non-interpreter raters was that the inter-rater variability was very high for the interpreter raters, so it was decided that the assessment had to be redone. Surprisingly enough, the non-interpreter raters had a very low inter-rater variability, but the means of the ratings were more or less the same for the two groups.

The surprising finding of this study was that there were only small differences in the ratings between the interpretings produced by the IT and LE 15 groups. In some cases the scores were even worse for the LE 15 interpretings than for the IT interpretings. The results did not therefore support the assumption that these interpreters had continued to develop interpreting experience that would make them perform better.

When it comes to the EU speech and the comparison with the cross-sectional data, the scores for the LE 15 interpretings were comparable with the SE interpretings and hence worse than the LE 25 interpretings. Furthermore, the scores of the LE 15 NATO and EU speech interpretings are remarkably similar; as the long-term subjects thus received similar scores for their three interpretings (IT NATO, LE 15 NATO and LE 15 EU), they did not evince any long-term development.

In terms of the analysis of the process, it is hard to draw any firm conclusions, as one of the retrospections was lost. But the little processing data that remained confirmed the findings of article 2 and Ivanova (1999), namely that experienced interpreters encounter fewer processing problems and have more strategies at hand to solve the ones they do encounter. No correspondence analysis was done in this analysis, as data from only two interpreters seemed too meagre.
4.3.4 Discussion

The results of the study are challenging. Do they suggest that the three experienced interpreters are not experts according to Ericsson and Smith’s (1991) characteristics? Do they indicate that the performance level reached at the end of the interpreting programme was some type of final stage? Or were perhaps the instruments used to measure performance too blunt?

The quantitative results from the cross-sectional material (discussed above in section 4.2.4) go against the assumption that the performance did not improve from the end of the interpreting programme. In fact, the results in article 2 strongly suggest improvement, at least inter-individually. The LE 15 interpreters have many superficial signs of expertise that support the assumption that they are experts, as explained above in section 3.2.2 (they work at EU institutions, are members of AIIC, are labelled “good interpreters” by their colleagues, and have long experience). The results from the study that tested the scales (4.1.1) and the cross-sectional study (4.2.2) also undermine the idea that the instruments are too blunt, as the instruments showed a clear difference between the groups in the former studies. Another fact that could be in play here is that the same individuals were tested on different occasions; perhaps intra- and inter-individual variation differ. Another factor that may change the results is interpreting style. In her study of translation expertise, Azbel-Schmidt (2005) found that style seemed to be established early on in an interpreters’s career. An interpreting style that favours message compression and thereby word deletion is not necessarily a low-quality interpreting style, but may result in lower scores in an assessment where differences on the word level may affect the overall judgment. It is thus possible that the interpreters developed their style early on, and that this style had a negative impact on the ratings.

Alternatively, the results may be due to a flaw in the design. The EU speeches were only interpreted by the experienced interpreters, perhaps
prompting the raters to assess them more strictly. The raters presumably contrast the different assessment units against each other, and if an assessment file conversely contains interpretations from both inexperienced and experienced subjects, then the latter will “look better”.

Whatever the reasons for the results of this study, they are interesting and unexpected and put the whole project in a different light.

4.4 Article 4 – “Expertise without deliberate practice? The case of simultaneous interpreters” (data set B)

The study reported in article 4 had a different methodological approach than articles 1–3, and carried out in-depth interviews with the LE 15 interpreters of data set B.

4.4.1 Background

Article 4 reports on the qualitative study of the PhD project. The method of in-depth interviews was chosen to cover aspects of deliberate practice, as has been described above in section 2.3.2. To the best of my knowledge, this is the only in-depth interview study on deliberate practice in interpreting, and it may well be the only study so far on interpreters’ engagement in deliberate practice in interpreting practice.

A subject’s engagement in deliberate practice could conceivably be observed in an experiment, but then only for short tasks and not for the extended periods of time that typically characterize deliberate practice. Moreover, the individual’s ability to set clear goals and be open to feedback are important features of expertise. The performer must be able to specify intentions, results or outcomes. Research in goal-setting has shown that practitioners perform better when they specify detailed goals or break a goal down into different sub-objectives (Zimmerman 2007).
A study of deliberate practice will need to discuss the interpreting skill and its different sub-skills, such as skills in linguistics, concentration, analysis and listening, speaking and reading. There is no exhaustive list of skills needed for interpreting, and none that is both empirically tested and generally agreed upon (see e.g. Jones 1998, New Jersey Courts 2007 and Corsellis 2008). Research by McNamara et al. (2011) and Napier and Bontempo (2011) in sign-language interpreting have singled out conscientiousness, emotional stability, self-esteem and openness to experience as factors that predict interpreting success. The openness to experience would be something like intellectual curiosity and thus not far from one of the components of deliberate practice. The aim of the study in article 4 was to explore whether the three LE 15 participants engaged in deliberate practice, set clear goals and received feedback, and whether and how they applied such experiences in their professional life.

4.4.2 Data and method

The participants were the three LE 15 interpreters in the long-term material, described above in section 3.2.2. It should be stressed that these participants were more than happy to participate in new recordings and in in-depth interviews; indeed, their very willingness may reveal something about their (unconscious) view of deliberate practice and learning through introspection.

As the study required a different methodology than the other studies, some time had to be spent on developing this method. In-depth interviews as a tool are described both in section 3.1.4 above and in section 2.2 of article 4. A mind map was also developed to serve as the basis of the interviews. In the above mentioned pilot study on sociological factors in expertise (Tiselius 2010), the participants clearly did not understand the concept of deliberate practice and considered it unimportant or unclear; however, the participants did in fact mention examples that the research leader classified as deliberate practice, goal-setting or openness to feedback. A list of
topics, presented in section 2.2.1 and figure 1 of article 2, was therefore developed that could be used as triggers for the core topics. The interviews were then conducted around these triggers. The interviews, lasting from an hour to ninety minutes, were carried out at the participants’ workplace immediately after the interpreting task reported in article 3. The interviews were recorded, transcribed and then analysed, with ATLAS.ti software used to examine core concepts. The protocols were coded following how the interpreters described learning and practising interpreting skills and sub-skills. The interviews were also reread together with a research colleague from another domain with experience in narrative analysis in order to look for topics that perhaps had been overlooked in the initial analysis.

4.4.3 Major findings

The three interpreters who participated in this study were all focused language learners, although two of them did not focus on language learning until a relatively late stage (late teens, early twenties). None of them was born or grew up bilingually, but they were all highly dedicated once they started focusing on language learning. All three also talked about how they constantly broadened their general knowledge by listening to the news and reading newspapers, books and so forth, in addition to the usual meeting preparations. Moreover, they all stressed the importance of teamwork and of listening to one another, both to help out and to learn. Another issue that stands out is their general ability to focus, with the interpreters talking about their skill to concentrate and to be present in the situation. When it comes to interpreting skill, if by that we mean the ability to transfer a message from one language to the other, none of these interpreters deliberately practised that particular skill. They also considered the interpreting skill to be more or less innate.

When it comes to the notion of deliberate practice, there are a multitude of examples in the interviews of the three interpreters practising several sub-
skills (though not the main skill). However, none of the examples seems to be “deliberate” in the sense defined by Ericsson et al. (1993), that is, that time is set aside with defined exercises and clear goals to refine the main skill. But the three interpreters talk about how they endeavour to improve their interpreting and how they absolutely do not want to work on autopilot, which could suggest that they are intuitively counteracting stagnation in an automatized mode (Ericsson 2007: 685). They also feel a sense of elation when performing well, which could be seen as a type of monitoring. Mood (cf. article 2) is a type of monitoring, where the interpreter reacts in positive or negative terms to his or her own interpreting. Many instances of mood in article 2 pertained to how satisfied the participant was with a certain solution or interpreting. In view of the results of the in-depth interview, there seems to be a connection between monitoring and the type of self-evaluation the participants talk about. Another point is that the goals they spoke of were task goals (i.e. goals for what they want to achieve when performing) and not training goals (i.e. goals for improving a certain skill). The feedback these interpreters talk about is not direct feedback from colleagues, but rather from evaluating themselves.

4.4.4 Discussion

The findings are interesting from several perspectives. Ericsson et al. (2007: 685) note the importance of social networks and support for deliberate practice and the development of expertise. The three LE 15 interpreters’ responses suggest a dearth of support and encouragement in their environment, where there are no competitions, rankings, coaches or performance-based salary increases. In order to develop, it is probably necessary for interpreters in such an environment (one that is typical for interpreters) to use their own strategies for improving their interpreting skill as they themselves understand this skill. The question then is whether such strategies – that is, such practising of sub-skills – can qualify as
deliberate practice. Or perhaps if the time for deliberate practice of the main skill, in the case of highly experienced simultaneous interpreters is merged with the time for work. If none of these explanations is valid, is it possible to be an expert without engaging in deliberate practice, or is it the very concept of “no expertise without deliberate practice” that is inapplicable to interpreting studies? The performers’ deliberate practice is a basic tenet of the expertise theory (Ericsson 2007b), which stipulates that practice should be deliberate and isolated from work (Ericsson, Krampe & Tesch-Römer 1993). One reason why the three participants of this study do not seem to engage in deliberate practice may quite simply be that they are not experts. Yet these interpreters seem very much engaged in practice and strive to improve themselves. But their practising is not done in isolation from work, and much of it seems more intuitive than deliberate. Assuming that the three interpreters are indeed experts, this indicates that expertise in interpreting is possible without deliberate practice.

For the PhD project as a whole, the findings in this study provide an interesting perspective on the three long-term interpreters. Articles 1–3 studied expertise from a quantitative perspective and did not take deliberate practice into account, but focused solely on performance and process during the task. The in-depth interview study provides a fuller picture of possible interpreting expertise and whether highly experienced interpreters engage in deliberate practice exercises.

On a pedagogical note, an interesting implication may be to introduce the notion of practice and skill development during an interpreter’s whole career, as well as the type of continued education that Bontempo and Napier also suggest for sign-language interpreting (2007: 295–296).
4.5 Summary of the results from all four studies

4.5.1 Methodological results

Holistic, Likert-type scales based on Carroll’s scales (1966) provided reliable and valid results when tested on interpreters with different levels of experience. The results of the participants’ retrospective recollection of the task and the researcher’s subsequent protocol analysis were also corroborated by Ivanova’s (1999) results. Finally, a mind map was developed as an interview guide for in-depth interviews on deliberate practice.

4.5.2 Research results

The results on the interpreting process showed that experienced interpreters (LE 15 and LE 25) encounter fewer processing problems and have more strategies at hand when they encounter problems than interpreters with short (SE) or no (NE) professional experience. This was true both for interpreters in the cross-sectional material (data set A) and in the long-term material (data set B). Furthermore, experience is decisive when it comes to monitoring: in the cross-sectional material (data set A), the most experienced interpreters (LE 25) were more associated with monitoring than the other two groups (SE and NE).

The results regarding the assessment of the interpretings are clear when it comes to the cross-sectional material (data set A). There is a statistically significant difference concerning product between the NE group and the SE and LE 25 groups. There is also a statistically significant difference in the transferred information (i.e. the content) between SE and LE 25 interpreters. However, the results from the long-term material (data set B) showed no difference between the product of the IT and LE 15 groups (that is, the same interpreters recorded fifteen years apart).
In-depth interviews with the three LE 15 interpreters showed that they were extremely goal-focused from early on in life and engaged in many practice-like activities (including for several of their sub-skills). They did not however give any indication of engaging in deliberate practice as defined by Ericsson et al. 1993. They did talk about constantly striving to produce better interpretations and also about the positive physical experience of performing well.
5. Discussion

This section discusses both the methodology and results of the dissertation.

5.1 Methodological discussion

First, it should be said that the combination of quantitative and qualitative methods has been fruitful and makes to a certain extent up for the low number of participants. The surprising findings in the various studies may be due to the research instruments used, but as discussed in section 4.3.4, there is also strong support for the functionality of the methodological design and thereby the validity of the results.

One of the factors that may skew the results is the scales chosen to measure the quality of the interpreting product. However, they were tested and retested and ought to be considered reliable, and they provided the expected outcome when they were used in the cross-sectional material that supports their validity.

Another factor that may affect the results is the elicitation tool. It is a fact that experts do not excel in routine tasks (see section 2.3), and there is a possibility that the interpreting task in the study was seen as a (too simple) routine task by the LE 15 participants; however, this was not observed in the LE 25 participants. Moreover, the elicitation tool (the speech) was to a certain extent adapted. Fairly short speeches were used for experimental reasons, and one of the criteria for choosing the speech was its generality, so as not to make it impossible for the laypersons participating in the experiment. The first speech was nevertheless tweaked in order to add difficulties such as figures, names and difficult reasoning. For the long-term group (LE 15), the second speech was the same one they had interpreted at the interpreting programme for comparative reasons; this speech presented difficulties to the LE 15 interpreters since it was dated
and dealt with matters outside their area of expertise. Except for the IT group in the long-term material, the speeches were not interpreted in a routine setting. LE 15 and LE 25 interpreters interpreted from an interpreting booth, but they did not have any ordinary listeners and the elicitation speech was recorded; they were also surrounded by recording equipment. The SE and NE groups performed their interpreting either at university or at the researcher’s workplace. The setting for the LE 15 and LE 25 interpreters does not indicate a routine task, which in this case would be in a booth with a live speaker, live audience and at least one colleague. In addition to the unusual conditions, the research leader was also present next to the participants listening to their performance. The experimental setting of the data collection event did not affect the performance of the LE 25 group, however, and the SE group performed at the same level as the LE 15 group. The elicitation situation was moreover a stressful event where experts could potentially excel because of their access to expert knowledge.

Furthermore, the result may be skewed since the interpreters’ voice was not part of the assessment in the ratings, which were done from transcripts with normalized spelling and syntax (based on intonation). Collados Aís et al. (2011) point out that voice quality is so important that it may actually overshadow other key factors. The aim of the rating in this study was to assess the interpreter’s ability to reproduce in the target language an understandable message that contained the information in the source language message, and it was assumed that the possibility of a rater identifying the rated subject through the voice would affect the evaluation. However, as important as voice may seem when grading interpreting, an experiment that compared grading from sound files and grading from transcripts showed that there was no significant difference between the two (Tiselius 2010). Finally, the same conditions were true for the cross-sectional material, and in that material the differences between the groups were clear.
Based on the arguments above I assume that the instruments were valid and reliable and that the data were representative. The following section (5.2) will discuss the results.

5.2 Discussion of the results

The quantitative results of the long-term material were the most surprising ones, as they went against the assumption that experience would enhance the interpreting performance and yield high assessment scores. As mentioned above, there are many superficial indications that the three LE15 interpreters are highly skilled. However, the assessment results indicate they are not experts as defined by Ericsson and Smith (1991). It cannot be excluded that their expert knowledge is found in other areas than the one tested in these studies. In the in-depth interviews they talk about adding languages, improving their general knowledge and working on delivery, so perhaps that is where their expert performance can be found.

Hervais-Adelman et al. (2011) showed that changes take place in an interpreter’s brain during training; given the results of the long-term study in this dissertation, we may wonder whether such neurological and cognitive changes take place during training and then remain fairly stable. As mentioned above in section 2.1, Hill and Schneider (2007: 675) say that as processing becomes automatized, the influence of the general control network (necessary while learning a task) either decreases or disappears entirely. If we assume that the automatization is completed during the training programme, and if automatized processes vs. non-automatized processes is what influence the results of the grading, then the difference between the subjects may be too small on the intra-individual level to be measurable, as the processes in this scenario were acquired and automatized during the interpreting programme and then perhaps unchanged over the years.
The experienced interpreters in the cross-sectional material (LE 25), who had more than twenty-five years of experience, received considerably higher scores than the experienced interpreters in the long-term material (LE 15), who had fifteen years of experience at their second recording. This may be due to experiment design, as discussed above, but it may also indicate that it takes longer than the supposed ten years, or even fifteen years, to gain expert knowledge in interpreting. In-depth interviews with the LE 25 group could have shed light on those issues, but practical constraints entailed that such interviews were unfortunately only carried out with the LE 15 group.

Finally, as is also discussed in article 4, there are few incentives for professional interpreters who have reached the highest level of their field to continue practising their skills. Having been accredited to international institutions, they are subject to constant quality monitoring so as not to perform below a certain minimum, but there are no mechanisms aimed at improving the main skill. Professional development consists of improving general knowledge or language skills (clearly important), but there are no rankings or pay raises for the best interpreters. The incentive for improvement lies instead in personal well-being and in the satisfaction of a job well done (as assessed by themselves). This is not necessarily a bad incentive, but the importance of the environment should not be underestimated.

5.3 The expertise theory and simultaneous interpreting

The findings are not conclusive concerning expertise in interpreting. In the cross-sectional material the results were clear and conclusive. There is a clear dividing line between the NE group on the one hand and the SE and LE 25 groups on the other, and there are also measurable differences both for process and product between the SE interpreters and the LE 25
interpreters. With these results we can assume that training and experience clearly matter in interpreting, and that extensive experience makes a difference.

The results from the cross-sectional material are however not confirmed by the interpreters in the long-term material (the LE 15 group), who did not interpret a given speech measurably better despite fifteen years of active experience. As years in the profession is a weak factor of expertise, the other parts of the superficial side of expertise was also taken into account, and all three of them had credentials that would put them in the expert category. But the LE 15 group outperformed neither the early recordings of themselves nor their experienced colleagues (the LE 25 group) in the cross-sectional material. Their performance was stable when improvement was expected. Either it must be assumed that these interpreters were not experts in absolute terms as defined by Ericsson and Smith (1991), or further investigations are needed to obtain more information.

The in-depth interview also shows that it is still an open question whether interpreters engage in deliberate practice as defined by Ericsson et al. 1993. Although the interpreters described examples of what would be labelled practice, none actually said that they practise and two of them even explicitly said that they never practise. This is supported by Vik-Tuovinen’s (2006: 308) finding that experienced interpreters are less occupied with preparation than their less experienced counterparts. None of the three interpreters in the present study talks about or provides any examples of working to improve their main skill. So although they practise, it is hard to label it deliberate practice in the sense assumed in the expertise theory. On top of that, they all more or less think there is an X factor or an innate talent in interpreting. If interpreters thus believe that their main skill is innate, there may be less reason to continue practising this skill. As seen above, students are taught to practise and assess their
interpreting skill, and individuals who prepare for an accreditation test for interpreters presumably also practise this skill. Once a student has graduated or a novice interpreter has passed the accreditation test, however, there are few if any courses or initiatives to refine the interpreting skill. Although interpreters learn to improve their language skill, learn new languages, enhance their general knowledge and possibly also practise their consecutive interpreting skill as they prepare to add that new language to their combination, there seem to be few incentives to improve the interpreting skill itself once the student has graduated or passed an accreditation test. There are no particular merits for improving the interpreting skill, such as higher salaries, prestigious prizes nor an improved ranking. Freelance interpreters can presumably get more assignments and thereby more money if their interpreting skill improves. But many other factors are involved when freelancers are assigned jobs, such as language combination, availability, geographical proximity and not least personal connections. Hunt (2007: 35) points out that since expertise requires both motivation and support, society greatly influences where expertise is produced: in areas where remunerations are high and excellence in a field is remunerated even higher, experts are likely to prosper. In the interpreting world, conference interpreters are paid the highest while staff conference interpreters at various institutions also receive a comparatively high monthly salary. According to Hunt’s theory, interpreting experts could thus be expected to be found as staff interpreters at international institutions.

For the expertise theory, the subjects’ engagement in deliberate practice is an absolute condition. Interpreting studies have only started to discover what characterizes an expert performance, and the findings here suggest that research must be carried out on deliberate practice in interpreting as well. These findings indicate either that interpreters practise their skills in a naive manner (that is, without a conscious understanding of the deliberate dimension, yet with features of deliberate practice), or that the
three interpreters do not engage in deliberate practice, something that might also apply to conference interpreters in general. More studies must examine deliberate practice in interpreting before a definitive conclusion can be drawn, but there may be a need to redefine either the concept of deliberate practice in interpreting in particular or the criteria of the expertise theory in general.

Moreover, it is clear from these results that the expertise label needs to be used cautiously. Liu’s overview of expertise in interpreting (2009) shows that much is known about interpreters with a certain amount of experience and what they do or not. But not much is known about what experts do, since the definition of an expert interpreter remains unclear. This is not to say that expertise in interpreting would have to take the Weiss and Shanteau (2003) definition that no measurable outcome exists concerning the expertise of interpreters, but the identification of expertise in interpreting, following the expertise theory (Ericsson, Charness & Hoffman 2007), requires more methodologically minute studies, before we can sketch the traits necessary for expertise in interpreting. In fact, it may be difficult to assess an outstanding performance because the group of highly experienced interpreters is small and homogeneous, and presumably they could be all excellent or all mediocre. These and other question marks must be addressed if interpreting studies is to adhere to the strictest definition of the expertise approach as described above.
6. Conclusion

The aim of this dissertation was twofold, as it contained both a methodological and a research side. In regard to methodology, it aimed to test holistic scales for assessing interpreting and develop an interview guide for in-depth interviews on deliberate practice. In regard to research, it aimed to establish a measurable difference in the interpreting skill (concerning both process and product) among interpreters with different levels of experience, and to explore what this difference consisted of. The data consisted of a cross-sectional material (n=9) and a long-term material (n=3). The following text repeats the questions in section 2.5 and provides answers to them:

1) Is there a measurable difference in the interpreting skill from the student level to the highly experienced level?
   a. The assumption that there would be a measurable difference in the interpreting skill among performers with little or no experience and performers with long experience was supported for one of the data sets, the cross-sectional data (A).

2) Is there a measurable difference in the interpreting skill both when it is measured cross-sectionally (i.e. inter-individually) and long-term (i.e. intra-individually)?
   a. The assumption that there would be a measurable difference in the interpreting skill regardless of data was not supported, as there was no measurable difference between the IT interpreters and the LE 15 interpreters in the long-term data (B).

3) If there is a measurable difference, what does this difference consist of?
   a. The assumption that there would be a difference in rating between participants with little experience and participants with long experience was supported by the cross-sectional material (A), but not for the long-term material (B).
   b. The assumption that there would be a difference in the interpreting process between participants with little experience and participants with long experience was
supported by the cross-sectional material (A) and could not be tested for the long-term material (B).

4) How do experienced interpreters perceive different factors in their long-term competence development?
   a. The assumption that experienced interpreters would claim that they practise a lot was not supported. However, the participants they talked frequently about other practice-like activities.
   b. The assumption that experienced interpreters would claim that they constantly strive to improve themselves was supported.
   c. The assumption that experienced interpreters would be able to talk about their goals, on both the micro and macro levels, was not directly supported (although they often talked about how they had achieved different goals and generally seemed goal-oriented).
   d. The assumption that experienced interpreters would claim that they made use of their colleagues for feedback and help was partly supported (interpreters help their colleagues in the booth and also listen for inspiration, but coaching outside the booth was never mentioned).
   e. The assumption that experienced interpreters would be able to describe how they solve issues under pressure was supported (all participants talked about their ability to focus and perform under difficult conditions).

For the methodological development part, the following questions were answered:

1) Can holistic scales for measuring intelligibility and informativeness be developed into a valid and reliable measuring instrument for quality in interpreting?
   a. Yes, the assumption that the scales could be developed and tested so as to form a valid and reliable measuring instrument for quality in interpreting was supported.

2) Will holistic scales work equally well as a measuring instrument whether used by laypersons or experienced raters?
   a. Yes, the assumption that the scales could be used by different raters and still generate valid results was supported.
3) How should an in depth-interview be carried out in order to yield results on the concept of deliberate practice?
   a. An interview guide was created and participants were prompted to discuss issues of deliberate practice through different trigger topics.

Interpreting performance was compared, in regard to both process and product, in many combinations between subjects without interpreting experience and subjects with different levels of interpreting experience, both cross-sectionally and long-term. As reported above, there was a measurable difference between the groups in the cross-sectional material, but not in the long-term material. Another conclusion, supported by other studies, is that experienced interpreters, when interpreting, have more strategies at hand and encounter fewer processing problems than less experienced interpreters or laypersons to interpreting non-interpreters. The results from the project supported the findings in Liu’s literature review (2009), where she notes that experienced interpreters seem to have developed well-practised strategies in the comprehension, translation and production processes that are specific to the needs of simultaneous interpreting. And finally, these experienced interpreters have developed an ability to efficiently manage their attention so that it can be switched between different processes (Liu 2009: 174). Vik-Tuovinen’s (2006) results are also supported by the results in this project. Her finding that more experienced interpreters were conscientious about their delivery was confirmed in the in-depth interview. From the interviews it could also be concluded that the experienced interpreters allocate much time for practice, although they don’t consciously label it as such. Furthermore, they were also highly goal-oriented both in life in general and when interpreting.

From a methodological point of view, the holistic scales that were adapted from Carroll (1966) and used in articles 1–3 are well-tested by now and await further testing by for instance examiners. The retrospective method used for exploring the interpreter’s process also produced valid and
reliable results, although, as shown by Englund Dimitrova and Tiselius (submitted), the use of retrospective protocols needs to be done with great exactitude. Retrospection is nevertheless warmly recommended as a method for investigating interpreting processes.

This is a large study, albeit with few participants. In order to really evaluate and map outstanding performance, further studies and more extensive data per interpreter are needed, with each participant ideally being recorded on several different occasions.

It is encouraging that the quantitative results from the cross-sectional study supported other results from other researchers (Ivanova 1999 and Vik-Tuovinen 2006). Hopefully, those results can in turn be supported by other researchers too.

Further longitudinal or long-term research would also be welcome in the field, as little is known within interpreting studies about intra-individual development beyond interpreting programmes. In order to understand the development of not only expertise in interpreting but also of interpreters after graduation, the field needs to see more longitudinal or long-term studies. Surely there must be many audio cassettes or mp3 files from various interpreter training programmes lying around waiting to be followed up.

Comparison with expertise in other fields that also lack rankings and reward systems would also be welcome, together with discussions on how to identify a practitioner’s main skill. The main skills of nurses, researchers or blacksmiths could presumably also be the object of discussion.

Finally, I repeat my call for more studies on deliberate practice. In order to study all the domains of expertise in interpreting, deliberate practice needs to be part of the tradition. In this thesis, interpreters’ deliberate practice has been studied through in-depth interviews, but it could also be studied
through for instance diary studies or on a micro-level with a quantitative design.
List of references


Isham, W. P. 1994. “Memory for sentence form after simultaneous interpretation: Evidence both for and against deverbalization”. In Bridging the Gap. Empirical Research in Simultaneous Interpretation


Article 1

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 indicate 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 literary 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 (cf. 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 larger 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)\(^3\) (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 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 deals strictly with simultaneous conference interpreting, and with the language combination English (C) into Swedish (A).\(^4\)

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 products 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 are two sub-sets of research questions that contribute to the overall purpose of the study:

---

\(^3\) [http://www.ncsconline.org/D_Research/Consort-interp/fcice_exam/index.htm](http://www.ncsconline.org/D_Research/Consort-interp/fcice_exam/index.htm)

\(^4\) 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](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 concerns 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 the 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 concerns who should do the grading: Professional interpreters usually have some experience of 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 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 the interpreting quality (e.g. Moser 1995 or SCIC customer survey 2008), but the way they grade and assess an interpreting product 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 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 and 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 instruments 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 is needed before any conclusions can 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 possibilities 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 my 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 is 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 to assumed for the purpose 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 group of graders were compared to each other.

3.1 Adaptation of the scales

As already mentioned, Carroll’s scales do not take issues 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 decided that “ordinary spoken Swedish” would encompass fluency, coherence and clarity. Therefore, no additional components were added.

Furthermore, as Cohen et al. (1996: 224) mention, 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 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 interpretation) 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).
Table 1. Scale of intelligibility, adapted version and original (Carroll 1966: 58)

<table>
<thead>
<tr>
<th>Original Scale of Intelligibility</th>
<th>Scale of Intelligibility (as adapted in the present study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.</td>
<td>6. The rendition is perfectly clear and intelligible. Like ordinary spoken Swedish with few if any stylistic infelicities.</td>
</tr>
<tr>
<td>8. Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily “corrected.”</td>
<td>5. Generally clear and intelligible but with minor grammatical or stylistic peculiarities or unusual word choices, though nothing that hampers the understanding.</td>
</tr>
<tr>
<td>7. Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>-</td>
</tr>
<tr>
<td>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.</td>
<td>-</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>-</td>
</tr>
<tr>
<td>2. Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.</td>
<td>2. Almost completely unintelligible, although it does not seem completely nonsensical and the addressee may, with great effort, discern some meaning.</td>
</tr>
<tr>
<td>1. Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.</td>
<td>1. Totally unintelligible and completely without meaning.</td>
</tr>
</tbody>
</table>
Table 2. Scale of informativeness, adapted version and original (Carroll 1966: 58)

<table>
<thead>
<tr>
<th>Original Scale of Informativeness</th>
<th>Scale of Informativeness (as adapted in the present study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.)</td>
<td>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.)</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>7. (Between 6 and 8.)</td>
<td>-</td>
</tr>
<tr>
<td>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.</td>
<td>-</td>
</tr>
<tr>
<td>5. (Between 4 and 6.)</td>
<td>4. Reading the original gives some additional information about syntax and words. It can also clarify minor misunderstandings in the rendition.</td>
</tr>
<tr>
<td>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.</td>
<td>-</td>
</tr>
<tr>
<td>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.</td>
<td>3. By correcting one or two meanings, mainly on the word level, the reading of the original gives only a minor difference in meaning compared to the rendition.</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>1. Not informative at all; no new meaning is added, nor is the reader’s confidence in his understanding increased or enhanced.</td>
<td>1. No new meaning is added by the original, nor is the addressee’s understanding of the rendition increased.</td>
</tr>
<tr>
<td>0. The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.</td>
<td>0. The original contains less information than the rendition.</td>
</tr>
</tbody>
</table>
Table 3. Scale of intelligibility on grading sheet

|--------------------------|----------------------------|----------------------|---------------------------|------------------------|--------------------------|

Table 4. Scale of informativeness on grading sheet

<table>
<thead>
<tr>
<th>0. Original contains less information than rendition.</th>
<th>1. Without any new information, strengthens the intended meaning.</th>
<th>2. No new information, gives some new information.</th>
<th>3. Minor changes in meaning.</th>
<th>4. Gives new information.</th>
<th>5. Original explains and improves.</th>
<th>6. Only new information.</th>
</tr>
</thead>
</table>

However, having adapted the scales as described above, it was estimated that they had a high componential element in them, and that 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 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.5

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 as follows

(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 for 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 one does not generally interpret),\(^6\) and the untrained interpreters studied English at the university level.

\(\text{Table 5. Age and experience of the interpreters}\)

<table>
<thead>
<tr>
<th>Group</th>
<th>Age span</th>
<th>Years at university</th>
<th>Interpreting school</th>
<th>Years of experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>20–30</td>
<td>4</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>No experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ii)</td>
<td>30–40</td>
<td>4</td>
<td>Yes</td>
<td>&lt;2</td>
</tr>
<tr>
<td>Short experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iii)</td>
<td>50–60</td>
<td>4</td>
<td>Yes</td>
<td>&gt;25</td>
</tr>
<tr>
<td>Long experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

\(^6\) http://www.aiic.net/ViewPage.cfm?page_id=199#langclassif.
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 their 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 avoid graders being able to recognize 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 students of translation (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-interpreters), 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 their 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, and 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 would be 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 no-experience interpreters. The interpreter graders’ scores also show significant difference for the renditions of short-experience interpreters versus the renditions produced by no-experience 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 ($n=6$)

<table>
<thead>
<tr>
<th>Renditions</th>
<th>Intelligibility</th>
<th>No-experience 3.79</th>
<th>Short-experience 5.25</th>
<th>Long-experience 5.42</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-experience 3.79</td>
<td>-</td>
<td>0.001**</td>
<td></td>
<td>0.001**</td>
</tr>
<tr>
<td>Short-experience 5.25</td>
<td>0.001**</td>
<td>-</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>Long-experience 5.42</td>
<td>0.001**</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**p<0.01
Table 7. Significance in gradings of intelligibility by interpreters (n=6)

<table>
<thead>
<tr>
<th>Renditions</th>
<th>No-experience 3.16</th>
<th>Intelligibility</th>
<th>Long-experience 5.11</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-experience</td>
<td>-</td>
<td>0.001**</td>
<td>0.001**</td>
</tr>
<tr>
<td>Short-experience</td>
<td>0.001**</td>
<td>-</td>
<td>0.1</td>
</tr>
<tr>
<td>Long-experience</td>
<td>0.001**</td>
<td>0.1</td>
<td>-</td>
</tr>
</tbody>
</table>

*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)

<table>
<thead>
<tr>
<th>Renditions</th>
<th>Non-interpreter graders</th>
<th>Interpreter graders</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-experience</td>
<td>3.79</td>
<td>3.16</td>
<td>0.018*</td>
</tr>
<tr>
<td>Short-experience</td>
<td>5.25</td>
<td>4.88</td>
<td>0.078</td>
</tr>
<tr>
<td>Long-experience</td>
<td>5.4</td>
<td>5.11</td>
<td>0.015*</td>
</tr>
</tbody>
</table>

*p<0.05

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

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

<table>
<thead>
<tr>
<th>Renditions</th>
<th>No-experience</th>
<th>Short-experience</th>
<th>Long-experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.42</td>
<td>3.15</td>
<td>2.31</td>
</tr>
<tr>
<td><strong>p&lt;0.01</strong> Note:</td>
<td>The lower score, the better performance.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Renditions</th>
<th>No-experience</th>
<th>Short-experience</th>
<th>Long-experience</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5.13</td>
<td>3.42</td>
<td>2.60</td>
</tr>
<tr>
<td><strong>p&lt;0.01</strong> Note:</td>
<td>The lower score, the better performance.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Renditions</th>
<th>Non-interpreter graders</th>
<th>Interpreter graders</th>
<th>Significance in gradings between non-interpreters and interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.42</td>
<td>5.13</td>
<td>0.001**</td>
</tr>
<tr>
<td></td>
<td>3.15</td>
<td>3.42</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>2.31</td>
<td>2.60</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>p&lt;0.01</strong> Note:</td>
<td>The lower score, the higher correspondence.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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

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. Quite a few 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 product of simultaneous conference interpreters. Furthermore, it investigated whether it was possible for graders who are not researchers in interpreting studies 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 that this tendency would hold in a larger sample, and it is not possible to speculate on the reason for it.

Second, 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 is likely to 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.

Third, in order to test the grading scales, an alternative set-up of the renditions would have been to manipulate the renditions on the grading sheets in order to ascertain 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 a 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 as well, especially 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 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 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.

6. 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


Lederer, M. 1978. “Simultaneous interpretation: Units of meaning and other


## Appendix 1 Carroll’s scales in Swedish

<table>
<thead>
<tr>
<th>Skala för förståelse (Intelligibility)</th>
<th>Skala för informativitet (Informativeness)</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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.)</td>
</tr>
<tr>
<td>5. I stort tydlig och förståelig men med små grammatiska eller stilistiska svagheter 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.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>3. Verkar vara en förståelig mening men är i själva verket mer oförståeligt 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.</td>
<td>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 the word level, the reading of the original gives only a minor difference in meaning compared to the rendition.</td>
</tr>
<tr>
<td>2. I princip helt oförståeligt. Verkar dock inte helt osammanhängande och lyssnaren kan möjliga 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.</td>
<td>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 förstår 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.</td>
</tr>
<tr>
<td>1. Helt oförståeligt och helt utan mening. Totally unintelligible and completely without meaning.</td>
<td>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.</td>
</tr>
</tbody>
</table>
| 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)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helt oförståeligt</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>I princip oförståeligt</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Verkar förståeligt</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Huvudtanken förståelig</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>I stort förståeligt</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

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 really consider to be a qualitative step forward in order to protect 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 entailed 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)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>orginalet innehåller mindre information än tolkningen</td>
<td>originalet contains less informatio n than rendition.</td>
<td>Utan någon ny information</td>
<td>Without any new information .</td>
<td>Ingen ny information stärker avsedd betydelse</td>
<td>No new information, strengthens the intended meaning.</td>
<td>Lite förändring i betydelsen</td>
<td>Minor changes in meaning.</td>
</tr>
</tbody>
</table>
Article 2

Process and Product in Simultaneous Interpreting
WHAT THEY TELL US ABOUT EXPERIENCE AND EXPERTISE

ELISABET TISELIUS
University of Bergen, Norway/Stockholm University, Sweden

GARD B. JENSET
Bergen University College, Norway

Abstract
The expertise approach (Ericsson 2008) has been used to explore the competence of translators and interpreters since the mid-1990s, and is now a well established sub-field in translation and interpreting process research (Jääskeläinen 2010). In the area of interpreting, Ivanova (1999), Liu (2001) and others have explored the expertise approach. The studies reported in this article follow up on this work, but go one step further and investigate both process and product. The aim of the two studies was to explore the differences in performance between interpreters with shorter and longer experience (possible experts). Participants (n = 9) with no, short or long experience interpreted the same speech and performed retrospection immediately after. The first study, dedicated to process, used Ivanova’s (1999) method for investigating the process. The second study, on product, let two groups, non-interpreters (n = 6) and interpreters (n = 6), rate the interpreting performances using Carroll’s (1966) scales for intelligibility and informativeness. It was found that the degree of experience influences the processing strategies used by interpreters and the types of problems they report. Experience also has an impact on how the product of experienced interpreters and that of less experienced interpreters is rated, both when rated by interpreters and by non-interpreters.

Keywords: expertise, simultaneous interpreting, retrospection, process, monitoring, strategies, rating.

The authors would first like to thank the interpreters that participated in this study. The authors are also indebted to Birgitta Englund Dimitrova, Åse Johnsen, Barbara Blair and Christer Johansson as well as the anonymous reviewers of this article for their insightful comments and corrections. Any faults or weaknesses that remain despite their close scrutiny are of course the responsibility of the authors alone.
1. Introduction

Expertise develops over time, as performers develop experience in their field. Ericsson (2000) outlines the expertise approach as follows:

*The theoretical framework of expert performance makes the fundamental claim that improvement in superior reproducible performance of adult experts doesn’t happen automatically nor magically. Performance improvement can be linked to changes in cognitive mechanisms mediating how the brain and nervous system control performance and in the degree of adaptation of physiological systems of the body. The principal challenge to attaining expert level performance is to induce stable changes that allow the performance to be incrementally improved (2000: 201–202).*

While experience can easily be described in temporal terms, e.g. the number of years in the profession, hours of training or repetitions of the task, it is more delicate to describe or identify expertise. In their work on a theory of expertise, Ericsson and Smith (1991) listed six criteria to be taken into consideration in order to study expert performance: long experience, consistent outstanding performance, access to expert knowledge when needed, deliberate practice, clear goals, feedback and advice. In competitive fields that involve ranking, such as tennis or chess, an examination of these criteria for the top performers will in all probability identify the experts in the field. Their performance can then be examined in order to identify the features of expert performance. However, it is more difficult to identify the possible experts in fields in which ranking is not available as a means to investigate their expertise. One method is to compare the work of experienced practitioners with that of novice practitioners, and map the former’s performance and background. The differences (if any) between the two groups may constitute clues to outstanding practice in that field. On the basis of those differences, it may be possible to describe expertise and expert performance in the field. As Ericsson and Smith (1991: 8) put it, “the expertise approach is an attempt to describe the critical performance under standardized conditions, to analyze it, and to identify the components of the performance that make it superior.” In order to investigate the performance of interpreters employing the expertise approach, it is necessary to consider both process and product: process, because these may alter with experience without necessarily showing in the product; and product, because the products of experienced practitioners most likely differ from those of less experienced practitioners. From these two angles it becomes possible to identify the components of superior performance and consequently describe expertise in the relevant field, in this case simultaneous interpreting.
1.1 Aim and Scope

The overall aim of the two studies presented in this paper is to explore differences in performance between inexperienced and experienced interpreters. This has certainly been done before (see section 2.1 below), but the field is not yet completely mapped; proceeding step by step, it may be possible to identify the components of superior performance.

The overall question that the two studies sought to answer was: Can a difference in performance be established between three groups of interpreters with different levels of experience?

Previous research suggests that there is likely to be a difference in performance that depends on the interpreter’s experience. The crucial question is: can this difference be identified and demonstrated in the context of an experimental design? There were two research questions:

First study: Is there a difference in the interpreting process (manifested through reported processing problems, instances of monitoring and strategies (see appendix 1)) when carried out by groups of interpreters with different experience? Previous studies indicate that there is a difference in the interpreting process employed by experienced interpreters and inexperienced interpreters, see e.g. Ivanova (1999) and Vik-Tuovinen (2006). Years of practice and knowledge of the field are likely to make a difference. However, can this difference be demonstrated in an experiment that taps into the process? Is it possible to identify what the difference consists of?

Second study: Is there a difference in the ratings of the interpreting products produced by groups of interpreters with different levels of experience? Again, on the basis of previous research, such as the ones mentioned above, it is reasonable to hypothesize that there is a measurable difference between the groups investigated. The question remains: can this assumed difference be demonstrated in the context of an experiment?

2. Background

2.1 Expertise Research in Interpreting Studies

An important focus of studies on conference interpreting has been on comparing inexperienced interpreters with experienced interpreters (e.g. Gerver 1971; Anderson 1979; Dillinger 1989). In the late 1990s Moser-Mercer (2000) took the concept of expertise from cognitive psychology into interpreting research, and since then several doctoral theses have been defended on that topic (see for instance Ivanova 1999 or Liu 2001). When exploring expertise, studies have, for instance, analyzed the interpreters’ performances either by studying the process (e.g. Ivanova 1999) or the product (e.g. Liu 2001) quantitatively, or by analyzing the product and the process both qualitatively and quantitatively (Vik-Tuovinen
Ericsson includes three steps in his original expertise approach (Ericsson 
& Smith 1991). The first is to capture the essence of superior performance under standardized laboratory conditions by identifying representative tasks. The second step is to provide a detailed analysis of the performance, and the third is to account for the acquisition of the characteristics and cognitive structures and processes that have been found to mediate the superior performances of experts (1991: 12). These steps may sound straightforward, but represent a major challenge for research into interpreting. Interpreting consists of a multitude of tasks all performed simultaneously (cf. the Moser model 1978). Only by testing interpreters’ performance under relatively standardized conditions is it possible to isolate and analyze the superior performance identified in step two. In Ericsson and Smith’s words, “by careful analysis of the expert performance in real life, we try to identify recurrent activities that can be reproduced under controlled conditions” (1991: 14). The two studies reported here are parts of Ericsson and Smith’s first and second steps. We study the performance of all the interpreters under reasonably standardized conditions. We contrast the different performances, analyze them and explore possible superiority.

Ivanova (1999: 182), whose method is described in greater detail in the next section, saw that experienced interpreters encounter fewer processing problems and have more elaborate problem representations. When it comes to problem solving and strategies, they have a wide range of context-sensitive solutions at hand. Ivanova concluded that an important feature of interpreting expertise is its adaptive nature relying on the cognitive flexibility of the experts. Liu (2001: 93), who compared 13 professional interpreters (with training and at least 2 years of experience) with two student groups (n = 12 and n = 11), found that experienced interpreters had developed domain-specific skills (ability of selective encoding, better monitoring of output, more efficient allocation of working memory), which were in turn crucial in outperforming less experienced subjects. Vik-Tuovinen (2006) investigated seven experienced interpreters with one to fifteen years of experience and compared them to two groups of six and eight students respectively. She found that beginners focus on the source text and the linguistic expression, whereas professional interpreters focus more on social aspects and situational factors (Vik-Tuovinen 2006: 313).

Liu (2008), in a general overview of research on interpreting expertise, concluded that expert interpreters differ from novices in that they “have developed well-practiced strategies in each of the comprehension, translation and production processes.” However, a crucial part of their expertise is not only the development of these strategies, but the skill to manage them during the task of simultaneous interpreting.
2.2 Studying the Process

Ericsson and Smith state that “comparison of think-aloud verbalizations by experts and novices is the best-known method of assessing differences in the mediating processes as functions of the subjects level of expertise” (1991: 20). When investigating the process of interpreting, there are not many methods at hand. A pure think-aloud method while the subject is performing a task is impossible; the interpreter cannot both interpret and verbalize on the process at the same time. There are very limited possibilities to tap into the ongoing process of interpreting, and none if we want the interpreting subjects to verbalize their process. Immediate retrospection, however, is a possibility since the subjects recall and verbalize their memories of the process after the task. The retrospective interview is usually prompted by a cue; for interpreting this can be either the source text or the interpreter’s own production (sometimes supported by the source text), as suggested by Ericsson and Simon (1996).

Both Ivanova (1999) and Vik-Tuovinen (2006) use retrospection in their studies of expertise in interpreting. The first study reported in this chapter closely follows the method used by Ivanova in her PhD thesis (1999). Her subjects (n = 16, including novices as well as possible experts) interpreted an English speech into Bulgarian; immediately after interpreting, the subjects performed a retrospective interview where they used only the source text as cue. Following a pilot study comparing the reliability of two types of cues (source text only and source text with a target text recording), Ivanova decided to use source text only since she found that use of the target text recording ran the risk of influencing subjects, who could deduce the process from the product rather than recalling the process in itself. For an examination of Ivanova’s method and retrospection for both translation and interpreting, see Englund Dimitrova and Tiselius (2009).

2.3 Evaluating the product

Quality and how to assess it has been discussed by researchers such as Moser-Mercer (1996), Shlesinger (1997) and Gile (2003). Shlesinger’s article referred to here sums up a discussion on quality in which many eminent members in the interpreting research community participated. There seems to be a general consensus that quality cannot be assessed in absolute terms: it has to be handled from many different angles, not only the communicative event, naturally, but also its fidelity to the information in the source speech and its value as a speech of its own.

Quality is a well researched topic in interpreting studies. Several investigations have been made in the field and many methods for evaluating the product have been developed. Bühler (1986) conducted the first larger-scale survey on interpreters’ own perception of quality, while Kurz (2002: 323) undertook a similar investigation in different user groups. The result of these surveys was somewhat similar in that interpreting is good if it serves its purpose;
an ideal interpreting is not an absolute value but depends on e.g. situationality and communicative context. The studies conducted by Bühler and Kurz are now being followed up by a doctoral thesis on quality and the role of the interpreter (Zwischenberger, Pöchhacker & Kurz 2008). Clearly, the quality in this case is investigated by means of audience response, with both interpreters and non-interpreters as the audience. The more recent, regular and larger studies of quality from this perspective are done by the Interpreting Services at the European Commission (SCIC) in its regular customer satisfaction survey (2007, 2010). Among the studies from the customers’ perspective are also Ng (1992) and Vuorikoski (1998). In the study reported here on assessment of the product, both the interpreter’s perspective and the user’s perspective were taken into account.

Relating quality to the investigation of expertise, Ericsson (1991: 15) states that “although judges can reliably assess the superior quality of the product, it is difficult to analyze such products in order to identify the measurable aspects capturing the superior quality of the product.” Hence, it is important to combine the assessment of quality with the investigation of the cognitive processes underlying the production of an interpreting of a certain quality.

In the second study presented here below, Carroll’s scales (1966) were used as tools to assess the products of the translation (see below). Carroll’s scales were originally designed to measure the quality of machine translation. In the present study they have been adapted to interpreting. There are two scales: one measures the intelligibility of the product on its own, while the other measures the informativeness of the product compared to the original.

For a detailed presentation and investigation of Carroll’s scales, their adaptation to interpreting and their appropriateness for this type of assessment, see Tiselius (2009).

2.4 The Studies

In this article we report on two studies, one focusing on the process of interpreting, and the other one focusing on the product. The same data were collected for the two studies: nine subjects with different levels of experience interpreted the same nine-minute-long speech. Immediately after interpreting, they carried out a retrospection task using a written transcription of the original speech as cue. The interpreting performances and the retrospections were recorded and transcribed. In the first study, the interpreting performances were compared across the different levels of experience. In the second study, the interpreting products were assessed by 12 different assessors, both interpreters and non-interpreters, and then the performance ratings were compared.
3. First Study: Interpreting Process

In the first study we investigated the differences in the interpreting process in the groups of interpreters with different levels of experience. We took the retrospections as a starting point in our analysis and classified the reported instances from the retrospections into processing problems, monitoring and strategies, following Ivanova’s (1999) method (see appendix 1). These instances were then compared across the three groups in order to investigate whether the pattern differed depending on the level of experience.

3.1 Material and Method

3.1.1 Input Material

The speech to be interpreted was held by Commissioner Byrne in the European Parliament in May 2002 and was approximately nine minutes long (00:09:35). For the purpose of the present study it was slightly adapted and rerecorded using another speaker. Some names and figures were added in order to increase the cognitive load, and the pace was adjusted to a slower speech rate in order to make it possible for all of the subjects to interpret. The original speech had an average speed of 141 words-per-minute and the speaker had a pronounced Irish accent. The re-recorded speech was delivered at an average speed of 119 words per minute in Received Pronunciation (RP).

As Ivanova (1999) points out, background knowledge is important for an expert performance, enabling the expert to anticipate and to draw conclusions from previous experience. For this reason Ivanova chose a bona fide conference speech. This is also the reason why a speech from a real-life situation was chosen for the present study. The written transcript of the speech is provided in appendix 2.

The speech was recorded on a CD in a soundproof studio, with a frequency range of about 20,000 Hz, which is an important criterion for good interpreting working conditions, as compared with the frequency range of a telephone line, which is about 4,000 Hz.

3.1.2 The Interpreting Subjects

The participants in this study (n = 9) had three different levels of interpreting experience; no, short and long. Those with short and long experience had both training and experience, while those with no experience did not have any

---

8 There seems to be a limit to how fast interpreters can actually process an utterance. Seleskovitch (1965) suggested that 100–120 words per minute (wpm) is a comfortable speed, and 150–200 the upper limit. Gerver’s study (1969, 1971) supported her estimate.
interpreting experience. The no-experience group was recruited among translation students beginning their second semester at the Institute for Interpretation and Translation Studies (TÖI) at Stockholm University. Students come to TÖI after at least one year of undergraduate language studies. They had seen the inside of an interpreting booth during an introductory lesson to conference interpreting, where they spent 15 minutes in the booth trying to interpret an easy text from English into Swedish. The interpreters with short experience were recruited among the former students of the European Masters in Conference Interpretation at TÖI. All of them had two years of experience at the time of the recording. The interpreters with long experience were recruited among the interpreters at the European Parliament, and all had at least 25 years of experience (see table 1).

The interpreters with long experience in this study clearly have credentials that would indicate some level of expertise: very long experience, accreditation at international institutions and membership in the international conference interpreters’ organization, AIIC, where prospective members must be sponsored by full members to obtain membership. However, this is just an indication of expertise and not support for factors such as outstanding performance or deliberate practice (see above under 1).

All professional interpreters had Swedish as their A-language and English as a C-language according to the language classification of the International Association of Conference interpreters, AIIC. The interpreters with no experience had English as one of their working languages in the translation program. All groups were thus highly proficient in English, but none of them judged themselves to be native or near-native.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age span</th>
<th>Years at university</th>
<th>Interpreting school</th>
<th>Years of experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>No experience</td>
<td>20–29</td>
<td>4</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>Short experience</td>
<td>30–49</td>
<td>4</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Long experience</td>
<td>50–60</td>
<td>4</td>
<td>Yes</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 1. Age and experience of the interpreters

9 A-language: the AIIC (International Association of Conference Interpreters) language classification of a language which one masters at mother tongue level and interprets into. B-language: a language other than the interpreter’s native language, of which she or he has a perfect command and into which she or he works from one or more of her or his other languages. C-language: languages of which the interpreter has a complete understanding and from which she or he works. http://www.aiic.net/ViewPage.cfm?page_id=199#langclassif (accessed October 14, 2009).
3.1.3 The Retrospection Procedure

In accordance with Ivanova’s method, the subjects were presented with a transcribed copy of the original speech immediately after the interpreting task. They were asked to go through the speech sentence by sentence and try to recall everything they had thought of while interpreting.

During both the interpreting and retrospection processes, the researcher sat slightly obliquely behind the subject, in order to be close should any technical problems occur, but not too close in order not to disturb the subject during the interpreting process. During retrospection the researcher tried to interfere as little as possible. The retrospection method used here is presented in depth in Englund Dimitrova and Tiselius (2009).

3.1.4 Transcription and Analysis

The input and output were taped simultaneously on a double-channel recorder and then digitalized as a WAV file. The retrospection interview was taped onto the same tape recorder and then digitalized into a WAV file. The speech, the interpreting and the retrospection were all transcribed and coded using CHILDES software (MacWhinney 1991).

The interpretations were analyzed taking the retrospection as a starting point. Any reported processing problem, instance of monitoring or use of strategy was classified according to Ivanova’s categories; although Ivanova only counts a strategy use if it has been coupled to a processing problem or an instance of monitoring, in this study all reported strategies have been counted.

The data were primarily analyzed using Correspondence Analysis (CA). CA is an exploratory multivariate statistical method that can be used to visualize the structure of the data (Greenacre 2007: 1). This constitutes a major methodological advantage over looking at percentages calculated from raw frequencies. Consider the difference between the frequencies 100 and 101: in this case, the difference of 1 observed unit is small, since it only amounts to a 1-percent increase. With low frequencies, on the other hand, adding a single observation makes a large difference: if we consider a fairly low frequency such as 5, adding 1 observation would amount to a 20-percent increase, something that makes it nearly impossible to compare percentages from small numbers of observations. CA, on the other hand, takes the total number of observations into account. For technical details, see Greenacre (2007).

The input to a CA analysis is a table, where the rows correspond to observations, the columns correspond to properties of these observations, and the cells contain the frequency of each property per observation (see table 1). CA is not a hypothesis test (hence, no p-values are produced for the CA analyses); instead, the idea is to describe as accurately as possible the structure in the data, primarily through graphical representations called biplots.
The biplots can be read as associations between row and column variables, that is, between observations and their properties. For instance, two row variables that are close to each other on the x-axis of the biplot are more closely associated than two variables that are further apart. Similarly, distance between variables on the y-axis can also be interpreted as associations.

The great advantage of CA is that it allows for simultaneous visualization of the interaction of a number of nominal row and column variables. Thus, the analysis of any one variable will be enlightened by its simultaneous interaction with all the other variables. All the statistical analyses were carried out using the software package R (R Development Core Team 2009); the CA analyses were carried out with the *ca* package in R (Greenacre and Nenadic 2007).

### 3.2 Results

Below, we present the results of the investigation of processing problems, instances of monitoring and use of strategies among the three different groups of interpreters.

#### 3.2.1 Processing Problems

As can be seen in figure 1 and table 2, the largest difference in processing problems is between experienced interpreters and interpreters with no experience. The difference between interpreters with long or short experience is less pronounced. Nevertheless, there are interesting differences in the types of processing problems.

Table 2. Processing problems (definitions of the headings are given in appendix 1)

<table>
<thead>
<tr>
<th>Processing problems (PP/)</th>
<th>Long-experience interpreters</th>
<th>Short-experience interpreters</th>
<th>No-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehension (C/)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perception (P)</td>
<td>8</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Lexical access in SL (L)</td>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Syntactic processing (Syn)</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Text integration (TC/integ/)</td>
<td>5</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Text comprehension (TC/bgkn)</td>
<td>5</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Translation (Tr/)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TL retrieval (TLr)</td>
<td>6</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>Equivalent (eqv)</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Simultaneity of tasks (Sim/)</td>
<td>7</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>TL delays (Tr.del)</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>30</td>
<td>108</td>
</tr>
</tbody>
</table>
Figure 1 shows a CA plot of the data in table 2. Most of the variation in table 2 is accounted for by the $x$-axis, and there is a clear difference between long and short experience on the one hand, and no experience on the other. All three groups are fairly strongly associated with the $x$-axis judging from their correlations with the $x$-axis (see table A.4, appendix 3), which suggests that there is a primary difference between degrees of experience and no experience at all.

Turning to the processing problems, the categories with the highest inertia are lexical access (Lexical_access), target language delays (TL_delays), equivalent (Equivalent) and source language input rate (SL_TL) problems. The first two are positively associated with no experience (and negatively associated with short or long experience), equivalents are positively associated with both short and long experience (and negatively associated with no experience), while source/target language problems are positively associated with long experience, but hardly associated with short or no experience at all. The other processing problems have very little explanatory value compared with these four main categories.

However, there is a further distinction between interpreters with short experience and those with long experience. When considering the combined effects of inertia and correlation with the axis, it becomes clear that it is primarily the source language input rate (SL_TL) problems that are responsible for the difference between long and short experience. This category is positively associated with long experience on the $y$-axis (and negatively associated with short experience). We find a similar (but weaker) pattern for syntactic processing.
that appears to help distinguish interpreters with long experience from those with short experience, but it is clearly secondary to the *source language input rate* (SL_TL). Turning to *equivalent*, the $y$-axis reveals that problems with this category appear to be somewhat more frequently associated with short experience than with long experience. See tables A.4 and A.5 in appendix 3 for the full numerical output.

In table 3 our findings are compared with Ivanova’s findings. The observed frequencies were compared by means of a Spearman rank sum test, which found no significant difference between the relative ordering of the categories in the two studies ($p > 0.05$). Thus, we can conclude that the figures vary along the same lines and that, despite some variations (for *text comprehension* and *translation*), the figures in the two studies clearly support each other.

*Table 3. Processing problems, long experience Ivanova vs. the present study*

<table>
<thead>
<tr>
<th>Processing problem</th>
<th>Ivanova’s experts (av. 9 years of interpreting experience)</th>
<th>Long-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Syn</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>TC</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Tr</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Sim</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>38</strong></td>
<td><strong>48</strong></td>
</tr>
</tbody>
</table>

Table 4 compares Ivanova’s novices and the short- and no-experience interpreters of this study. As above, the categories were compared using Spearman’s rank test, and again we found no significant difference between the two studies.

*Table 4. Processing problems, Ivanova’s novices vs. short- and no-experience interpreters in this study*

<table>
<thead>
<tr>
<th>Processing problem</th>
<th>Ivanova’s novices (three months interpreter training)</th>
<th>Short-experience interpreters</th>
<th>No-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>24</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>L</td>
<td>13</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Syn</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>TC</td>
<td>39</td>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td>Tr</td>
<td>21</td>
<td>12</td>
<td>28</td>
</tr>
<tr>
<td>Sim</td>
<td>26</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>129</strong></td>
<td><strong>30</strong></td>
<td><strong>108</strong></td>
</tr>
</tbody>
</table>
3.2.2 Instances of Monitoring

In table 5 it is also clear that the interpreters with long experience differ from the other three groups in types of monitoring. We can assume that the long-experience interpreters have enough attention to spare (cf. Gile’s effort model 1985) to allow themselves to reflect on the accuracy of the utterance before pronouncing it, while this is not the case for less experienced interpreters.

Table 5. Instances of monitoring

<table>
<thead>
<tr>
<th>Monitoring code (M/)</th>
<th>Long-experience interpreters</th>
<th>Short-experience interpreters</th>
<th>No-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation (tr)</td>
<td>16</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Inner speech monitoring (insp)</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Time (tm)</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Internal commentary (int. com.)</td>
<td>14</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Mood</td>
<td>12</td>
<td>16</td>
<td>39</td>
</tr>
<tr>
<td>Id</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>45</strong></td>
<td><strong>30</strong></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>

Figure 2 presents a CA biplot of the data in table 5. The plot shows a distinction between interpreters with long experience on the one hand, and those with short or no experience on the other. Long experience is associated with a high inertia and is strongly correlated with the x-axis. Short and no experience are more correlated with the y-axis than the x-axis, indicating that they are primarily distinguished from each other on the y-axis. The difference between short and no experience should not be underestimated, since the y-axis accounts for 37.6% of the variance. However, the primary difference appears to be that between long experience, and short and no experience.

The monitoring observations that have the highest explanatory value (judging from inertia) are translation (Transl), time, inner speech monitoring (InnerSpMon) and non-analyzed problems (Id). Internal commentary contributes as much as we would expect given its relative frequencies, while mood appears to have less explanatory value than its relative frequency would suggest. The strongest contribution to the x-axis is made by translation, which is positively associated with long experience and negatively associated with the other two categories. There is also some contribution from mood, that is, emotive self-evaluations (primarily associated with no experience). On the y-axis, short and no experience are distinguished by internal commentary, Id (non-analyzed problems) and time. Time and to a certain extent internal commentary are positively associated with short experience (and negatively associated with no experience), while Id is positively associated with no experience (and negatively associated with short experience).
Thus, *translation* is by far the most frequent monitoring observation for interpreters with long experience, a feature that distinguishes long-experience interpreters from the other two groups. The most frequent monitoring observation for interpreters with short experience is *time*, while those with no experience seem to experience *non-analyzed problems* (Id), instances on monitoring that it has not been possible to classify.

See tables A.6 and A.7 in appendix 3 for a full numerical output.

![CA biplot of the association between interpreter experience and instances of monitoring. The first axis accounts for 62.4% of the total inertia, the second for 37.6%.

The types of monitoring instances of translation also differ, despite the equal numbers for long-experience and no-experience interpreters. The interpreters with long experience reflect on better ways to interpret a certain utterance, whereas the no-experience interpreters try to find a general coherence in the output.

The similarities to Ivanova’s findings are not as obvious for interpreters with long experience as they are when comparing processing problems. The two most important instances, viz. translation and mood, are in fact similar; however, whereas Ivanova’s experts report both on time and inner speech monitoring, our interpreters with long experience have few and no such instances. Instead, our interpreters with long experience have a number of instances for internal commentary, which Ivanova’s interpreters with long experience have considerably less of. However, as with processing, the Spearman rank test found no significant difference between the two studies, which suggests that the observed differences are so small that the two studies can be compared.
Table 6. Instances of monitoring, Ivanova’s expert vs. the long experience interpreter in the present study.

<table>
<thead>
<tr>
<th>Monitoring code</th>
<th>Ivanova’s experts</th>
<th>Long-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>Time</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Insp</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Int. com.</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Mood</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Id</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>46</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 7 compares Ivanova’s novices to the interpreters with short and no experience in this study. *Mood* is a category that stands out both in table 7 and table 6. Simultaneous interpreting seems to be a very emotional experience, both for experienced interpreters and novices, although a little less so for the former. As in the previous cases, we used Spearman’s rank test to compare the results of the two studies, and the result was not significant, which indicates that also in this case, no overall difference exists between the two studies.

Table 7. Instances of monitoring, Ivanova’s novices vs. the short- and no-experience interpreters in the present study

<table>
<thead>
<tr>
<th>Monitoring code</th>
<th>Ivanova’s novices</th>
<th>Short-experience interpreters</th>
<th>No-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Time</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Insp</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Int. com.</td>
<td>1</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Mood</td>
<td>7</td>
<td>16</td>
<td>39</td>
</tr>
<tr>
<td>Id</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>16</td>
<td>30</td>
<td>24</td>
</tr>
</tbody>
</table>

### 3.2.3 Strategies

Table 8 shows the raw frequencies for strategies employed by the groups of interpreters. To assess the correlation of length of experience with individual strategies, we again made use of CA. The result is shown in figure 3. The *x*-axis accounts for 85.1% of the variance, and the *y*-axis for the remaining 14.9%. On the *x*-axis, there is a continuum from no experience on the right, via short experience near the middle to long experience on the left-hand side.
Table 8. Instances of reported strategies

<table>
<thead>
<tr>
<th>Strategy code</th>
<th>Long-experience interpreters</th>
<th>Short-experience interpreters</th>
<th>No-experience interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection (S)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Summarization (SUM)</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Restructuring (Rest)</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Creative Interpretation (CR)</td>
<td>1</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Overgeneralization (Overgen)</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Deletion (D)</td>
<td>18</td>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>Explication (Expl.)</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Compromise (ACC)</td>
<td>7</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>41</strong></td>
<td><strong>62</strong></td>
<td><strong>63</strong></td>
</tr>
</tbody>
</table>

It should be noted that the short- and long-experience groups are closer to each other than either of them is to the no-experience group. However, the numerical analysis reveals that long and no experience are primarily responsible for the $x$-axis, while short experience draws up the $y$-axis; thus, it is the opposition between long and no experience that has the greatest explanatory value. Furthermore, the short experience point is close to the origin, indicating that that category has less explanatory value than the other two.

Deletion (D) is the most common strategy and is not strongly associated with any of the groups; hence, we must turn to other strategies to distinguish the interpreter groups.

Interpreters with long experience are strongly associated with overgeneralization (Overgen). Interpreters with no experience, on the other hand, are most strongly associated with creative interpreting (CR), i.e. guessing. For the interpreters with short experience, restructuring (Rest) and summary (SUM) are most characteristic, and the latter in particular appears to have some explanatory value with respect to short experience. Compromises (ACC) are also somewhat correlated with this group, but have too low inertia to have much explanatory value. Similarly, selection (S) plays virtually no role in defining the projection seen in figure 3.

See tables A.8 and A.9 in appendix 3 for a full numerical output.
Figure 3. CA biplot of the association between interpreter experience and strategies. The first axis accounts for 85.1% of the total inertia, the second for 14.9%.

Strategies have not been compared to the results in Ivanova, since Ivanova only counts strategies in relation to a processing problem, whereas in this study all strategies were counted. Indeed, for the strategies in relation to processing problems, the instances were as low as 13 for interpreters with short experience and 16 for interpreters with long experience. The figures are therefore far too low to make a comparison.

4. Second Study: Assessment of the Interpreting Product

In the second study, we investigated the assessment of the products of the interpreting through ratings. The ratings were done by two groups of raters, professional interpreters (n = 6) and students (n = 6), using Carroll’s (1966) scales to rate.

The results of the interpreting assessments were then compared across the groups of interpreters with different levels of experience, with the aim of investigating whether the interpreter’s experience made a difference.

4.1 Material and Method

The data for the rating study was the same as in the process study: transcriptions of the nine interpretations of the same nine-minute-speech (see sections 3.1.2 and 3.1.3 above). In an initial, unpublished study, the assessments of interpreting from transcripts and sound files of the same interpreting were compared and no significant difference was found. Therefore, transcripts were used in order to avoid recognition of the interpreters in the study by the interpreter raters.
4.1.1 The Rating Files

In order to assess the product, all of the transcribed interpreting performances were divided into shorter units, in total 162. These units were then mixed randomly and assembled to six assessment files consisting of 54 units (meaning that each unit got rated twice by different assessors). Each assessment file consisted of randomly mixed units from all the original interpreting performance and without a logical inner order from start to end. These steps were done in order to create an assessment with as even conditions for each assessor as possible. It was assumed that having one rater rate the same interpreter from start to end could possibly distort the rating. The rater could get used to the interpreter’s style and this could either improve or impair the scores.

4.1.2 The Raters

The raters consisted both of experienced interpreters (n = 6) and non-interpreters (n = 6) in order to obtain as broad assessments as possible and to even out differences that might be due to whether the assessor had interpreting experience or not. Gile (1999), for instance, found in one experiment that interpreters tended to be more lenient in their assessment of their peers than non-interpreters. The assessors were trained for the rating immediately prior to the task. Interpreter graders had 8–9 years of interpreting experience and both teaching and testing experience. Non-interpreter graders were university students from various university programs with at least 3 years of university studies.

4.1.3 The Rating Session

Each assessor was given a file of 54 units to grade according to Carroll’s scales for intelligibility and informativeness. When assessing intelligibility, the raters only saw the interpreted version; once they had rated the speech for intelligibility, they rated for informativeness by comparing the interpreting with the original speech.

4.1.4 The Scales and the Analysis

The scale for intelligibility is a six-point scale ranging from one to six, where six is the best score. When the interpreting unit is assessed for intelligibility, it is assessed as an independent speech segment of its own. This scale covers assessments of qualities such as fluency, clarity and so forth. In contrast, the scale for informativeness is used to assess the interpreted speech in relation to the source speech, evaluating how much it differs in terms of information from the original speech. This scale covers assessment of fidelity and accuracy. It runs from six to one, with one being the best score. It also has a step 0 to cover explicitation (cf. Englund Dimitrova 2005), where the interpreting contains more information than the original.

These scales were designed to assess the interpreters’ rendition from a perspective of general understanding; therefore, they can be said to have holistic qualities. The scales are provided in appendix 3. For a detailed presentation of
the methodology of this part of the study, see Tiselius (2009). In the present study, the Friedman rank sum test has been used to test whether a difference exists between the groups of interpreters in terms of their product.

4.2 Results

In this section we present the results for informativeness and intelligibility from the interpreting transcripts.

4.2.1 Intelligibility and Informativeness

Table 9 gives the scores from the assessment for intelligibility and informativeness of the interpretations. As pointed out in section 4.1.4 above, a high score for intelligibility means a more comprehensible interpreting product, whereas a low score for informativeness means little difference between source speech and target speech.

Table 9. Mean evaluation scores of intelligibility and informativeness.

<table>
<thead>
<tr>
<th>Interpretings</th>
<th>Intelligibility</th>
<th>Informativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-interpreter</td>
<td>Interpreter</td>
</tr>
<tr>
<td></td>
<td>grader</td>
<td>grader</td>
</tr>
<tr>
<td>No experience</td>
<td>3.79</td>
<td>3.16</td>
</tr>
<tr>
<td>Short experience</td>
<td>5.25</td>
<td>4.88</td>
</tr>
<tr>
<td>Long experience</td>
<td>5.42</td>
<td>5.11</td>
</tr>
</tbody>
</table>

To investigate the possible differences between the groups of interpreters, we carried out two Friedman rank sum tests on the ratings for intelligibility and informativeness, testing the null-hypothesis that there is no difference between the interpreter-experience groups. One test was carried out for intelligibility (Friedman \( \chi^2 \text{ (df = 2)} = 38.5, p < 0.01 \)), and one for informativeness (Friedman \( \chi^2 \text{ (df = 2)} = 53.7, p < 0.01 \)). There was a significant difference at the 0.01 level in both cases, indicating that there are differences between the groups somewhere in the data. To pinpoint the between-group differences, we applied a Nemenyi-Damico-Wolfe-Dunn post-hoc test for pairwise comparisons.\(^\text{10}\) We found that for intelligibility, there is a significant difference between no experience and long experience, and between no experience and short experience. No difference was found between long and short experience with respect to intelligibility. For informativeness, we found that all three groups are significantly different from each other. Figures 4 and 5 show boxplots of the differences between groups and give the results of the post-hoc tests. The plots show differences between groups: a result close to zero indicates no difference.

\(^{10}\) R code for this test was provided by T. Galili from http://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code/.
The test results suggest that for *intelligibility*, experience vs. lack of experience has an effect, but once a minimum experience threshold is reached, there is no added effect. For *informativeness*, on the other hand, there appears to be a cumulative benefit of experience, so that a little experience is better than no experience, and long experience is better than short experience.

![Figure 4. Intelligibility: boxplot of between-group differences. Significant differences are gray. Note the non-significant difference between short and long experience.](image1)

![Figure 5. Informativeness: boxplot of between-group differences. Significant differences are gray. All differences are significant, but the largest differences are experience vs. no experience.](image2)

### 4.2.2 The Effect of Raters

In order to assess the reliability of the results reported above, it was important to quantify the influence that the selection of raters had on the results. As table 9 above indicates, the scores for non-interpreter (i.e. student) graders are a little more extreme than those produced by experienced graders. Since the Friedman rank sum test employed above is robust and makes very few assumptions about the data, it is a good indicator of any differences between groups. However, it does not take the effect of the raters into consideration.

In order to measure the effect of the raters, a more complex test type known as a mixed-effects model was employed. For our purposes, it is sufficient to note that such models allow for complex interactions between multiple factors, while simultaneously incorporating information about group correlations in the data. A full treatment of such models is not possible here; see instead Pinheiro and Bates (2000) or Baayen (2008) for introductions and explanations. The models were fitted with the lme4 package (Bates and Maechler 2009) in R.

The problem to be solved with this approach was as follows: the raters were drawn from two different groups, which might influence the outcome of the
rating session. Treating the ratings as coming from a homogeneous group of assessors could potentially lead to misleading results if these rater groups perceived the products in very different ways. For both data sets (viz. intelligibility and informativeness), the same approach was followed: first, a model was fitted that measured the simple effect of interpreter experience (no, short, long) and rater group (student, interpreter) on the ratings; second, a more complex model measuring the interaction between interpreter experience and rater group was fitted to the same data. All models were found to fit the data well, judging by standard diagnostic procedures.

Overall, the results were congruent with the results presented in the previous section: for intelligibility, we found a statistically significant difference between no experience and long experience (p < 0.01), but not between long experience and short experience (p = 0.23). For informativeness, we found statistically significant differences between no experience and long experience, as well as between long experience and short experience (all p < 0.01). This supports the conclusions from the previous sections with respect to the effect of interpreter experience.

In the first, simple model fitted to the data sets, a significant difference between the rater groups was found both for intelligibility and informativeness (both p < 0.05). This was expected, based on the numbers reported in table 9. The real question, however, was whether this had any systematic effects on the overall outcome of the assessments. The second, more complex model, which measured the interaction between rater group and interpreter experience, found this not to be the case (all p > 0.05). Once the interactions were taken into account, the effect of rater group was no longer statistically significant. This result corroborates the impression from table 9, namely that although the student raters as a group give more extreme assessments than the experienced raters, this has no effect whatsoever on the relative ranking of the products by the three groups of interpreters. In conclusion, there is a difference in magnitude between student and experienced assessors, but the overall trend of the ratings is the same, which suggests that the approach in section 4.2.1 above was justified.

5. Discussion

The aim of the two studies presented here was to explore the difference (if any) of both process and product between interpreters with no experience, short experience and long experience. This section discusses the result of this study as well as some limitations. Finally, some conclusions regarding the initial research questions are offered.

11 The p-values were obtained through MCMC sampling with the pvals.fnc function from Baayen (2009).
5.1 Discussion of Results

Our study involved relatively few participants; hence, it is difficult to draw exhaustive generalizations from our material. However, until larger studies are conducted, the material allows for certain tentative generalizations. First, the patterns that emerged from our material were relatively clear, and consistent with previous research and theoretical assumptions. Second, the material was gathered under controlled, experimental conditions. This implies that all the participants were tested under the same conditions, which should bring to the forefront any differences due to amount of experience by removing other confounding factors that might otherwise arise. The experimental set-up has the added benefit of offering compatibility with future studies that employ a similar approach.

With respect to processing problems, we found that the no-experience group struggled mostly with lexical access and target language delays. The experienced interpreters (long and short) were clearly distinct in this matter from those with no experience. For those with short experience, equivalents were found to be the most characteristic processing problem. The interpreters with long experience had no difficulties with delays or lexical access, but experienced some difficulties with source language input/target language output rate (SL_TL).

In the case of monitoring, a different pattern emerged, where long experience was distinguished from the other two categories; the amount of experience appears in other words to play a more active role. For long experience, there was a preference for translation (i.e. ascertaining accuracy of translation). Interpreters with short and no experience were grouped together, but could be further distinguished by the fact that short experience was associated with time and no experience with Id.

In the case of translation strategies, long experience and no experience stood out, whereas interpreters with short experience were fairly close to the overall average. The most common strategy in all groups was deletion, which is not surprising considering the fact that interpreters work under a time constraint and a certain amount of deletion comes from compressing the message. Consequently, this deletion could not be used to distinguish the groups. Instead, the most distinctive strategy for interpreters with long experience was overgeneralization, that is, using a more general term in the translation (such as “mammal” for “cat”). For interpreters with no experience, creative interpreting was the most important strategy, whereas those with short interpreting experience (which was close to the expected average) showed a preference for summary.

It was found that the results concerning processing problems in our study supported the findings in Ivanova (1999). For monitoring, the figures were not as unambiguous as for processing problems. With respect to the categories translation and mood, we could also corroborate Ivanova’s results. However, it
should be noted that many of the monitoring tendencies may be due to personality or interpreting norms, which may differ according to nationality, language and interpreting school.

The experience levels of the interpreters participating in the present study and that of Vik-Tuovinen (2006) are not completely similar. However, our results support Vik-Tuovinen’s (2006) conclusion that beginners focus on source text and linguistic expression, since the subjects with no experience in the present study struggle with problems of lexical access. Among the interpreters with extensive experience in the present study, the monitoring category translation dominated, and this might very well mirror Vik-Tuovinen’s experienced interpreters, who focused more on situational factors than less experienced interpreters.

Liu (2001) found that interpreters with extensive experience monitored the output better. Although her test subjects were not fully comparable in terms of experience to the subjects participating in the present study, Liu’s results were supported in our study of the interpreting process: interpreters with long experience make more use of monitoring strategies, and in particular translation and internal comments.

When it comes to the analysis of the product, the data for intelligibility and informativeness were tested using the Friedman rank sum test and the Nemenyi-Damico-Wolfe-Dunn post-hoc test. The results clearly indicate that experience results in higher intelligibility once a minimum level of experience has been attained, but that there is no added benefit of longer experience beyond this minimum. For informativeness, on the other hand, longer experience has a cumulative effect, so that a short experience leads to better scores than no experience, and long experience gives better scores than short experience. Through mixed-effects models, it was demonstrated that rater experience did not influence the overall trend of the assessment of the product.

The fact that intelligibility is not a distinguishing factor between interpreters with short and long experience is not surprising considering the fact that interpreters with training and some professional experience have been taught how to produce a comprehensible message. To find differences in the interpreting product between interpreters with long and short experience, it is necessary to look beyond intelligibility, to informativeness.

5.2 Conclusion
We have demonstrated that the impact of experience in interpreting is dependent on the task or problem at hand. Both interpreting experience and general background knowledge are important for the solution of processing problems in different situations. In all cases, it would seem that some minimum of experience is desirable, but only in some areas did we find a difference between the long- and short-experience groups.
It seems there is a clear difference between the dichotomy experience vs. lack of experience both in process and product. It should be noted that a major difference, in terms of experience between the student group (no experience) and the two groups of interpreters with some degree of experience, is that the two groups of interpreters both had experience interpreting at the European Parliament and the speech chosen was from the European Parliament. This might give them a double edge not only in terms of simultaneous interpreting technique, but also in knowledge of the European institutions. Thus, the groups of experienced interpreters had experience in that field and perhaps a certain amount of expertise as well.

Degrees of experience (short and long) proved a less profound opposition in our material, although some differences could be seen in the preferences of strategies and perceived problems of the long- and short-experience groups. The exceptions to this were monitoring and informativeness, where long experience appears to have a larger impact (and give more of an advantage) than in the other cases. It should also be stressed that we still do not know for sure whether the interpreters with long experience are experts or not. This is a limited study and all criteria for studying expertise (see section 1) have not been covered here. Expertise must necessarily be studied through proxy indicators. Although the indicators studied could be expanded, it can be assumed that these very experienced interpreters have at least some degree of expertise. When we combine the results of the two studies we see that for strategies, long-experience interpreters report more instances of overgeneralization, whereas short- and no-experience interpreters report more deletion and creative interpreting. Since subjects rate products from interpreters with long experience higher, this can indicate that the interpreting is perceived to be closer to the original when overgeneralization is employed, compared with deletion or creative interpreting.

To sum up, certain processing problems, instances of monitoring and strategies may indeed tell us something about expertise. They may indicate which components of the performance should be studied in detail to find the superiority Ericsson and Smith (1991) encourage us to look for. It may very well be that monitoring and informativeness are the components that make the expert performance superior. The challenge ahead is to dig deeper into these differences in order to understand and define expertise in interpreting.
References


## Appendix 1

### Table A.1. Processing problems. Ivanova (1999)

<table>
<thead>
<tr>
<th>Processing problems (PP/)</th>
<th>Perception (P)</th>
<th>Problems with hearing (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehension (C/)</td>
<td>Lexical access in SL</td>
<td>Failure to access meaning of a SL chunk, which has been identified as familiar</td>
</tr>
<tr>
<td>Syntactic processing (Syn)</td>
<td>Text integration (TC/integ/)</td>
<td>Failure to recognize syntax patterns</td>
</tr>
<tr>
<td>Text comprehension (TC/bgkn)</td>
<td>Difficulties in constructing a coherent representation for SL chunks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comprehension difficulties due to lack of background knowledge</td>
<td></td>
</tr>
<tr>
<td>Translation (Tr/)</td>
<td>TL retrieval (TLr)</td>
<td>Problems in rendering a SL chunk in TL</td>
</tr>
<tr>
<td></td>
<td>Equivalent (eqv)</td>
<td>Problems in selecting an appropriate equivalent when there is a choice</td>
</tr>
<tr>
<td>Simultaneity of tasks (Sim/)</td>
<td>(SL, TL)</td>
<td>Problems due to high SL input rate in relation to interpreter’s own output rate</td>
</tr>
<tr>
<td></td>
<td>TL delays (Tr.del)</td>
<td>Delays in TL product due to translation</td>
</tr>
</tbody>
</table>

### Table A.2. Monitoring observations. Ivanova (1999)

<table>
<thead>
<tr>
<th>Monitoring (M/)</th>
<th>Translation (tr)</th>
<th>Inner speech monitoring (insp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ascertaining accuracy of translation at the conceptual level against an ST representation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verification of the TL message against TL rules prior to articulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time™ Awareness of the timing relative to the TL production</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internal commentary (int. comm.) Affective commentary to ST/ST producer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mood Emotive self-evaluation of performance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Id Non-analyzed problems</td>
<td></td>
</tr>
</tbody>
</table>

### Table A.3. Strategies. Ivanova (1999)

<table>
<thead>
<tr>
<th>Strategy code (SC/)</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection (S)</td>
<td>Selection of one SL chunk for further processing because it is more informationally or pragmatically salient</td>
</tr>
<tr>
<td>Summarization (SUM)</td>
<td>Rendering the gist of a SL segment</td>
</tr>
<tr>
<td>Restructuring (Rest)</td>
<td>Changing the original syntactic structure of a SL segment (usually by transposing clauses or segments within the clause), in order to improve the expression in TL, or anticipation of problems</td>
</tr>
<tr>
<td>Creative Interpretation (CR)</td>
<td>Compensating for missing information by guessing on the basis of previous knowledge</td>
</tr>
<tr>
<td>Overgeneralization (Overgen)</td>
<td>Rendering an aspect of ST (e.g. a lexeme) by selecting a more abstract, hence less committing representation in TL</td>
</tr>
<tr>
<td>Deletion (D)</td>
<td>Omitting SL chunks because of time constraints without reference to their semantic or pragmatic role</td>
</tr>
<tr>
<td>Explication (Expl.)</td>
<td>Explicitly expressing information implied in the ST</td>
</tr>
<tr>
<td>Compromise (ACC)</td>
<td>Lower the acceptability standards for a TL production in order to minimize processing costs</td>
</tr>
</tbody>
</table>
Appendix 2

Original Speech


Mr. President, honorable Members,

It is with great pleasure that I reply to this debate on the Committee report concerning the Tobacco Products Directive. As you are all well aware, this directive is the culmination of several years of hard work by the Commission, and has been developed on the basis of the best scientific advice.

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.

I need hardly remind Parliament of the level of public damage to the health of our population from tobacco. This damage exceeds many other threats to public health that are far more widely debated. The World Health Organization, for instance, estimates tobacco-related deaths in the Community to exceed half a million a year, and this number does not even include the number of smokers who suffer from serious and debilitating illnesses, such as emphysema, heart failure, TB, lung and liver cancer. Because of this major impact on human health, tobacco control emerges as the most important tool in reducing death and disease in our midst.

I must here compliment Parliament and in particular the rapporteur, Mr. Maaten. As you all know, the amendments he has tabled have substantially improved and completed the directive. The Commission was happy to accept many of these amendments. Mrs. Maes went as far in her contribution to Parliament as to make a statement that this was a textbook example of cooperation between the three great legislative institutions of the European Union. I would echo that and go even further and say that this issue is worthy of close examination by any serious student of the European Union or its legislative procedure. Both the advertising directive that I will be presenting to you soon, and this particular tobacco directive on which you will vote today, are worthy of such an examination. The directive provides a clear example of how all the institutions can work well together, both in formulating policy, and in bringing that policy forward in the form of valuable legislation. And all this has been done within the confines of the European legislation from the Treaty of Rome and onwards. Also, it will surely calm anybody who fears that the European institutions are becoming too powerful.

This directive recasts several existing directives from the late 1980s. We have used this proposal to update and complete existing instruments on the basis of scientific developments.

First of all I should like to mention the setting of a reduced tar ceiling in cigarettes and the creation for the first time of a ceiling for nicotine and carbon monoxide. These ceilings will directly attack the agents in cigarettes responsible for cancer, addiction, and cardiovascular disease. I wish to point out that we are not, however, aiming at producing a safe cigarette, since, as everybody knows, such a product cannot exist.

Importantly, the directive now applies to all imported products and to all products manufactured in the Community. We will no longer export any products that we consider inappropriate for our citizens. This will also eliminate the risk of unsafe exports coming back onto the Community market through smuggling.
The tobacco industry has tried to use the spectre of job losses to threaten this measure, and these cynical efforts have made me particularly upset. The same threats were also expressed when the original tobacco directives were adopted in the late 1980s. These threats were not true then and they are not true now. There is no room for double standards for our own citizens or others, when we are talking about addictive and deadly products.

Another innovation in the text of this directive is to require that the tobacco producers declare what additives they put in tobacco products – potentially the single most important measure contained in this legislation. Currently we do not know what these additives are in the Member States or, more importantly, why they are added to the product. Manufacturers and importers in future will be obliged to declare these constituents.

As a result of an amendment from this Parliament on the 24th of July, 1999, the introduction of colour photos on product warnings is to be introduced for the first time. This will greatly increase their impact. In recent years, research has shown the importance of avoiding misleading product descriptors on tobacco products, because such descriptors may mislead the consumer into believing that one product is safer than another. This directive introduces a requirement not to use such misleading terms on tobacco products. As such it aims to protect smokers and non-smokers alike from misleading and dangerous descriptors such as light, low tar, double filtered, ultras, etc.

In a previous declaration to the Parliament, I undertook to consult tobacco experts. Richard Peto, who is statistician and epidemiologist at Magdalen College of the University of Loughborough reported that BAT, British American Tobacco, recently gave more than £3 million to sponsor Nottingham University. Now, every cigarette sold makes a profit for the company of about 3p. To get its donation money back, which we have to assume BAT wants to do, it has to sell 100 million extra cigarettes. Every million cigarettes causes about one death. So to break even, they’ve got to sell enough cigarettes to cause about 100 deaths. The money, incidentally, was given to fund a professorship in corporate responsibility.

I can also inform you today that in the coming weeks, I hope to meet my promise to present a new proposal on tobacco advertising and sponsorship. This will replace the directive annulled by the Court of Justice on the 18th of October last year.

I also regard the Community’s legislative activity as a complement to the legislation work under way in the World Health Organization Framework Convention on Tobacco Control. I welcome the presence of an observer from the European Parliament in the Community delegation, which early this month met in Geneva during the second round of negotiations for this convention.

Finally, I should like to pay tribute in particular to Mr. John Ryan and the significant contribution he has made to this issue over the years.

I am convinced that it will be seen in coming years that this directive is a keystone of our efforts to reduce the damage to health caused by smoking.

Mr. President, honorable Members, thank you for your attention.
### Appendix 3

Table A.4. CA output for the biplot in figure 1 (columns), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>ProcProblem</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prc</td>
<td>142</td>
<td>101</td>
<td>999</td>
<td>122</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L__</td>
<td>119</td>
<td>324</td>
<td>981</td>
<td>384</td>
<td>19</td>
<td>36</td>
</tr>
<tr>
<td>Sy_</td>
<td>17</td>
<td>23</td>
<td>33</td>
<td>1</td>
<td>967</td>
<td>127</td>
</tr>
<tr>
<td>Ttx_n</td>
<td>159</td>
<td>41</td>
<td>133</td>
<td>7</td>
<td>867</td>
<td>208</td>
</tr>
<tr>
<td>Ttx_c</td>
<td>136</td>
<td>13</td>
<td>661</td>
<td>10</td>
<td>339</td>
<td>25</td>
</tr>
<tr>
<td>TL_r</td>
<td>216</td>
<td>31</td>
<td>73</td>
<td>3</td>
<td>927</td>
<td>165</td>
</tr>
<tr>
<td>Eqv</td>
<td>74</td>
<td>162</td>
<td>997</td>
<td>195</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>SL_</td>
<td>91</td>
<td>147</td>
<td>492</td>
<td>87</td>
<td>508</td>
<td>434</td>
</tr>
<tr>
<td>TL_d</td>
<td>45</td>
<td>159</td>
<td>998</td>
<td>192</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table A.5. CA output for the biplot in figure 1 (rows), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>Experience</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>LnE</td>
<td>216</td>
<td>381</td>
<td>816</td>
<td>375</td>
<td>184</td>
<td>409</td>
</tr>
<tr>
<td>ShE</td>
<td>170</td>
<td>300</td>
<td>662</td>
<td>239</td>
<td>338</td>
<td>590</td>
</tr>
<tr>
<td>NEx</td>
<td>614</td>
<td>320</td>
<td>1000</td>
<td>386</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.6. CA output for the biplot in figure 2 (columns), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>Monitoring</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trn</td>
<td>206</td>
<td>324</td>
<td>906</td>
<td>471</td>
<td>94</td>
<td>81</td>
</tr>
<tr>
<td>ISM</td>
<td>14</td>
<td>49</td>
<td>828</td>
<td>65</td>
<td>172</td>
<td>22</td>
</tr>
<tr>
<td>Tim</td>
<td>43</td>
<td>103</td>
<td>390</td>
<td>64</td>
<td>610</td>
<td>167</td>
</tr>
<tr>
<td>InC</td>
<td>213</td>
<td>228</td>
<td>267</td>
<td>98</td>
<td>733</td>
<td>445</td>
</tr>
<tr>
<td>Mod</td>
<td>475</td>
<td>213</td>
<td>884</td>
<td>302</td>
<td>116</td>
<td>65</td>
</tr>
<tr>
<td>Id</td>
<td>50</td>
<td>83</td>
<td>5</td>
<td>1</td>
<td>995</td>
<td>220</td>
</tr>
</tbody>
</table>

Table A.7. CA output for the biplot in figure 2 (rows), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>Experience</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>LnE</td>
<td>319</td>
<td>418</td>
<td>977</td>
<td>656</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>ShE</td>
<td>213</td>
<td>351</td>
<td>392</td>
<td>220</td>
<td>608</td>
<td>567</td>
</tr>
<tr>
<td>NEx</td>
<td>468</td>
<td>231</td>
<td>335</td>
<td>124</td>
<td>665</td>
<td>408</td>
</tr>
</tbody>
</table>
Table A.8. CA output for the biplot in figure 3 (columns), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>18</td>
<td>5</td>
<td>707</td>
<td>5</td>
<td>293</td>
<td>11</td>
</tr>
<tr>
<td>SUM</td>
<td>42</td>
<td>198</td>
<td>847</td>
<td>197</td>
<td>153</td>
<td>203</td>
</tr>
<tr>
<td>Rst</td>
<td>36</td>
<td>53</td>
<td>775</td>
<td>48</td>
<td>225</td>
<td>79</td>
</tr>
<tr>
<td>CR</td>
<td>90</td>
<td>181</td>
<td>1000</td>
<td>213</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ovr</td>
<td>96</td>
<td>427</td>
<td>906</td>
<td>454</td>
<td>94</td>
<td>267</td>
</tr>
<tr>
<td>D</td>
<td>506</td>
<td>87</td>
<td>819</td>
<td>84</td>
<td>181</td>
<td>105</td>
</tr>
<tr>
<td>ACC</td>
<td>211</td>
<td>50</td>
<td>4</td>
<td>0</td>
<td>996</td>
<td>333</td>
</tr>
</tbody>
</table>

Table A.9. CA output for the biplot in figure 3 (rows), giving relative frequencies (‰), inertia, correlation with the x- and y-axes, and contribution to the x- and y-axes per factor.

<table>
<thead>
<tr>
<th>Experience</th>
<th>RelFreq</th>
<th>Inertia</th>
<th>X-cor</th>
<th>X-contrib</th>
<th>Y-cor</th>
<th>Y-contrib</th>
</tr>
</thead>
<tbody>
<tr>
<td>LnE</td>
<td>247</td>
<td>471</td>
<td>923</td>
<td>511</td>
<td>77</td>
<td>242</td>
</tr>
<tr>
<td>ShE</td>
<td>373</td>
<td>103</td>
<td>108</td>
<td>13</td>
<td>892</td>
<td>613</td>
</tr>
<tr>
<td>NEEx</td>
<td>380</td>
<td>427</td>
<td>950</td>
<td>476</td>
<td>50</td>
<td>144</td>
</tr>
</tbody>
</table>
Appendix 4

Table A.10 Scale of Intelligibility as Adapted for the Present Study

6. The interpreting is perfectly clear and intelligible. Like ordinary spoken Swedish with few if any stylistic infelicities.

5. Generally clear and intelligible but with minor grammatical or stylistic peculiarities or unusual word choices, though nothing that hampers the understanding.

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.

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.

2. Almost completely unintelligible, although it does not seem completely nonsensical and the addressee may, with great effort, discern some meaning.

1. Totally unintelligible and completely without meaning.

Table A.11. Scale of Informativeness as Adapted for the Present Study

6. Reading the original changes the whole understood meaning. (6 should be given when reading the original completely changes the meaning that the interpreting gave).

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.

4. Reading the original gives some additional information about syntax and words. It can also clarify minor misunderstandings in the interpreting.

3. By correcting one or two meanings, mainly on the word level, the reading of the original gives only a minor difference in meaning compared to the interpreting.

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. No new meaning is added by the original, nor is the addressee’s understanding of the interpreting increased.

0. The original contains less information than the interpreting.
Article 3

The development of expertise – or not

THREE SIMULTANEOUS INTERPRETERS’ DEVELOPMENT OVER TIME

Elisabet Tiselius

Department of Foreign Languages,
University of Bergen
Centre for Research on Bilingualism,
Stockholm University

Abstract

In this article, the performance of three simultaneous interpreters is studied. Excerpts of their performance at interpreting training and today are studied. Furthermore, their present day performance is compared with other experienced interpreters. It is assumed that the three interpreters may be possible experts according to the expertise theory (Ericsson, Charness & Hoffman 2007). However, the results in this study indicate that their development over time may not support the assumption that they are experts in the strictest sense of the theory.
1. Introduction

In research on interpreting, factors such as interpreters’ performance, their growing experience and the possible development of expertise over longer periods of time are rarely investigated intra-individually. Instead, experienced interpreters are typically compared with inexperienced subjects, or students are investigated during their training. It would therefore be of great interest to investigate the development of an individual over longer periods. There are drawbacks with such a study, however, with time constraint being the most prominent. It is rare that researchers in Translation Studies have the time and money to carry on a study over for instance fifteen years. In some exceptional cases, researchers are fortunate to be able to use either their own research data from earlier studies in a longitudinal design, such as Hansen for translation (2008), or the data of a colleague, as in the case of the study reported here. Individuals develop differently and at different speeds (Ericsson et al. 2007), and studying the growing experience intra-individually may illuminate other features of how expertise develops in simultaneous interpreting than when it is studied inter-individually.

Another method of exploring such development is retrospective interviews, where subjects recall what types of exercises they have used, what type of training they have had and so on (Sosniak 2007). Though this provides much information about the expert’s background, in the study reported here there are no protocols or recordings of the performance at earlier stages. Such interviews could be combined with performance excerpts to give a fuller picture of the individual’s evolution.

Following Ericsson and Smith (1991), this report defines experts as outstanding performers who have long-term experience, access to particular expert knowledge when needed, and a special way of practising that includes setting clear goals and openness to feedback.

1.1 Aim and scope

The aim of this study is to investigate how the performances of three simultaneous conference interpreters have developed over time. Does their interpreting improve? Are there any differences in their interpreting performance over time?

The interpreting product of the three interpreters is compared diachronically, through two identical excerpts that were interpreted fifteen years apart, and synchronically with identical interpreting excerpts from peers. The product data are hence compared both intra- and inter-individually and may reveal how experience and expertise develop over time. In the second part of this study, the processes of the three interpreters are investigated synchronically through retrospection compared with their peers.

The starting hypothesis is that the products of the three longitudinal participants will show improvement over time and furthermore that they will be equivalent to their experienced peers.
2. Background

As process and product are closely interlaced in the reproduction of a source language utterance and hence also in the evaluation of that reproduction, both were investigated in the study reported here (see Tiselius and Jenset (2011) for a further discussion on the importance of studying both aspects). The definition of process in the present study follows Englund Dimitrova (2010) as the cognitive activity of producing a target text in one language on the basis of a source text in another language. Product is defined here as the target text, or target speech in this case; because of the nature of the collected data, this article will focus more on this aspect.

Both product and process presumably refine over time. Ericsson says about this development that “although all individuals improve with experience in any domain, some individuals develop faster than others and after years of experience are viewed as experts and masters” (2000: 189). The expertise theory as put forward by Ericsson and Smith also states that these individuals who are viewed as experts and masters all use the same strategies to reach the upper echelons of their field. As noted by for instance Tiselius and Jenset (2011), it is relatively straightforward to study some of the expert criteria, such as years of experience, whereas others may be more elusive: outstanding performance, for example, is a requirement that may be easily mapped in areas with ranking, but for interpreting or other areas without ranking it is the researcher (rather than for example a collective ranking body) who must measure and identify the outstanding performance in comparison with other performances. Ericsson and Smith (1991) outline a three-step method for investigating fields without ranking. First, analyse the investigated domain and the skills necessary for experts in that domain, systematically map the cognitive processes for the specific skill, and then analyse the performance within the framework of general cognitive theory. Second, identify the systematic process and its link to the structure of the task and the behaviour of the performers. And third, demonstrate the superior performance by means of the given cognitive processes, how these processes were acquired, and the structure of the relevant domain knowledge. This three-step model is thorough and all-embracing, and therefore perhaps not possible for only one researcher or one project to cover; as more studies are made on high-profile interpreting and interpreters, however, they contribute to the suggested mapping of the skill.

The data set in the study reported here is small and constitutes but one part of a larger project. The larger project explores how twelve interpreting subjects develop their expertise through their processes and products. Three participants make up the longitudinal data, with the nine other interpreting subjects representing three different levels of expertise and constituting a cross-sectional data. The amount of longitudinal data is minuscule and insufficient for properly investigating the development of expertise from a longitudinal perspective. However, the possibility of using existing, unique longitudinal data seemed too important to pass up. The longitudinal project was therefore incorporated in the cross-sectional project with the aim of drawing conclusions from one set of data in the light of the other.
Quality is inevitably a factor when studying product. As in the present case, quality is often studied from an expectancy approach, meaning that quality is defined by the expectations of different recipients, whether researchers, interpreters or clients. Research in interpreting studies has variously assessed quality on the basis of both different groups’ opinions (see for instance Bühler 1986; Kurz 1993; Moser-Mercer 1996; Kurz 2001; Zwischenberger 2010; Collados Aís, Iglesias Fernández & Stévaux 2011) and the product itself (see for instance Gerver 1971; Barik 1975; Gile 1985; Setton & Motta 2007). Whichever gauge is used, the notion of good, better or superior quality can be said to be based on inter-subjectivity, and a challenge is to reach inter-subjective consensus. Though quality would intuitively seem to improve with growing experience and competence, studies vary in their actual findings: Setton and Motta (2007), for instance, support this assumption, while other studies (e.g., Dillinger 1994) show surprisingly little difference between novices and professionals.

From an expertise perspective (cf. Ericsson 2000), the condition of outstanding performance in the field of interpreting is linked to the notion of superior quality. An outstanding interpreting performance is presumably a performance of exceptionally high or superior quality compared with other performances. Studies comparing interpreters with different experience are mostly cross-sectional (Ivanova 1999; Liu 2001; Setton & Motta 2007). In Interpreting Studies, only longitudinal studies on students have been carried out, usually covering parts of the period of their studies.

Outstanding performance is presumably a result of both a high-quality product and an efficient process. Hence, in order to study excellence or expertise in interpreting, it is crucial to measure the quality of the product in some way. Another reason for measuring product in this particular study is that there were no process data from the earlier recording. Measuring the quality of the product is notoriously difficult, however, as many contributors have shown before (e.g., Grbić 2008; Moser-Mercer 2009; Collados Aís et al. 2011). Grbić explores the concept of quality as it has been treated in interpreting literature and interpreting studies. She divides the construct of quality into different dimensions depending on how it is perceived. First, quality can be defined as either an exception (i.e., it is either something exclusive that only a small elite can produce) or as something that complies with high standards. Second, quality can also be defined as perfection, something that can be achieved and maintained by hard work. This definition divides product into right and wrong, where there is a correct interpreting that should be delivered; the other side of quality as perfection is quality as culture, where quality is created by the joint efforts of all the participants in a certain context. Third, quality is defined as something that is fit for a certain purpose, with a premium placed on customer satisfaction and value for money.

Despite this typology, the evaluation of quality remains notoriously difficult, as shown by Collados Aís et al. (2011), since the assessors’ conscious perception of quality may diverge from their unconscious perception of quality. In order to approach quality in a perhaps more consistent way, a more holistic approach to evaluation has been advanced (cf. Lamberger-Felber 2011).
Evaluating quality according to different criteria such as accuracy, fidelity and fluency is akin to language-testing theory’s use of discrete points, that is, that one particular feature will indicate something about the whole product (Oller 1979). However, just as was claimed in language testing (Byram 2004) and also as said above, assessment may need to be less componential in order to reflect the assessors’ perception of quality. In the present study we combined the second and third definitions of quality from Grbić’s typology, that is, quality as perfection and quality as something that is fit for a certain purpose. Although our chosen testing method was holistic, we retained the idea that one interpreting would be more perfect than the other. However, the interpretations were not compared with a given model; rather, they were evaluated according to how well they served their purposes.

None of the studies cited above link process to quality. However, Tiselius and Jenset (2011) showed that the experienced interpreters in their study both received higher scores for quality and had more developed processes at hand to solve processing problems that turned up.

3. Material and Method

3.1 Participants

The participants in the present study were three female conference interpreters who worked on the staff of various European Union institutions. They all had Swedish as L1 (A-language in interpreting terminology) and English as one of their working C-languages. After completing the same interpreting programme in the mid-1990s, they were accredited to the European institutions and succeeded in joining the staff a couple of years later. At the interpreting programme they participated in a research project on interpreting (Williams 1995) where they were recorded several times, both interpreting (simultaneous and consecutive) and free speech; although the recordings were not used at that time, they were preserved. The second round of recordings was carried out when the participants had fifteen years of experience. In the report of this study, the subjects will be labelled in-training interpreters (IT interpreters) when the first recordings of the subjects are reported, and long-experience interpreters with 15 years of experience (LE 15 interpreters) when the current date recordings are reported to.

3.2 Control group

The control group consisted of nine different interpreters at three different levels of experience. This is the cross-sectional material used in an earlier study (Tiselius and Jenset 2011). The levels were:

1) Three highly experienced (25+ years) conference interpreters, one male and two female. This group only comprised freelance interpreters, but all of them also had long-term experience with the European institutions. All three had Swedish as their mother tongue and English as one of their working C-languages, and all three had an interpreting diploma. The experienced control group will be referred to as LE 25 interpreters.
2) Three recent graduate interpreters. This group consisted of one male and two females. They had an interpreting diploma and two years of professional experience. They were accredited to the European institutions. Just as the LE 25 group they had Swedish as their mother tongue and English as one of their working C-languages. This group will be referred to as SE.

3) The third group consisted of three subjects who had neither interpreting training nor experience. They were recruited at an introductory course at the Institute for Interpretation and Translation Studies at Stockholm University. They had English as a strong foreign language, according to their own appreciation. This group will be referred to as NE.

3.3 Speeches
Two speeches were used in the test. The first one is a NATO speech, just over ten minutes long (1,225 words), that is adapted for students. The first time the longitudinal participants interpreted this speech was fifteen years ago at an interpreting programme, with one of their teachers acting as the speaker; the speech was interpreted live in class approximately two months before the exam. The second time the participants interpreted this speech was at present, at a session reserved for interpreting two speeches and doing retrospection and an in-depth, post-interpreting interview; for this second time around, a recording of the speech from the first performance was used.

The second speech, a nine-minute-long EU speech (1,093 words), was interpreted both by the longitudinal participants and the cross-sectional control group. The speech was originally held at the European Parliament’s plenary session and has been adapted for the purpose of the present study: the pace and accent of the original were adjusted, some figures and challenging lines of reasoning were added, and the speech was re-recorded with an English speaker with Received Pronunciation. The EU speech was only interpreted once, on the same occasion as the second interpreting of the NATO speech.

3.4 Rating
The interpretings were transcribed and broken down into units according to intonation and core idea, with each unit consisting of an idea or logical whole. The units were divided into rating files, which each consisted of separate units from each interpreter randomly put together into a file. This means that the files did not represent the interpreting in chronological order, and each file also contained units taken from all participants. Each rating file was then rated by two different groups of raters, a student group (non-interpreters) and an interpreter group: each unit was thus rated by four different raters. The raters rated the units using an adapted version of Carroll’s scales for intelligibility and informativeness (1966; for an account of adapting and using Carroll’s scales for interpreting, see Tiselius 2009). The scales are a Likert-type scale from one to six, giving a holistic assessment of the interpreting. Intelligibility should be understood as understandable and idiomatic Swedish, and informativeness as the least possible difference in meaning between the original and the interpreting. The scales can be found in appendix 2.
The ratings were done on transcriptions and not on sound files for several reasons. First, because the rater group included interpreters and because the Swedish conference interpreting community is small, the aim was to avoid bias caused by recognition. Second, voice is one of the interpreters’ principal tools, and it can blur the raters’ perception of the interpreting, as has been discussed by Rennert (2010) and Collados Aís et al. 2011. Collados Aís and her team have also shown that prosody, speed and accent affects (even unconsciously) how the interpreting is evaluated. All vocal aspects are thus confounding variables that are difficult to control. In order to test the hypothesis, the sought-after improvement was delimited to the ability to transfer a message from one language to another, and we therefore decided to use transcripts. Furthermore, the adapted scales were tested both on sound files and on transcripts (Tiselius 2009). These tests showed that the assessments done with sound files received slightly worse ratings, though not statistically significant. There was thus no significant difference between the ratings from sound files and the ratings from transcripts.

Assessments were first tested on a cross-sectional data set of nine interpreters with different levels of experience (inexperienced (NE), somewhat experienced (SE), experienced (LE25)). The results from this cross-sectional study showed that experienced interpreters receive better ratings than interpreters with less experience. The difference is statistically significant for informativeness between experienced interpreters and somewhat experienced interpreters, and between the inexperienced group and the two other interpreter groups (for a full report, see Tiselius & Jenset 2011). The control group described above consists of the highly experienced interpreters from this cross-sectional study. In conclusion, the method was precise when it came to making a clear distinction between the groups, and it was assumed that it should be apt for discerning differences between the longitudinal participants’ interpreting performances.

3.5 Retrospection

After the second (or present-day) interpreting round, the three longitudinal participants did retrospection of the two speeches immediately after each interpreting and with a transcript of the original speech as cue. The participants went through the transcript and recalled aloud their interpreting, problems and strategies. The retrospections were recorded, transcribed and then analysed using Ivanova’s (1999) three categories of processing problems, monitoring and strategies. Processing problems are problems that arise while interpreting and that cause automatized processes to falter, such as problems with perception, speed and language retrieval; monitoring occurs when the interpreter reflects on the process or production, for instance an evaluation of the speaker: finally, strategies are solutions of the processing problems and can be anything from word or sentence deletion to generalization or the use of aids. (For a discussion on using Ivanova’s categories, see Englund Dimitrova & Tiselius 2009; the categories with explanations and examples can be found in Appendix 1). The earlier cross-sectional study (Tiselius & Jenset 2009) showed that experienced interpreters encounter fewer processing problems and have access to more strategies to solve the problems they encounter, and that this tendency is stronger with longer experience.
Retrospection was used in the study presented here in order to compare the longitudinal participants in the present study with the experienced control group in the cross-sectional data. However, a technical mishap led to only two of the retrospections being recorded, which means that the comparison reported below is only tentative.

4. Results

This section reports the results from the two analyses of the interpretings, first the assessments and then the tentative results from the retrospection. The results are reported in the different sections below, before the results are discussed under section 5.

4.1 Assessment of the interpreting

The interpretings were assessed by both interpreters (n=12) and non-interpreters (n=12) using the adapted scales. Table 1 shows that for the longitudinal group there is no significant difference between the three participants as novices and as experienced interpreters. In raw numbers, two of the three have improved, while both groups of assessors give the third member (Filippa) a worse score as an experienced interpreter than as a novice; it should be noted that for intelligibility the highest score is six, whereas for informativeness the highest score is one. The NATO speech was rated by both interpreter and non-interpreter raters; as the inter-rater variability was conspicuously high for the interpreter raters (0.6), the rating was done once again with non-interpreter (0.3). The second time the inter-rater variability was good, but as can be seen from table 1, the results are similar.

| Table 1. Intelligibility and informativeness: NATO speech, comparison of IT-interpreters vs. LE 15 interpreters as rated by non-interpreter raters and interpreter raters. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | Intelligibility | Informativeness |
|                 | non-interpreters| non-interpreters| Interpreters    | Interpreters    |
|                 | n=12            | n=12            | n=12            | n=12            |
|                 | Scale 1–6, 6=highest| Scale 1–6, 6=highest| Scale 1–6, 6=highest| Scale 1–6, 6=highest |
| Gabriella, IT   | 4.7             | 3.6             | 3.6             | 4.0             |
| Gabriella, LE 15 | 4.9             | 3.2             | 4.5             | 3.8             |
| Ingrid, IT      | 4.3             | 3.4             | 3.9             | 3.9             |
| Ingrid, LE 15   | 4.4             | 3.4             | 4.0             | 3.5             |
| Filippa, IT     | 4.9             | 3.1             | 4.3             | 3.2             |
| Filippa, LE 15  | 4.7             | 3.2             | 3.8             | 3.5             |

Table 2 shows the ratings of the EU speech. The EU speech was only rated by non-interpreter raters, as students were more easily accessible at the time of the rating; an earlier study (Tiselius 2009) had shown that the difference in rating between non-interpreter raters and interpreter raters was not significant, so we decided to use only non-interpreter raters. The ratings done by students, reported
in table 2, are quite similar to the figures in table 1, which together indicate the longitudinal participants’ consistent performance.

*Table 2. Intelligibility and informativeness: EU speech, non-interpreter rating only.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>Intelligibility</th>
<th>Informativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriella LE 15</td>
<td>4.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Ingrid LE 15</td>
<td>4.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Filippa LE 15</td>
<td>4.6</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Table 3 gives the assessment scores of the longitudinal and the cross-sectional data sets. As can be seen here, the experienced cross-sectional interpreters receive better scores than the longitudinal participants. Again it should be noted that six is the highest score for intelligibility while one is the highest score for informativeness.

*Table 3. Intelligibility and informativeness: cross-sectional and longitudinal data sets.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>Intelligibility</th>
<th>Informativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE</td>
<td>3.5</td>
<td>4.8</td>
</tr>
<tr>
<td>SE</td>
<td>5.1</td>
<td>3.3</td>
</tr>
<tr>
<td>LE 25</td>
<td>5.3</td>
<td>2.4</td>
</tr>
<tr>
<td>Gabriella LE 15</td>
<td>4.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Ingrid LE 15</td>
<td>4.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Filippa LE 15</td>
<td>4.6</td>
<td>3.7</td>
</tr>
</tbody>
</table>

4.2 Quantitative data: interpretngs and retrospection

Tables 4 and 5 provide quantitative data on the two interpretngs and the retrospections. Both the number of words and the time correlate with the number of words of the original, while the time, number of words and number of reports per 100 words vary between the interpreters in the retrospection.

*Table 4. Length of interpreting, length of retrospection: NATO speech, IT and LE 15.*

<table>
<thead>
<tr>
<th>Participant</th>
<th>No. of words IT LE 15 (orig. 1225)</th>
<th>Time of int. IT LE 15 (orig. 11:10)</th>
<th>Time for retro</th>
<th>No. of words in retro</th>
<th>No. of reports per 100 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriella</td>
<td>997</td>
<td>00:10:52</td>
<td></td>
<td>1 960</td>
<td>1.3</td>
</tr>
<tr>
<td>Ingrid</td>
<td>1 127</td>
<td>00:11:05</td>
<td></td>
<td>3 425</td>
<td>2.0</td>
</tr>
<tr>
<td>Filippa</td>
<td>1 031</td>
<td>00:10:57</td>
<td></td>
<td>1 698</td>
<td>1.1</td>
</tr>
<tr>
<td>Mean</td>
<td>1 052</td>
<td>00:11:03</td>
<td>00:18:16</td>
<td>2 361</td>
<td>1.5</td>
</tr>
</tbody>
</table>

In table 4 it can be observed that the number of words in the interpreting increases slightly with experience. The interpretngs took slightly less time than the original, except in one case (Ingrid as IT).

Table 5 shows the quantitative data of the interpreters in the long-term data (LE 15 and the experienced interpreters in the cross-sectional data (Bettina, Malin and Folke, LE 25). The interpretngs range from 1,015 to 1,177 words and from 9:18 to 9:34 minutes, with Filippa using more words than the others in both the NATO and EU speeches. With regard to retrospection, the time and words vary greatly between the different interpreters of both groups.
Table 5. Length of interpreting, length of retrospection: EU speech, long-term (LE 15) and cross-sectional (LE 25) interpreters.

<table>
<thead>
<tr>
<th>Participant</th>
<th>No. of words interpreting (original 1 093)</th>
<th>Time of interpreting (original 00:09:35)</th>
<th>Time for retrospection</th>
<th>No. of words in retrospection</th>
<th>No. of reports per 100 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabriella</td>
<td>1 148</td>
<td>00:09:28</td>
<td>00:25:59</td>
<td>3 691</td>
<td>1.6</td>
</tr>
<tr>
<td>Ingrid</td>
<td>1 015</td>
<td>00:09:26</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Filippa</td>
<td>1 177</td>
<td>00:09:18</td>
<td>00:15:02</td>
<td>1 888</td>
<td>1.8</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>1 162</strong></td>
<td><strong>00:09:24</strong></td>
<td><strong>00:20:30</strong></td>
<td><strong>2 789</strong></td>
<td><strong>1.7</strong></td>
</tr>
<tr>
<td>Bettina</td>
<td>1 116</td>
<td>00:09:34</td>
<td>00:33:17</td>
<td>2 738</td>
<td>1.9</td>
</tr>
<tr>
<td>Malin</td>
<td>1 078</td>
<td>00:09:29</td>
<td>00:09:53</td>
<td>1 380</td>
<td>3.5</td>
</tr>
<tr>
<td>Folke</td>
<td>1 044</td>
<td>00:09:29</td>
<td>00:13:45</td>
<td>1 177</td>
<td>2.5</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>1 079</strong></td>
<td><strong>00:09:31</strong></td>
<td><strong>00:19:08</strong></td>
<td><strong>1 765</strong></td>
<td><strong>2.6</strong></td>
</tr>
</tbody>
</table>

4.3 Processing problems, monitoring and strategies

Tables 6 and 7 show the processing problems reported in retrospection of the NATO and EU speeches. Filippa has the lowest number of reports of processing problems for both retrospections. In table 6, Gabriella and Ingrid are similar both in numbers of reports and in the distribution of the reports. For the NATO interprettings (table 6), most reports are given for comprehension/text integration, that is, background knowledge.

Table 6. Processing problems and categories: NATO speech, long-term participants.

<table>
<thead>
<tr>
<th>Comprehension</th>
<th>Gabriella LE 15</th>
<th>Ingrid LE 15</th>
<th>Filippa LE 15</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3 (7.1%)</td>
</tr>
<tr>
<td>Lexical access in SL</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1 (2.4%)</td>
</tr>
<tr>
<td>Syntactic processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text integration</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>10 (23.8%)</td>
</tr>
<tr>
<td>Text comprehension</td>
<td></td>
<td></td>
<td></td>
<td>5 (11.9%)</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TL retrieval</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6 (14.3%)</td>
</tr>
<tr>
<td>Equivalent</td>
<td>8</td>
<td>5</td>
<td>13</td>
<td>13 (31.0%)</td>
</tr>
<tr>
<td>Simultaneity of tasks</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3 (7.1%)</td>
</tr>
<tr>
<td>SL,TL</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TL delays</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1 (2.4%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
<td><strong>17</strong></td>
<td><strong>5</strong></td>
<td><strong>42</strong></td>
</tr>
</tbody>
</table>

Table 7 gives the processing problems for the two participants in the long-term data set and the processing problems for the three experienced interpreters in the cross-sectional control group (LE 25, i.e. Bettina, Malin and Folke). Unfortunately, the third retrospection could not be used because of a technical mishap. The two informants in the longitudinal data set mainly report on problems with target language retrieval, whereas the problems reported by the cross-sectional informants are more evenly spread out over the different categories.
Table 7. Processing problems and categories: EU speech, LE 15 and LE 25.

<table>
<thead>
<tr>
<th></th>
<th>Gabriella LE 15</th>
<th>Filippa LE 15</th>
<th>Total longitudinal</th>
<th>Bettina LE 25</th>
<th>Folke LE 25</th>
<th>Malin LE 25</th>
<th>Total cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>8 (21%)</td>
</tr>
<tr>
<td>Lexical acc. in SL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Syntactic processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Text integ.</td>
<td>6</td>
<td>1</td>
<td>7 (20%)</td>
<td>3</td>
<td>2</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Text comprehens</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Translation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>TL retrieval</td>
<td>8</td>
<td>5</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Equivalent</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Simultaneity of tasks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>SL,TL</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>TL delays</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>10</td>
<td>35</td>
<td>23</td>
<td>9</td>
<td>6</td>
<td>38</td>
</tr>
</tbody>
</table>

Tables 8 and 9 show the reported monitoring for both speeches. Table 8 shows that neither Gabriella nor Filippa have many reported instances of monitoring, while Ingrid reports many more.

Table 8. Reported monitoring categories: NATO speech, LE 15 participants.

<table>
<thead>
<tr>
<th></th>
<th>Gabriella LE 15</th>
<th>Ingrid LE 15</th>
<th>Filippa LE 15</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Inner speech</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Time</td>
<td>3</td>
<td>2</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Monitoring</td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>Internal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>commentary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Id</td>
<td>7</td>
<td>1</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>21</td>
<td>9</td>
<td>36</td>
</tr>
</tbody>
</table>

When compared with the cross-sectional control group (Bettina, Folke and Malin), Filippa still stands out with all her monitoring reports in mood, as seen in table 9. Folke reports more monitoring than any of the other informants.
Tables 10 and 11 show the strategy use reported by the informants. The informants only rarely report instances of creative interpreting. Ingrid reports the highest level of strategy use and has the most instances of deletion and acceptability (lowering of standards).

Table 10. Strategies: NATO speech, LE 15.

<table>
<thead>
<tr>
<th>Strategy Code</th>
<th>Gabriella LE 15</th>
<th>Ingrid LE 15</th>
<th>Filippa LE 15</th>
<th>Total Longitudinal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Restructuring</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Creative Interpreting</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Overgeneralization</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Deletion</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Explicitation</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>28</td>
<td>5</td>
<td>39</td>
</tr>
</tbody>
</table>

In table 11 the reports from the two groups of informants support each other. The longitudinal group has considerably fewer instances of deletion than the other group, though Gabriella reports many instances of overgeneralization.

Table 11. Strategies: EU speech, LE 15 and LE 25 interpreters.

<table>
<thead>
<tr>
<th>Strategy Code</th>
<th>Gabriella LE 15</th>
<th>Filippa LE 15</th>
<th>Total Longitudinal</th>
<th>Bettina LE 25</th>
<th>Folke LE 25</th>
<th>Malin LE 25</th>
<th>Total Cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Summary</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Restructuring</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Creative</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Overgeneralization</td>
<td>10</td>
<td>2</td>
<td>12</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Deletion</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>Explicitation</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Acc.</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>9</td>
<td>31</td>
<td>12</td>
<td>17</td>
<td>10</td>
<td>39</td>
</tr>
</tbody>
</table>
which is also the speech where all the interpreters were provided with an information sheet featuring names and key terms. There is a wider variety of macrostrategies reported for the EU speech as compared to the NATO speech.

Table 12. Macrostrategies: NATO speech, LE 15 interpreters.

<table>
<thead>
<tr>
<th>Macrostrategy</th>
<th>Gabriella</th>
<th>Ingrid</th>
<th>Filippa</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pre-act</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Text Technique</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5</strong></td>
<td><strong>3</strong></td>
<td><strong>1</strong></td>
<td><strong>9</strong></td>
</tr>
</tbody>
</table>

Table 13. Macrostrategies: EU speech, long-term and cross-sectional interpreters.

<table>
<thead>
<tr>
<th>Macrostrategy</th>
<th>Gabriella LE 15</th>
<th>Filippa LE 15</th>
<th>Total longitudinal</th>
<th>Bettina LE 25</th>
<th>Folke LE 25</th>
<th>Malin LE 25</th>
<th>Total cross-sectional</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pre-act</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Text Technique</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Aid</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3</strong></td>
<td><strong>5</strong></td>
<td><strong>8</strong></td>
<td><strong>4</strong></td>
<td><strong>1</strong></td>
<td><strong>3</strong></td>
<td><strong>8</strong></td>
</tr>
</tbody>
</table>

5. Discussion

The study reported here examined the products and processes of three experienced interpreters, as compared with both their performance fifteen years earlier in interpreting training and with three other highly experienced interpreters (each with twenty-five or more years of experience). As said above, the data set is small, and the results reported here pertain to only three different cases. Nonetheless, these three cases show similar results, and the data set is unique as it spans over fifteen years.

The results of the product assessment of the interprettings yielded little difference between the performance at the training programme and the performance fifteen years later (tables 1 and 2 above). For one longitudinal participant, the scores were worse when interpreting as a professional than when interpreting as a student. There may be several reasons for this. First of all, the instruments may not be sensitive or accurate enough to correctly measure differences. However, the instruments were tested on several occasions prior to the present study, both on different interpreter groups and on different rating groups (see Tiselius 2009; Tiselius & Jenset 2011). Inter-rater reliability is an important issue in testing, as discussed by for instance Lamberger-Felber (1997). In the present study, the inter-rater variability for the interpreter raters was high in a Pearson’s r-test, as discussed in section 4.1. The assessment was repeated with student raters, whose inter-rater variability was low; the assessment results were nevertheless similar, as can be seen in table 1. The assessment results may also have varied because of individual aspects. Since it is the same individual who is rated at different points
of her career, the intra-individual difference may not be distinctly perceived by the rater. The interpreter acquires a certain style during interpreting training, and that style is then kept throughout the interpreter’s career. The reasons for Filippa’s low scores may be also be that her interpreting style is not favourable in this type of rating. Another reason for the slight difference may be that voice was excluded from the design. As was discussed above, however, voice is a highly confounding variable.

There is also a difference between the scores of the longitudinal participants and the control group. When it comes to intelligibility and informativeness (cf. table 3 above), the longitudinal participants are in fact closer to the somewhat experienced level, rather than the experienced level, though this should be interpreted in light of the rating design. In the control group, rating units from all experience levels were mixed together into one assessment file. This means that the experienced interpreters were probably rated in light of the somewhat experienced interpreters and vice versa, whereas the longitudinal participants’ rating units were also mixed but only represented a fairly homogeneous group of experienced interpreters. This design may lead to the experienced interpreters in the control group receiving a more favourable rating.

The process data confirmed the findings of earlier studies (e.g. Ivanova 1999; Tiselius & Jenset 2011), in that the experienced longitudinal participants also encounter few processing problems and have more diverse strategies at hand for solving processing problems. This supports the assumption that the longitudinal participants are experienced, high-level interpreters, although their assessment scores were lower. It can be assumed that their performance is fairly stable.

Looking at the time and the number of words in retrospection and the number of reports per 100 words, we see that the retrospection types vary (see tables 4 and 5 above). For the NATO speech, which was done first, Gabriella and Filippa have a similar number of words in retrospection as well as a similar number of reports per 100 words. For the EU speech, the second retrospection, Gabriella has a considerably higher number of words than Filippa, yet their reports per 100 words do not vary much. Ingrid, whose retrospection for the EU speech could not be used for technical reasons, has both more words in her NATO retrospection and a higher number of reports per 100 words. The interpretations are close to the original in regard to time, probably because of the norm that interpreting should end as soon as possible after the speaker ends.

When it comes to monitoring, it seems that just as for retrospection there are differences that are probably due to different personalities, interpreting styles and retrospection styles. For the strategies, creative interpreting (i.e., guessing) is the least represented category for both the longitudinal participants and the control group; perhaps it can be assumed that these highly experienced interpreters do not have to guess that much, as they can use other strategies to create a consistent message.

All but one interpreter from both groups used aids (cf. table 13), not least because they all had access to a support sheet for the EU speech. All the interpreters also used a greater variety of macrostrategies when interpreting the
EU speech, perhaps because of a greater knowledge of EU matters or again because of the information sheet.

As shown in Hill and Schneider’s (2007) overview of different neurological experiments in expertise, the brain changes with expertise, and changes in connecting strength and neuron size are present in adult subjects too. Ericsson also points out in the same volume (2007) that experts remain in what he calls the cognitive/associative phases through “increasingly complex mental representations to attain higher levels of control of their performance”, whereas some experts at some point in their career stop doing that, which “results in premature automation of their performance” (2007: 685). The longitudinal participants in the present study have reached a very high level in their field. It is possible they had already reached these high levels when they completed their final exam and were admitted to the European institutions. As they started working for the European institutions, they surely refined their technique and gained experience, but they did not have any incentive to refine their interpreting skill, as that does not lead to higher remuneration or specific distinctions; rather, compensation is given for adding languages or taking on administrative tasks. It is therefore possible that these interpreters, although highly skilful and experienced, reached this premature automation of their performance at some point after interpreting training, as shown in this case in the assessment scores.

6. Conclusion

The study reported here sought to study how the performance of three simultaneous conference interpreters had developed over time and whether their interpreting had improved. Their product was assessed on the basis of three interprettings: a NATO speech from interpreting training, the same NATO speech interpreted today and an EU speech interpreted today. Their product was compared both intra- and inter-individually, as well as with a control group of three experienced interpreters. Moreover, their process was studied through retrospection, and their retrospection was compared with three other experienced interpreters. The starting hypothesis was that the products of the three longitudinal participants would show improvement over time, and furthermore that they would be equivalent to their experienced peers. The first part of the hypothesis was not supported, while the second part was.

The results showed that although their processes seem comparable to other experienced interpreters, their product had not developed significantly over time. This may be because of methodological imprecision or what Ericsson (2007) calls premature automation, which may cause highly experienced performers to stagnate prematurely.

Since the study reported here is a small case study, it is difficult to make generalized conclusions as to interpreters’ development of expertise. It should however be stressed that the longitudinal participants in the present study all showed superficial signs of expertise, such as long-term experience, a high level of professional achievement and peer appreciation. Further research into intra-individual development over time is needed in order to confirm or reject these
tentative results. Moreover, as the methodology used in this study seems to be stable, it could preferably be used again with more participants.
References


## Appendix 1. Ivanova’s (1999) categories

*Table 14. Processing problems. Classification according to Ivanova (1999).*

<table>
<thead>
<tr>
<th>Processing problems (PP)</th>
<th>Examples from interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception (P) Problems with hearing</td>
<td>I didn’t have time to catch that it was tobacco-related deaths. (8/Amadeus)</td>
</tr>
<tr>
<td>Lexical access in SL (L) Failure to access meaning of an SL chunk, which has been identified as familiar</td>
<td>I said TB, but I’m not sure what it means. (8/Amadeus)</td>
</tr>
<tr>
<td>Syntactic processing (Syn) Failure to recognize syntax patterns</td>
<td>It’s difficult when they start by saying “which”, which, like you start having those subordinate phrases and so on, that is difficult. (15/Amadeus)</td>
</tr>
<tr>
<td>Text integration (TC/integ/) Difficulties in constructing a coherent representation for SL chunks</td>
<td>It’s like you understand what it’s all about, but you miss certain details. (21–22/Amadeus)</td>
</tr>
<tr>
<td>Comprehension (C) Comprehension difficulties owing to lack of background knowledge</td>
<td>Yes, “community” – again I became confused over which community. (47/Amadeus)</td>
</tr>
<tr>
<td>Text comprehension (TC/bgkn) Problems in rendering an SL chunk in TL</td>
<td>I got stuck there, that is, “I must here compliment…”, you know, I thought, “Oh, how do you say that?” (10/Amadeus)</td>
</tr>
<tr>
<td>TL retrieval (TLr) Problems in selecting an appropriate equivalent when there is a choice</td>
<td>I was thinking about whatever that is called in Swedish, I considered “guidelines” or “descriptions”. (37/Amadeus)</td>
</tr>
<tr>
<td>Equivalent (Eqv) Problems owing to high SL input rate in relation to interpreter’s own output rate</td>
<td>Then there were too many words, sort of, like terms, like “tobacco advertising” and “sponsorship”, so I don’t know what I said. (45/Amadeus)</td>
</tr>
<tr>
<td>Simultaneity of tasks (Sim) TL delays (TL delays) Delays in TL product owing to translation</td>
<td>You know, I heard this but I didn’t get it out, everything that is written here. It was really difficult to, you know, use these two processes at the same time, to listen and understand and then speak, it was like a catch or an obstacle. (10/Amadeus)</td>
</tr>
<tr>
<td>Monitoring (M/)</td>
<td>Examples from interpreters</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Translation (tr)</td>
<td>Ascertaining accuracy of translation at the conceptual level against an ST representation</td>
</tr>
<tr>
<td>Inner speech monitoring (insp)</td>
<td>Verification of the TL message against TL rules prior to articulation</td>
</tr>
<tr>
<td>Time (tm)</td>
<td>Awareness of the timing relative to the TL production</td>
</tr>
<tr>
<td>Internal commentary (int. comm.)</td>
<td>Affective commentary to ST/ST producer</td>
</tr>
<tr>
<td>Mood</td>
<td>Emotive self-evaluation of performance</td>
</tr>
<tr>
<td>Id</td>
<td>Non-analysed problems</td>
</tr>
</tbody>
</table>

**Table 15.2. Monitoring observations (Ivanova 1999).**

- **Translation (tr):** Well, I thought I was almost saying “TBC”, but then I thought, no, he only said TB. (8/Amadeus)
- **Inner speech monitoring (insp):** And it’s the same syntax, so it is easy, isn’t it, to just transfer it directly to Swedish, so that becomes easy. (45/Lisa)
- **Time (tm):** There were many such details that you got stuck on, but it’s like I have taken them all in, all those words in between, but it was so fast. (4/Amadeus)
- **Internal commentary (int. comm.):** I thought it was funny with “gaps” because I just felt that there were gaps all over my own interpretation. (5/Amadeus)
- **Mood:** Oh well, it started to feel good here when he was beginning to conclude. (45/Amadeus)
- **Id:** Yes, this was also too complicated. (15/Amadeus)
### Table 16. Strategies (Ivanova 1999).

<table>
<thead>
<tr>
<th>Strategy code (SC)</th>
<th>Brief description</th>
<th>Examples from interpreters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection (S)</td>
<td>Selection of one SL chunk for further processing because it is more informationally or pragmatically salient</td>
<td>I think I only said the date and nothing else. (34/Kristina)</td>
</tr>
<tr>
<td>Summarization (Sum)</td>
<td>Rendering the gist of an SL segment</td>
<td>The only thing I got was that I took a quick glance and saw “Court of Justice”. (46/Kristina)</td>
</tr>
<tr>
<td>Restructuring (Rest)</td>
<td>Changing the original syntactic structure of an SL segment (usually by transposing clauses or segments within the clause), in order to improve the expression in TL, or anticipation of problems</td>
<td>Here I did not hear the Greeks, but I heard the Germans and that made everything clear... so I had to re-arrange the structure. (Ivanova, N16)</td>
</tr>
<tr>
<td>Creative Interpretation (Cr)</td>
<td>Compensating for missing information by guessing on the basis of previous knowledge</td>
<td>But you miss some details, so sometimes you take a guess and maybe you render it incorrectly but you want to give it some context anyway. (21/Amadeus)</td>
</tr>
<tr>
<td>Overgeneralization (Overgen)</td>
<td>Rendering an aspect of ST (e.g., a lexeme) by selecting a more abstract and hence less committed representation in TL</td>
<td>So I probably did some simpler version of it. (21/Amadeus)</td>
</tr>
<tr>
<td>Deletion (D)</td>
<td>Omitting SL chunks because of time constraints without reference to their semantic or pragmatic role</td>
<td>It just disappeared completely. (1/Amadeus)</td>
</tr>
<tr>
<td>Explication (Expl.)</td>
<td>Explicitly expressing information implied in the ST</td>
<td>I added “associated members”, although I did not hear it then, because it seemed logical that if they are negotiating with them, they have to be associated members (Ivanova, N11)</td>
</tr>
<tr>
<td>Compromise (Acc)</td>
<td>Lower the acceptability standards for a TL production in order to minimize processing costs</td>
<td>But then I thought, no, he only said “TB”. (8/Amadeus)</td>
</tr>
<tr>
<td>Macrostrategy</td>
<td>Example from interpreters</td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Knowledge about the speaker (Sp)</td>
<td><em>It was always at the back of my mind that This person is a native speaker and has a strange usage.</em> (Ivanova, E4, p. 184)</td>
<td></td>
</tr>
<tr>
<td>Pre-activation of relevant domain knowledge (pre-act)</td>
<td><em>So I thought this was so terribly important at the very end too, because that is what this is all about, like their attempt to reduce the injuries of smoking.</em> (50/Amadeus)</td>
<td></td>
</tr>
<tr>
<td>General characteristics of the SL text (text)</td>
<td><em>I thought it was good with adverbs and such words, like “importantly” and “recently”, and I know that it was something more, “finally”, it is good because then you have something to sort of hang on to.</em> (24/Amadeus)</td>
<td></td>
</tr>
<tr>
<td>Use of general interpreting/translating technique (technique)</td>
<td><em>I sort of thought, yes, but now I won’t bother about what I don’t have before, just take this sentence, and I tried to pull myself together a little bit then.</em> (37/Kristina)</td>
<td></td>
</tr>
<tr>
<td>Use of interpreting/translating aids (aid)</td>
<td><em>The only thing I got was that I took a quick glance [i.e., at the word list] and saw “Court of Justice”.</em> (46/Kristina)</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2. Carroll’s (1966) scales in Swedish

<table>
<thead>
<tr>
<th>Skala för förståelse (Intelligibility)</th>
<th>Skala för informativitet (Informativeness)</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>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.)</td>
</tr>
<tr>
<td>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, though nothing that hampers the understanding.</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>3. Verkar vara en förståelig mening men är i själva verket mer oförståeligt 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.</td>
<td>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 the word level, the reading of the original gives only a minor difference in meaning compared to the rendition.</td>
</tr>
<tr>
<td>2. I princip helt oförståeligt. Verkar dock inte helt osammankopplade 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.</td>
<td>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 verkligten 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 she really comprehends the meaning intended.</td>
</tr>
<tr>
<td>1. Helt oförståeligt och helt utan mening. Totally unintelligible and completely without meaning.</td>
<td>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.</td>
</tr>
<tr>
<td>0. Originalet innehåller om möjligt mindre information än tolkningen. The original contains less information than the rendition.</td>
<td></td>
</tr>
</tbody>
</table>
Article 4

Expertise without Deliberate Practice?

THE CASE OF SIMULTANEOUS INTERPRETERS

Elisabet Tiselius

Department of Foreign Languages, University of Bergen
Centre for Research on Bilingualism, Stockholm University

Abstract

Deliberate practice (Ericsson 2007) is a type of focused, goal-oriented practice that is part of the process of developing expertise. A less explored area in interpreting research, deliberate practice is a construct that is not easily investigated using an experimental research design.

This article reports on in-depth interviews with three interpreters. By exploring their background, training, views on interpreting, and perceptions of core areas of deliberate practice (such as practice, setting clear goals and being open to feedback), an impression of their practice habits emerges. The article concludes that deliberate practice as defined by Ericsson is not consciously employed by these interpreters. Some of the implications of these findings for the application of expertise theory in interpreting are outlined in the discussion.

1. Introduction

The expertise approach was introduced to interpreting studies in the late 1990s. Several important publications on expertise in interpreting appeared around 2000, such as Ivanova (1999), Ericsson (2000), Moser-Mercer (2000), Moser-Mercer et al. (2000) and Liu (2001). Most research into expertise in interpreting has been done on conference interpreting, in particular simultaneous interpreting. The expertise theory originates from psychology (Ericsson & Smith 1991) and argues that the reason for experts outperforming other performers in their field is a combination of various characteristics. Expertise is thus not just a result of
talent or aptitude, but years of extended practice involving a combination of different tactics for acquiring, developing and maintaining a specific skill. These characteristics (which include but are not restricted to “long experience in the task domain”, “regular outstanding performance”, “access to expert knowledge when needed”, “deliberate practice”, “clear goals” and “openness to feedback”) are common among expert performers regardless of field. The first three characteristics can be observed to a greater or lesser degree by the researcher, whereas the latter three cannot.

The findings reported in this article result from in-depth interviews with three conference interpreters, hereafter referred to as the informants. The aim of the interviews was to investigate their personal and professional backgrounds as well as their views on their profession, preparation, practice and goals. The rationale for doing this was to approach the more elusive concepts of deliberate practice, clear goals and openness to feedback.

2. Background

Deliberate practice is a particular type of practice, summarised by Horn and Masunaga (2007: 601) as “focused, programmatic, carried out over extended periods of time, guided by conscious performance monitoring, evaluated by analyses of level of expertise reached, identification of errors, and procedures directed at eliminating errors.” According to Ericsson, “the core assumption of deliberate practice is that expert performance is acquired gradually and that effective improvement of performance requires the opportunity to find suitable training tasks that the performer can master sequentially” (Ericsson 2007: 692). Deliberate practice is crucial for achieving levels of expertise in a domain. Ericsson et al. (1993: 368) divide any activity into three parts: work, play and deliberate practice. Work is defined as performing in public and often also for remuneration, play is defined as an enjoyable activity without any particular goal and deliberate practice is defined as an activity that includes processes designed to improve the current level of performance. Ericsson et al. also suggest that deliberate practice can be used to discern experts from other performers.

An important part of deliberate practice is having clear goals. The performer must be able to specify intentions, results or outcomes. Research in goal-setting has shown that performers perform better if they are able to specify detailed goals or can break a goal down into different sub-objectives (Zimmerman 2007).

Experts are open to feedback, whether from coaches, trainers, colleagues or the performer’s own results. Being open to feedback helps the performer to evaluate performance, improve performance and set new goals (Horn & Masunaga 2007: 601).

Deliberate practice as described above can materialize during training or education, and also when the performer steps out into the professional world. In expertise theory, a performer is not an expert when he or she graduates from school or a training programme. Budding experts continue to refine their skills by deliberate practice.

Studies of interpreting expertise have typically studied the performance of highly skilled interpreters and compared the features of their performance to that of less experienced interpreters (see for instance Ivanova (1999); Liu (2001); Köpke & Nespoulous (2006); Vik-Tuovinen (2006)). This type of design favours measurable aspects of expertise, such as “outstanding performance” or “access to
expert knowledge when needed”, but is less suitable for studying different aspects of “deliberate practice”, “clear goals” and “openness to feedback”.

Few, if any, studies of expertise in interpreting have used qualitative methods, although researchers in other fields have made use of qualitative methods when studying the expertise theory. For example, Sosniak (2007) reviewed different studies using retrospective interviews (i.e. dealing with events that occurred a long time ago, such as in childhood or adolescence) to study how expertise developed. Deakin et al. (2007) used diaries in studies of time management in practice and its links to expertise. Sosniak reported that habits of deliberate practice were formed during childhood, while Deakin et al. found that experts practise more and with a higher intensity than other performers.

Interpreting is made up of skills and sub-skills. The primary skill is the elusive interpreting skill, and sub-skills include language knowledge (both foreign and mother tongue), general knowledge (popularly referred to in interpreting as “culture générale”), communicative skills (i.e. analysing, speaking, presenting and voice), concentration, memory and the ability to deal with stress. Many more skills can be added to this list. In a literature survey on aptitude testing, Russo (2011: 13) identified three specific areas: a) language knowledge and cognitive skills, such as general mental ability, general and culturally specific knowledge, ideational fluency, verbal and associative fluency and working memory; b) interpreting-related skills that can be acquired, such as simultaneous note-taking and simultaneous transfer; and c) personality traits.

When students acquire these skills, Moser-Mercer says that they “develop flexible understanding of when, where, why, and how to use their declarative and procedural knowledge to solve new problems” (2008: 13).

3. Methodology

3.1 Participants

The three informants in this study were all female who grew up in Sweden with Swedish as their mother tongue. After graduating from the same interpreting programme, they became staff interpreters for various European institutions, where they have been active for the past fifteen years. The interview study was a complement to a larger longitudinal project, and the three participants were recruited on basis of their early recordings as well as their professional success. They were regarded “good interpreters” by their colleagues. They had experience from teaching and peer reviewing of other interpreters. On the basis of their previous experience both on and off task, it was assumed that they would have developed expertise. They were also willing to participate both in new recordings as well as in-interviews, which in turn indicate willingness to expose themselves both to scrutiny and in-depth reflection. The participants were informed of what their participation implied and signed a form of informed consent.

3.2 Procedure for conducting the in-depth interviews

The in-depth interviews lasted between an hour and ninety minutes and were conducted in an unstructured way following a map of topics; Koskinen (2008) used a similar method in her study of translators in the European Union (see Kaijser & Öhlander 1999 for a thorough description of the use of unstructured interviews). These interviews were structured insofar as both parties agreed that an interview was to take place and they set time aside for it. In all other respects they were unstructured in order to be as free as possible. Traditional definitions of an interview are also applicable, for example that an interview is a form of communication where one person recounts something and answers another person’s questions, and the material is recorded in some way (Fägerborg 1999: 55). Quinn-Patton (2002: 342) refers to this type of interview as informal conversational, defining it as the most open-ended approach to interviewing and the type that offers maximum flexibility to “pursue information in whatever direction appears to be appropriate” (2002: 342). Quinn-Patton stresses that unstructured does not mean unfocused and that such interviews should rather be highly focused.

For the purpose of this study, an interview model was developed by means of discussions, mind maps, a pilot interview and pilot focus group interviews. On the basis of early discussions with research colleagues and pilot studies, different areas of interest were identified, the main ones being “deliberate practice”, “clear goals” and “feedback”. Concepts relating to these areas were identified in the preparation phase. The focus group study (Tiselius 2010) showed that expertise concepts like “deliberate practice”, “clear goals” and “openness to feedback” were not clearly perceived by those taking part in that study. These different characteristics of expertise were ranked below concepts like “render a complete interpreting” or “not change the information in the message” (Tiselius 2010: 12–13). From the discussions in the focus groups, it was also clear that the participants did not really understand the concepts. The experiences from this focus group study helped to create a more open interview form with which to investigate the three core areas in question.

3.2.1 Identification of topics and core phenomena

This section presents the topics and core areas that were included in the interview mind map (see figure 1, below) and the reasons for including them. Childhood and teenage dreams and goals were included because studies in expertise show that expert characteristics are present during childhood (see above Sosniak 2007). Learning languages is a sub-skill of interpreting, but strategies for learning languages reveal practice habits, goals and so forth. Experiences from the interpreting programme were included because interpreting skills (e.g. consecutive and simultaneous interpreting) are taught at interpreting programmes, along with different sub-skills such as preparation, practice or terminology work, and habits promoting expertise. Testing, that is, interpreters’

---

13 Deliberate practice was perceived as only practice of the interpreting skill, in the booth in front of a microphone. No consensus was achieved of what a goal would be, and it was concluded in the focus group that the concept was unclear. Openness to feedback was also dismissed as unclear.
attitudes to tests and being tested, reveals their approach to goal setting, practice habits or relations to colleagues. Practice and preparation are logical starting points for discussions about deliberate practice. Colleagues, listeners and clients provide feedback that the participants could be more or less open to.

The above topics and core areas were included in the interview mind map, which then served as the basis for the questions in the more structured yet unmoderated focus group study mentioned above. After being tested in a pilot interview, the mind map was furthermore used as a road map for the interview. Using a mind map rather than already formulated questions entails that the informants are not necessarily asked exactly the same question, because many of the questions are guided by the answers; since the same concepts were covered, however, the questions were more or less the same for all three respondents. For the purpose of this article, the mind map and its concepts are presented in figure 1 below.

Figure 1. Thematic sketch of topics covered in the in-depth interviews. The heart-shaped themes reflect the focal points of the study.

3.2.2 The interview setting

The interviews were conducted at the informants’ workplaces. It should be mentioned that the author of this article is a colleague of the informants. As Fägerborg (1999) points out, the role of the interview leader in an ethnographic in-depth interview is that of a discussion partner, that is, the interview is co-constructed by the interviewer and the interviewee. The implications for such an interview can be that the participants are less honest when discussing difficult topics, such as mistakes, tests or relationships with colleagues. Answers may be formulated with the intent of making an impression on their peer (me) or hiding weaker aspects; these mechanisms may even be unconscious. Moreover, as in all interviews, the informant is aware that the material will be used for a certain purpose and thus has the power to choose what to say or not to say in this situation. All this has to be taken into consideration when analysing the collected data. The informants in this study were candid in their responses, however, and did not shy away from difficult topics. In my experience, if a trustworthy atmosphere is created, honesty and openness will follow.
3.3 Coding and analyses of the interview data

The recorded interviews were fully transcribed and analysed. The analysis took its starting point in the various skills and sub-skills considered to be crucial for interpreting. All the occurrences that had any bearing on the identified skills, topics and core concepts were coded. The skills that were discussed with the interpreters were used as indicators of deliberate practice, goal setting or openness to feedback. For example, even though all the respondents explicitly stated that they did not use deliberate practice, they did give examples of practices such as the following: “It’s normal – if you’re just hanging around waiting, you can always go and listen to your colleagues, and for instance reflect about what makes that interpreter so pleasant to listen to.” Instances like this were classified as supporting the different core areas or skills. In the above cited case, it was classified as supporting deliberate practice, in accordance with Ericsson’s (2000: 214) claim that listening to or studying the performance of highly experienced peers helps to improve your own performance. It should be pointed out that the interpreter’s perception of practice and the construct of deliberate practice as it is explored in this study is not necessarily the same thing. Therefore, the many instances that are classified as deliberate practice by the researcher may not be regarded as practice by the interpreter.

In order to capture other narratives, topics or tendencies that may have arisen in the interviews, the interview transcripts were re-read together with another research colleague.

4. Qualitative Sides of Expertise

This section presents the analyses of the various topics discussed in the interviews. Fictitious names (Filippa, Ingrid and Gabriella) have been used in order to protect the identities of the informants.

4.1 Language learning and language knowledge

Contrary to the common belief that interpreters grow up bilingually, the three informants grew up in monocultural and monolingual environments and did not focus on language learning early on in life, although this had clearly not affected their ability to interpret well. Ingrid was the only one to display an early desire to communicate in other languages, as she tried to learn different languages with dictionaries as her only sources of reference. Ingrid also recalled how her dad used to say that what he remembered from her middle school years and throughout secondary school was “a murmur from my room when I read texts and glossaries aloud”. In contrast, Gabriella was focused on natural sciences, and only decided to study languages when she was in her twenties, after having gained a university degree. Filippa started secondary school by studying the natural sciences, but then switched tracks during secondary school and focused more on languages. The common denominator for the three participants was that when they did decide to learn a language, they focused intently on the language learning task. Ingrid studied an additional fourth language on her own in

14 All the quotations have been translated from Swedish to English by the researcher.
secondary school. Both Filippa and Gabriella went abroad soon after starting their language studies, for longer periods of time to study their chosen language at University.

At the interpreting programme all three informants experienced the need to improve their mother tongue and not “merely” learn foreign languages. In Gabriella’s words, “my focus had been on learning foreign languages and now I suddenly felt that I had to learn Swedish”.

The language profiles of the informants differ with respect to the age at which they began their active work with L2. They started acquiring their L2 past the critical age, in fact for two of them, this is a process which began in early adolescence. However, once they became interested in languages, they pursued their studies with unusual dedication and focus, seeking different opportunities to enrich their knowledge of both foreign and native languages.

4.2 General knowledge

All the informants talked about improving their general knowledge, albeit not in those exact words. Filippa said that when she left secondary school, “the idea was to get a complete, general foundation that I would then be able to do anything with – whatever that might be”. All the informants said that at the interpreting programme they read newspapers, listened to the radio and watched TV in new ways, both in their mother tongue and in their foreign languages. Gabriella added that when she talked about reading in general, it meant that “I read differently than I would do otherwise [i.e. if not an interpreter], it’s not like reading in a deckchair” (meaning that reading much more focused).

The informants were all curious and well informed about world events. They also considered curiosity and general knowledge to be necessary for being a good interpreter. When Filippa talked about general interpreting abilities, she mentioned “a general curiosity and openness, striving to always absorb everything and a genuine desire to understand everything”. When Ingrid talked about what made a good interpreter, she mentioned “intellectual curiosity, general knowledge and fast thinking”.

4.3 Communicative skills

Filippa said that “interpreting is very personal depending on who you are – we all have our personal way of expressing ourselves, and when [we started to work] we were able to listen to experienced interpreters who worked differently, but who were all equally good, and that was very useful”. The informants listened to their colleagues interpreting when they worked together in order to improve their own communication skills, which included good formulations, solutions and terminology use. Ingrid’s statement summed this up well:

I listen because I may have to help out with a word or maybe something else, or maybe even take over, it happens sometimes. Sometimes I listen because it’s a pleasure to listen, and it’s a joy to hear how somebody solves a tricky situation, and I also try to – even if I don’t think that you can just assimilate somebody else’s system – get inspiration for different solutions.

Gabriella emphasized the interpreter’s communicative relationship to clients and listeners in particular. Ingrid and Filippa, on the other hand, stressed the importance of being understood when interpreting no matter who was at the other end of the headphones. Although Gabriella initially stated that she did not
have a relationship with her listeners, she went on to say that she almost had a crush on everyone who made contact with the interpreters, for example by waving or smiling to them in the booth or just saying thank you. While she stressed that interpreters at the European Parliament are primarily there to provide a service, she felt it was a great boost to discover that “our service is used, they listen to us”, or to hear a client say, “Oh, it’s you again, that’s great!”

4.4 Focus

In their responses, all three respondents came across as being focused when young, although in different fields. Filippa had focused on sport and spent most of her youth practising and competing at a high level, at high school she studied natural sciences. Gabriella had specialized in the natural sciences too, and in middle school she had forced her parents to find out how she should prepare for secondary school and university in order to work in this field. Ingrid had a particular interest in learning languages, sometimes with a dictionary as her only support. The common denominator here is not their initial field of interest, but rather the intensity of the interest.

Another striking similarity is that although the informants were determined and had clear goals with regard to sports, hobbies or school results, as children or young adults they had no clear goals or visions about their future, and they had little idea about what to study after leaving secondary school. Even after obtaining a university degree, their future profession was not obvious.

Focus, in this section, has been approached from a macro perspective, goal focus in life. It can also be approached from a micro perspective, meaning the ability to focus on task. At the micro level, as is also indicated in section 3.5, the informants talk about being good at concentrating on the task, in the here and now.

4.5 Coping with stress

Interpreting can be both psychologically and cognitively stressful, and an inability to cope with stress will have a significant impact on one’s interpreting skills. Interestingly, none of the informants talked about particular types of stress management or learning to deal with stress, although all three seem to cope positively with stress.

An area in which coping with stress was discussed was test situations. Interpreting tests are stressful, because the candidate has to interpret one or several unprepared speeches in front of an examination board, often with five or more assessors present. The informants approached tests differently. Filippa said that she had “a very good ability to concentrate and be present in what I do”. Gabriella said that she did what she was told to do, namely, “pass the test”. Ingrid stressed the importance of routines for test preparation and not “over-preparing”. Filippa also said that tests were good because several people listened to the performance and gave the interpreter feedback.

Ingrid also talked about the demands and stresses of the interpreting programme, which according to her “were of a different nature; it felt like you were inside your brain and tampered with it much more [than in traditional university training]”. This intuitive impression of the learning process of interpreting gets support in the results in the brain imaging study of Hervais-Adelman et al. (2011) which found indications of change in the bilingual brain of interpreters.
4.6 The interpreting skill

The question of whether interpreting is an innate or an acquired skill has been discussed by both researchers and interpreters (see for instance Mackintosh 1999). All the informants claimed that the ability to interpret had a certain innateness to it, and they considered their profession to be close to their nature or personality. Ingrid explained this as follows: “And then I believe there is a certain factor X, as there is in all recipes, you can use some of this, this and that, and then there is something, that little extra, which is also needed and which cannot be defined”. To some extent this may have a bearing on how the informants viewed the need for practice.

If they consider the interpreting skill to be innate to a certain extent, they may not need to practise the main skill, so that practising their sub-skills would suffice. However, Ingrid also talked about improving her interpreting skill: “I also believe that to continue to add new languages is also a way to improve. Because I believe that if I master more languages, then I can disconnect from the original languages in some way. That it forces the actual interpreting process to be stronger.” Ingrid made the connection between the sub-skill (language learning) and the main skill (interpreting). During the interview she returned to the skill of interpreting when talking about interpreting programme and how they were taught and how to teach interpreting:

Because I think that this process – and I have to say that I’m not even sure it can be taught, I have not made up my mind yet – but this process – well, I suppose that everyone can develop a certain skill – but what makes it really come to life has probably to do with aptitude. Because [the development] of this process cannot be rushed.

This was not unique to Ingrid, with all three informants talking about “an X factor”, “something innate” or “a particular skill”.

They all said that they practised consecutive interpreting (although more as a tool for language learning rather than actually improving the consecutive skill) when preparing for a test with a new language. Gabriella was the only one who said anything about practising an interpreting skill. She said that “I still do à vistas (interpreting from a written text) when I discover a good text, or feel that I have to hammer in some terms, not every week, but maybe twice a month”.

The finding that the informants did not practise the interpreting skill is supported by Leis’ (2003) conclusions from her questionnaire study on self-assessment and self-evaluation among trained and un-trained Estonian interpreters. Her study showed that trained interpreters prioritized improving sub-skills such as language learning or background knowledge over refining the interpreting skill.

5. Deliberate practice, clear goals and openness to feedback

The informants seemed to have been highly focused from an early age on areas that interested them: sport for Filippa, language learning for Ingrid and science for Gabriella. They all mentioned setting goals and the importance of practice when talking about their childhood activities. Ever since childhood the informants took time to prepare and practise, although none of them explicitly defined this as deliberate. The determination displayed in mastering different
skills since childhood characterized how they now mastered the various skills necessary for interpreting.

With regard to interpreter training, they all mentioned different types of practice, although they did not specifically state that they practised their main skill. Without being taught to do so, and without regarding it as practice, they talked about different types of activities performed regularly under practice-like conditions, such as Filippa’s newspaper reading or Gabriella’s radio listening. But they did not seem to consciously or even unconsciously practise in a way that could be defined as deliberate in terms of Ericsson’s definition. They simply did not engage in activities outside the actual interpreting activity (work in Ericsson’s words) that were solely aimed at improving their interpreting skills (contrast this with how for example athletes, singers, actors or chess players regularly practise, i.e. with time set aside for practise, with a precise goal for the practise session, often with a coach and so forth). When they talked about practice, they all said that they did not practise per se, that is, they did not practise their main skills in consecutive and simultaneous interpreting. But they did all talk about reading plenty of newspapers and listening to the radio, which indicated that they do practise sub-skills.

On the other hand, they talked about how they struggled to improve and how their improvement was rewarding. Ingrid labelled herself as a perfectionist and said that she always tried to improve herself, and that her worst professional nightmare would be to discover that she was working on autopilot. Filippa said that it’s a kick. For me, it’s particularly when I really understand, for instance, a strange line of reasoning, and I manage to sort it out, then I get a huge kick. Both because it’s my job, which is the only important thing really, but also for me personally, when everything falls into place, I’m in harmony, it’s a very physical experience.

Ingrid said:

There are days when I am better, when I strain every nerve, and then it’s very rewarding when I feel that my performance is better. It feels good in my whole body. It’s harmony, it’s more like I create order in the chaos of universe.

Getting a perceived physical reward from producing good interpreting creates a virtuous circle. This feeling of producing high-quality interpreting is self-perpetuating, in that the interpreter is motivated to perform better and spend more time on the task.

In the case of clear goals, the informants all said that the most important goal in every interpreting situation was to understand and be understood. It should be stressed that the goals mentioned here are task goals, i.e. what to achieve while on task, and not training goals, i.e. goals related to structuring practise in order to improve performance. Before the discussion about goals arose, Ingrid repeatedly mentioned that she constantly tried to perform better because she was never totally satisfied and always had a desire to improve her performance. Ingrid also said about goals that “there is no absolute goal, but that is also something that is satisfying, that you will never get there”. Gabriella, who practised by doing an â vista interpreting, set goals like reading most of the *Economist* and similar sub-skill goals. Filippa said that when she started working she did not use all her languages, but broke the work down into different part-time goals, mastering one
language at a time. The goals the informants talked about did not necessarily pertain to improving the interpreting skill but the different sub-skills, such as learning or improving languages, because these are the areas that are publicly rewarded.

For the informants, feedback came from evaluating themselves according to their own standards or from listening to their colleagues’ performances, rather than from receiving comments on their performances from colleagues. This finding aligns with that of Leis (2003), who found that Estonian conference interpreters evaluated themselves according to their own standards learned in interpreting training, rather than from clients’ feedback (in that case possibly a lower standard). In terms of deliberate practice in expertise theory, however, openness to feedback from peers and trainers is a tool that the informants only partially made use of. In their view, listening to highly experienced peers was beneficial for improving their own performance (Ericsson 2000: 214).

6. Discussion and Conclusions

The interviews analysed in this article constitute a case study on deliberate practice. They represent an enquiry into the practises engaged in by three interpreters, which places emphasis on exploring in depth their perception about interpreting, and the process where by they have acquired and perfected their skill. The analysis highlights interesting findings, which emerge from the informants’ stories that align with findings in other studies.

Many superficial indicators suggest that the informants in this study fulfil the criteria of experts as defined by expertise theory, for example that they have long experience and have passed challenging accreditation and qualification exams. But experts are also defined by other qualities, including deliberate practice and the activities linked to such practice. Deliberate practice is not easily or immediately investigated in fields lacking obvious needs or incentives for improving the main skill. For employed interpreters at larger institutions, a personal physical positive reward (cf. the quotes in section 5) may be the only reward available, especially as there is little hope of higher remuneration, prizes or other recognition. Staff interpreters at larger institutions do not get a pay increase for producing better interpretations than their colleagues, there are no prizes for outstanding interpreters or interpretations, and outstanding simultaneous skills do not automatically lead to promotion. Instead, it is additional languages or administrative skills that have the potential to increase a staff interpreter’s salary. Freelance interpreters could theoretically get more jobs if their interpreting skills are outstanding, which may in itself be an incentive for practising the skill. But for freelance interpreters who are accredited to the European institutions and who are placed in the highest quality category, the only criteria that matter for recruitment are geographical distance and number of languages. There is not much incentive here for continued refinement of the interpreting skill. Interpreters cannot be compared with translators in this area, because several different translation awards are available.

This does not mean that interpreters are not interested in improving their performance. On the contrary, the in-depth interviews reported here show that although the respondents had not been taught deliberate practice, they did make use of deliberate practice strategies to improve their sub-skills. They also seem to have made use of these strategies at a young age. But whether this can be defined
as deliberate practice as it has been defined by Ericsson et. al. (2007) is open to discussion, especially as none of the informants participated in activities in order to improve their main interpreting skill.

The in-depth interviews have showed, however, that these interpreters engage in (although unconsciously) deliberate practice strategies. They practice their language skills and strive to enhance their general knowledge, they actively learn from their peers by listening to them. Moreover, they also consider at least some part of the interpreting skill as innate, or dependent on an x-factor. This view of the interpreting skill may have effects on practise, which did not come up in the interviews. Presumably an innate skill would need less practise than an acquired one. However, the fact that the participants engage in so many other practice activities argues, at least partly against that argument.

The narratives that emerged during the interviews formed a uniform pattern. As the informants came from similar backgrounds, were more or less the same age, attended the same interpreting programme, had similar language combinations and the same professional backgrounds, it is fair to assume that they shared the same norms and the same professional habitus. Their stories nevertheless say something about their interpreting expertise. From a superficial perspective they are highly experienced interpreters who have reached the highest levels of the interpreting profession, and are regularly evaluated by their superiors. Nevertheless, they are unable to make more money, win competitions or become famous by improving their interpreting skills. From their narratives it is clear that their goals to perform better, or at a level that was acceptable to them, revolved around their own personal ranking or pride and no one else’s. They were also convinced that the interpreting skill was mostly innate. In other words, there was scant external or internal incentive that could motivate them to engage in deliberate practice with clear goals and regular feedback from colleagues in order to improve their main skill of interpreting.

The above conclusion might not be valid for interpreters who aim towards passing accreditation tests for larger institutions, as they may well have an incentive to improve their interpreting skills. But if this conclusion proves to be true for the crème de la crème of the interpreting community, it will have implications for the application of expertise theory in interpreting. The definition of experts in interpreting research is very varied (see Liu 2008). Findings in this study indicate that experienced interpreters do not engage in deliberate practice the same way as other professions. If this is case, the theoretical framework will need to be adapted both in terms of how an expert is identified and also in terms of how the expertise concept of deliberate practice can be applied to interpreting research.

The findings of this study raise the following questions: Is it possible to be an expert without deliberately practising the main skill? Would it be enough to refine one’s sub-skills? Is expertise theory still applicable to interpreting studies? In order to answer these questions, more studies of simultaneous interpreters’ deliberate practice must be conducted.


