VNU Journal of Science: Foreign Studies, Vol. 32, No. 4 (2016) 12-20

Piloting an Assessment Model of Interpreting Quality

Nguyen Ninh Bac*

Faculty of English Language Teacher Education, VNU University of Languages and International Studies, Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 21 June 2016
Revised 29 November 2016; Accepted 30 November 2016

Abstract: How to assess interpreting quality in conferences remains a question not yet satisfactorily answered. When disputes arise upon interpreters' performance in conferences, the related parties do not have a consistent ground to base their assessment on. This research, completed under the sponsorship of the University of Languages and International Studies (ULIS, VNU) in the VNU research grant No QG.15.35 "Models for English-Vietnamese translation assessment", has piloted Kurz's model in 1989 with eight criteria in assessing simultaneous interpreting quality in three conferences. The findings show that this model allows comprehensive, accurate and objective assessment of interpreting quality. They also help pointing out interpreter's strengths and weaknesses. However, there are certain limitations in the model, especially regarding large scale applicability and the incorporation of external quality factors.

Keywords: Interpreting, quality, assessment, Kurz, model.

1. Introduction

The era of globalization generates an increasing need for exchange between local people and foreigners. In Vietnam, interpreting has become a profession that is ever more important. This is reflected in a large number of international conferences which require interpretation service organized every day in Ha Noi and Ho Chi Minh City - the two hubs of the country. The number of interpreters has also increased to meet this demand.

However, how to assess interpreting quality in conferences remains a question not yet satisfactorily answered. In fact, when disputes arise upon interpreters' performance in conferences, the related parties do not have a consistent ground to base their assessment on. Most of the time, the complaining party only bases on their subjective, arbitrary "feelings" on the interpreter's output. This method of assessment is of course not acceptable to professional interpreters. But these interpreters themselves, in their turns, may not be able to defend their position with convincing arguments [1:768].

While translation has been done for thousands of years, simultaneous interpreting has only appeared since 1927 and become more popular after 1945 [2:30]. That partly explains why there has been intensive research on the quality assessment of translation, "the quality of interpreting services is an issue which confronts interpreters, interpreting trainers, users and

* Tel.: 84-904245158
Email: bacvnu@gmail.com

researchers with considerable problems" [1:768].

This research is part of a larger project (QG.15.35) to recommend a model that is reliable, valid, and feasible in assessing simultaneous interpreting quality for English-Vietnamese language pair. In this research, Kurz model [3:143-148] will be piloted to assess the quality of interpreting at three different conferences.

2. Quality and quality assessment in simultaneous interpreting

According to the European Organization for Quality Control, quality is defined as "the totality of features and characteristics of a product or service that bear on its ability to satisfy a given need" [cited in 5:404]

Marketing experts also claim that customer satisfaction depends not only on the product's/service's performance but also on that customer' expectations. There may be different degrees of satisfaction. The customer is dissatisfied if the product's/service's performance is lower than expectations, is satisfied if it matches, and is highly satisfied if it exceeds his/her expectations [4:553].

From this definition, Kurz [5:405] came up with the following formula on quality:

Quality of service (customer satisfaction) = service quality delivered – service expected

In other words:

Quality = Actual Service – Expected Service

This formula even increases the complication of interpreting quality assessment and proves that interpreting quality is highly subjective [5:405].

Besides user expectations, interpreter quality is also influenced by external factors such as low voice quality, lack of documents for preparation, speakers' speed of delivery, view obstruction from interpreters' booth to projector screen, non-native speaker accent, speakers telling personal stories or highly contextual jokes, strange idioms, etc. However, it is hard to explain these difficulties to those who are not familiar with the interpreting profession [6].

3. Some assessment model of simultaneous interpreting quality

Despite difficulties in assessing interpreting quality, especially simultaneous interpreting, a lot of authors have tried to propose a number of models.

According to Chiaro and Nocella [7:279], "although there is considerable agreement in the literature regarding criteria that are involved in assessing quality in this field, there appears to be little harmony concerning which perspective to take when undertaking research: whether it is best to explore the success of an interpretation from the perspective of the interpreter or from that of the user is a debatable issue."

The development of a model to assess conference interpreting quality started somewhere in the 1980s with efforts led by Bühler [8: 231-235]. She came up with 16 criteria and conducted a survey on members of International Association of Conference Interpreters (AIIC). These criteria range from linguistic factors, such as "sense consistency with original message", "correct grammatical usage", "fluency of delivery", "native accent" to extra-linguistic factors, such as "pleasant voice", "thorough preparation of conference documents", "pleasant appearance", and "positive feedback of delegates". Professional interpreters were asked to rank the importance of these criteria from their own perspective.

In his model, Viezzi [cited in 13:123] included four goals: equivalence, accuracy, appropriateness and usability. Quality is defined as the level of which these four goals are achieved.

Pöchhacker [9:97] came up with a model of quality standards ranging from lexico-semantic core to socio-pragmatic sphere of interaction. He defined good interpreting quality as accurate

rendition of source, adequate target language expression, equivalent intended effect, and more broadly: successful communicative interaction.

Late 2008, Pöchhacker was commissioned on another AIIC member targeted Survey on Quality and Role as part of a larger research project on Quality in Simultaneous Interpreting. His findings share some points in the ranking of quality criteria with the earlier model by Bühler.

Among these models, Viezzi's may provide overall view on interpretation quality. However, model users may have difficulties in quantifying interpretation quality as his criteria are relatively broad. The one by Bühler really established the ground for many researchers later looking into assessing interpretation quality. However, her survey may have problems with reliability and validity as the sample size is very small (47 interpreters). Basing on Bühler's work, Pöchhacker was able to produce a much more reliable model with much larger sample size (704 interpreters). However, both Pöchhacker and Bühler have only looked at interpretation quality from professional interpreter's perspective while it is not yet clear if that can represent the opinion of other important target groups, including the audience.

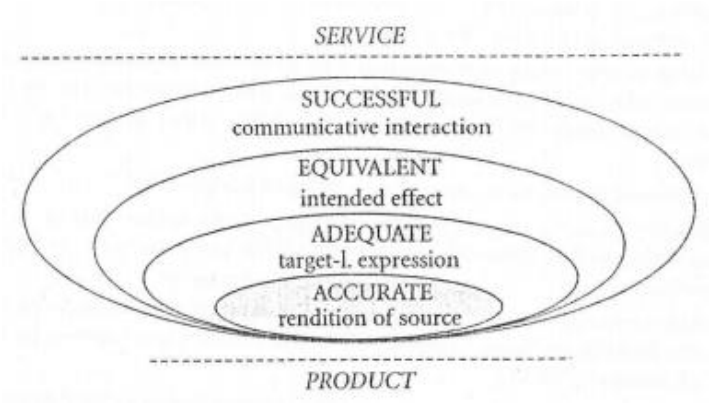

Figure 1. Pöchhacker's model of quality [9:97].

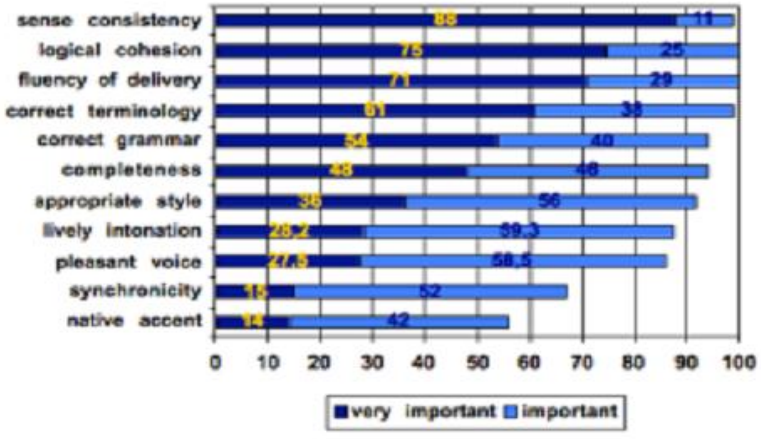

Figure 2. Rating of Quality Criteria, N=704 [10:311].

4. Kurz's model

While Bühler focused on interpreter's perspective, Kurz conducted a survey on the expectation of interpretation service users in 1989.

Relating to the assessment approach from service user perspective, Kalina [11:123] claimed that "the content of the ST can be judged only by listening to it in the original language. If the user listens to the TT, equivalence between ST and TT can be assessed only on the basis of general criteria, such as logical coherence and plausibility. These factors alone, crucial as they are for interpreting quality, are not enough to allow a broader assessment of quality examining the TT in relation to the ST. Users' understanding of ST content is at best vague, since they would not need interpreters if they could understand it without difficulty."

In an attempt to be comparative to Bühler, Kurz [3:143-148] also used eight criteria from the former's research, including "sense consistency with original message", "logical cohesion of utterance", "correct grammatical usage", "completeness of interpretation", "fluency of delivery", "correct grammatical usage", "native accent", and "pleasant voice". In her research in 1989, Kurz deployed the survey questionnaire to 47 delegates in a medical conference and asked them to rate the importance of different quality criteria on a four level scale (4 = most important, 1 = least important). She continued her research in 1993 on 19 delegates from a quality control conference and 48 delegates from a European Council meeting [12:13-21].

It is interesting that the ranking of criteria by both groups are mostly similar in terms of importance order. Linguistic-semantic criteria are given higher importance than extra-linguistic ones in both research findings. The differences are only in the last criteria: interpreters attach higher importance to "grammar" and "terminology" than delegates do [7]. This is relatively explainable as professional interpreters may be more technically critical towards their own quality.

Kurz's model is selected for this pilot for the following reasons:

Firstly, this model is based on user's perspective. This approach should be prioritized as, to sell a product/service, the producer/supplier has to satisfy the user. If the user is not satisfied and willing to pay, the product/service cannot be viable despite the fact that it may be acceptable to researchers.

Secondly, Kurz's model includes eight criteria which are rather easily quantifiable. This is very important, because an assessment model does not only need validity and accuracy but also feasibility.

5. Data sources and methodology

5.1. Data sources

Data for analysis is recorded from interpreters in three international conferences. Each recording extends to 10-15 minutes, approximately the length of one interpreting turn.

Conference 1: Experience of Non-Governmental Organizations in policy advocacy for gender-based violence issue

Conference 2: Developing green house gas emission mitigations in building sector

Conference 3: Improving budget revenue collection from natural resources

5.2. Methodology

In this research, the quality criteria that Kurz recommended in 1989 and piloted in 1989 and 1993 are used. To make scoring and comparison more consistent, the significance of the least important criterion ("native accent") is used as the base point (it is assigned the weighting of 1). In other words, the significance of seven other criteria reflects they are how many times more important than "native accent".

Table 1. Criteria weighting

No	Criteria	Significance (out of 4)	Weight
1	Sense consistency with original message	3.69	1.6
2	Logical cohesion of utterance	3.458	1.5
3	Correct terminology usage	3.4	1.4
4	Completeness of interpretation	3.2	1.4
5	Fluency of delivery	3.1	1.3
6	Pleasant voice	2.6	1.1
7	Correct grammatical usage	2.6	1.1
8	Native accent	2.365	1

Note: Weight = Significance/2.365 (2.365 is the significance of the least important criterion: "native accent"; Weight is rounded to 0.1 for convenience)

Conference recordings are assessed basing on these eight criteria on the scale of 10.

Score for each criterion is converted to Weighted score.

Average (scale of 10) = total Weighted score /10.4 (10.4 is the sum of Weight).

Results are calculated using a Microsoft Excel table with given formulas (see appendix for further details).

Table 2. Recording scoring sheet

No	Criteria	Significance (out of 4)	Weight	score	weighted score	average (10 scale)
1	Sense consistency with original message	3.69	1.6			
2	Logical cohesion of utterance	3.458	1.5			
3	Correct terminology usage	3.4	1.4			
4	Completeness of interpretation	3.2	1.4			
5	Fluency of delivery	3.1	1.3			
6	Pleasant voice	2.6	1.1			
7	Correct grammatical usage	2.6	1.1			
8	Native accent	2.365	1			
	TOTAL	24.413	10.4			

Recordings, including source speech and interpretation, are transcribed precisely to each pause or sound produced. Transcriptions of source speech and the relevant interpretation are put into a table with two parallel columns for easier comparison. Highlighted criteria (number one – "sense consistency with original message", number three – "correct terminology usage", and number four – "completeness of interpretation") are assessed by comparing transcriptions of source speech and interpretation. The other criteria can be assessed on the basis of the interpretation alone.

Assessment steps:

Step 1: Listen and precisely transcribe the source speech, enter it into the left column.

Step 2: Listen and precisely transcribe the interpretation, enter it into the right column, in

parallel to the left column for easier comparison. While transcribing, the assessor also marks (using New Comment and Text Highlight Color tools in Microsoft Word) the noticeable details, including mistakes and/or errors made by the interpreter.

Step 3: Review the interpretation to scan for any noticeable details that have not been marked.

Step 4: Aggregate noticeable details (evidence) in a Microsoft Excel template.

Step 5: Make comments on each quality criteria, score each criteria, calculate the average score and make overall quality conclusion.

6. Assessment result

6.1. Interpreter at conference 1: Experience of Non-Governmental Organizations in policy advocacy for gender-based violence issue

- Average (scale of 10): 8.356

- Criteria score (detailed comments and evidence are provided in Appendix 10.1):

Criteria	1	2	3	4	5	6	7	8
Score (scale of 10)	8.5	8	8	8	7	9	9	10

- General comment: Basically, the interpreter ensures sense consistency between source speech and interpretation. The interpretation is also clear and cohesive. Most of the details are interpreted. Target language terms are used accurately. Fluency is relatively good. The voice is at moderate volume and pleasant. Grammar use is correct and the accent is exactly native-like. However, the interpreter should improve further on fluency, minimizing "fillers" such as "ah", "uh", etc.

- Conclusion on quality: The interpreter at conference 1 well completed her job.

6.2. Interpreter at conference 2: Developing green house gas emission mitigations in building sector

- Average (scale of 10): 7.875

- Criteria score (detailed comments and evidence are provided in Appendix 10.2):

Criteria	1	2	3	4	5	6	7	8
Score (scale of 10)	7.5	8.5	8.5	7	7.5	8.5	8.5	7

- General comment: The interpreter basically ensures sense consistency between source speech and interpretation. The interpretation is relatively clear and cohesive. Details are interpreted relatively fully but quite a lot of details are missed (partly because the speaker spoke too fast and repeated himself sometimes). Most of the target language terms are used accurately. Fluency is relatively good but there were segments when the interpreter was a little bit struggling (partly because of the "interpreter unfriendly" way of presenting by the speaker. The voice is at moderate volume and pleasant. Grammar use is correct most of the time. However, the accent is not exactly native-like. In addition, two other weaknesses in this interpretation are completeness of interpretation and fluency. It is worth noted, however, that in the source speech recording, the speaker spoke too fast. His ideas were also clumsy and unintentionally repeated for many times. Without cooperation from the speaker, it is very hard for the interpreter to improve these two issues.

- Conclusion on quality: In general, the interpreter at conference 2 completed her job at good quality.

6.3. Interpreter at conference 3: Improving budget revenue collection from natural resources

- Average (scale of 10): 7.794

- Criteria score (detailed comments and evidence are provided in Appendix 10.3):

Criteria	1	2	3	4	5	6	7	8
Score (scale of 10)	8	8	8	8	7.5	9	8	7

- General comment: The interpreter basically ensures sense consistency between

source speech and interpretation. The interpretation is also clear and cohesive. Most of the details are interpreted. Target language terms are used accurately most of the time. Fluency is relatively good. The voice is at moderate volume and pleasant. Grammar use is relatively correct but the accent is not native-like. The interpreter should improve further on terminology use. Besides, fluency should also be improved. However, this would require cooperation from speakers (speakers need to speak more slowly, clearly and limit their self-repetition).

- Conclusion on quality: In general, the interpreter at conference 3 completed her job at good quality.

7. Comments on the applicability of Kurz's model

From the piloted analysis of interpretation in the three conferences, it can be seen that the criteria in Kurz's model helps make quality assessment clearer and less subjective. The assessment result is also in line with audience's preliminary observation (in all three conferences, interpreters were complimented and highly appreciated by service users). The result partly helps interpreters identify their strengths and weaknesses.

There are also limitations to the use of this model. Firstly, the assessor needs recordings of both speakers and interpreters. This is not always available if the conference organizers do not intend to have quality assessment or do not want to disclose it to a third party for a variety of reasons. However, this limitation may not exist if this model is used in an interpreter training or recruitment test. In these cases, the organizers often pro-actively keep and provide recordings needed for assessment.

The second limitation is that the assessment is very time-consuming. The steps that take most of the time are precisely transcribing speaker and interpreter (transcriptions are very long: the content of the first conference amounts to 4700 words, the second conference 4000 words, and the third conference 3000 words). In this research, it took on average four working hours to finish the assessment of 10 minutes recording (step 1: 1 hour, step 2: 1.5 hour, step 3: 0.5 hour, step 4: 0.5 hour, step 5: 0.5 hour). Although the time needed may be shortened when the assessor becomes more familiar with the procedure, it is still too time-consuming to be applied on large scale.

Thirdly, the assessor needs to master both languages and be knowledgeable about the conference technical topic and about the interpretation profession. These conditions help the assessor to make accurate and objective observations on interpreting quality, especially for criterion 1 (sense consistency), criterion 3 (accurate term usage), and criterion 4 (interpretation completeness).

The last limitation of Kurz's model is that it has not taken into account external factors that may influence interpreting quality. There is no mechanism of "score compensation" in the model when the presenter speaks too fast, unclearly and clumsily, the presenter's accent and/or pronunciation is too difficult, documents are not provided in advance, the sound system encounters technical issues, the interpreter's booth is not convenient for seeing presentation screen, or the conference's time is prolonged, etc. Among others, this limitation is the hardest to be resolved as there are so many such factors of which the influence quantification is not easy.

8. Conclusion

This research has piloted Kurz's model (1989) with eight criteria in assessing simultaneous interpreting quality in three conferences. The findings show that this model allows comprehensive, accurate and objective assessment of interpreting quality. They also help pointing out interpreter's strengths and weaknesses. However, there are also limitations in the model, especially regarding large scale applicability and the incorporation of external quality factors.

After this research, further works are recommended in the following directions:

(1) Combining Kurz's model (assessment from user perspective) with Pöchhacker's model (assessment from interpreter perspective) to have more comprehensive observations.

(2) Assessing more samples, including conferences where interpreters do not well perform. This is to see if the recommended model can help distinguish different levels of performance by interpreters.

(3) Shortening the time needed to assess each sample.

(4) Recommending a mechanism to quantify the influence of external quality factors, e.g. speaker's delivery speed and accent, availability of reading materials, sound equipment problems, obstruction from interpreters' booth to projector screen, etc.

References

[1] Kalina, S., Quality Assurance for Interpreting Processes. Journal des traducteurs / Meta: Translators' Journal, vol. 50: 768-784, 2005, Retrieved from https://www.erudit.org/revue/meta/2005/v50/n2/011017ar.pdf on March 21st, 2016.

[2] Gaiba F., Origins of simultaneous interpretation: the Nuremberg Trial, University Press: Ottawa, Canada, 1998.

[3] Kurz I., Conference Interpreting: User Expectations, ATA Proceedings of the 30th Annual Conference: 143-148, 1989.

[4] Kotler, P. and G. Armstrong., Principles of Marketin,. 6th ed., Prentice-Hall, Englewood Cliffs (NJ), 1994.

[5] Kurz, I., Conference Interpreting: Quality in the Ears of the User. Journal des traducteurs / Meta: Translators' Journal, vol. 46, n° 2: 394-409, 2001.

[6] Kahane, E., Thoughts on the quality of interpretation, International Association of Conference Interpreters, 2000, Retrieved from http://aiic.net/page/197/thoughts-on-the-quality-of-interpretation/lang/1 on March 21st, 2016

[7] Chiaro, D. and Nocella, G. 2004. Interpreters' perception of linguistic and non linguistic factors affecting quality: A survey through the World Wide Web, Translators' Journal, vol. 49, no2, 2004, 278-293, Retrieved from https://www.erudit.org/revue/meta/2004/v49/n2/009351ar.pdf on March 21st, 2016.

[8] Bühler, H., Linguistic (Semantic) and Extra-Linguistic (Pragmatic) Criteria for the Evaluation of Conference Interpretation and Interpreters, Multilingua 5-4: 231-235, 1986.

[9] Pöchhacker, F., Researching interpreting quality – Models and methods. Interpreting in the 21st Century – Challenges and opportunities: Selected papers from the 1st Forlì Conference on Interpreting Studies: 95-106. John Benjamins Publishing Company, 2000.

[10] Pöchhacker, F., and Zwischenberger C., Survey on quality and role: conference interpreters' expectations and self-perceptions. International Association of Conference Interpreters. 2010, Retrieved from http://aiic.net/page/3405 on March 21st, 2016

[11] Kalina, S., Quality in interpreting and its prerequisites - A framework for a comprehensive view. Interpreting in the 21st Century – Challenges and opportunities: Selected papers from the 1st Forlì Conference on Interpreting Studies: 121-130, John Benjamins Publishing Company, 2000.

[12] Kurz, I., Conference Interpretation: Expectations of Different User Groups, The Interpreters' Newsletter, 5: 13-21, 1993.

Thử nghiệm mô hình đánh giá chất lượng phiên dịch đồng thời

Nguyễn Ninh Bắc

Khoa Sư phạm tiếng Anh, Trường Đại học Ngoại ngữ, ĐHQGHN,
Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Tóm tắt: Vấn đề đánh giá chất lượng phiên dịch tại các hội thảo vẫn là câu hỏi chưa có câu trả lời thoả đáng. Khi có bất đồng xảy ra về chất lượng phiên dịch, các bên liên quan thường không có một cơ sở chung để đưa ra nhận định của mình. Nghiên cứu này, được hoàn thành với sự bảo trợ của Trường Đại học Ngoại ngữ - Đại học Quốc gia Hà Nội (ĐHQGHN) trong đề tài cấp ĐHQGHN mã số QG.15.35 "Nghiên cứu mô hình đánh giá dịch thuật Anh-Việt", đã thử nghiệm việc đánh giá chất lượng phiên dịch tại ba hội thảo khác nhau sử dụng mô hình của Kurz (1989). Kết quả thử nghiệm cho thấy mô hình của Kurz cho phép đánh giá toàn diện, chính xác và khách quan chất lượng phiên dịch. Bên cạnh đó, kết quả đánh giá còn giúp chỉ ra những điểm mạnh và điểm cần cải thiện của phiên dịch. Tuy nhiên, mô hình cũng có nhiều điểm hạn chế, nhất là ở khả năng ứng dụng đại trà và việc tính tới các yếu tố khách quan ảnh hưởng tới chất lượng của phiên dịch.

Từ khóa: Phiên dịch, chất lượng, đánh giá, Kurz, mô hình.

APPENDIX - ASSESSMENT DATA

10.1. Conference 1: Experience of Non-Governmental Organizations in policy advocacy for gender-based violence issue

Source speech recording: https://goo.gl/3pbkbE

Interpretation recording: https://goo.gl/SKBHda

Transcription of source speech and interpretation: https://goo.gl/oQudLP

Excel file containing detailed comments and evidence: https://goo.gl/R3tSx9

10.2. Conference 2: Developing green house gas emission mitigations in building sector

Source speech recording: https://goo.gl/Ntj0QQ

Interpretation recording: https://goo.gl/902Zy1

Transcription of source speech and interpretation: https://goo.gl/hpIqCC

Excel file containing detailed comments and evidence: https://goo.gl/r8Oojt

10.3. Conference 3: Improving budget revenue collection from natural resources

Source speech recording: https://goo.gl/NSejhL

Interpretation recording: https://goo.gl/whrupJ

Transcription of source speech and interpretation: https://goo.gl/itfGVw

Excel file containing detailed comments and evidence: https://goo.gl/3IMHyf