*Daniel Gile**

Variability in the perception of fidelity in simultaneous interpretation

Abstract

In a series of distinct operations, an English speech made at a conference and its interpretation into French were presented auditorily and visually to professional interpreters, students and other assessors to study what they considered errors and omissions, as well as their assessment of the overall fidelity of the target speech. Results include high intra-group variability in all categories of assessors, marked differences between the number of errors and omissions reported after auditory vs. visual presentation, a generally more lenient assessment by interpreters than by other assessors, and a lack of clear correlation between the number of errors and omissions reported and the general fidelity rating given by assessors. These results suggest that fidelity assessment performed by single assessors and by very small groups of assessors may be very unreliable, and that variability in fidelity norms may be an important factor in fidelity assessment.

1. Introduction: the difficulty of fidelity assessment

Fidelity to the information content of the source text is generally viewed as one of the major determinants of translation quality. At first sight, the concept is intuitively clear to the practitioner, but its actual use in quality assessment is problematic.

Firstly, fidelity norms are anything but invariant: as is increasingly recognized by translation scholars, the very concept of translation may mean different things to different social groups, depending on the historical period and cultural environment (see for instance Toury 1995). The translation studies community has come a long way from the idea

* I am indebted to Miriam Shlesinger for her valuable comments and criticism on an earlier version of this paper.

* *Daniel Gile*
Université Lumière Lyon 2
F-69365 Lyon Cedex 07
Home address: 46, rue d'Alembert
F-92190 Meudon
DGile@compuserve.com

of linguistic correspondence as a fidelity criterion. Even semantic fidelity at the "molecular" level of small groups of words is no longer considered a necessary or sufficient condition for fidelity, as both source texts and target texts are viewed as statements with specific objectives, and fidelity to objectives may call for some information addition and some changes in metaphors, allusions, etc. With Vermeer's skopos theory (see for example Reiss and Vermeer 1984/1991), there is even an explicit normative rule which subordinates the relationship between the content of the source text and the content of the target text to the function of the latter in the target culture. It follows that the existence or absence of a given informational element from the source text in the target text can be perceived as either positive or negative, depending on the assessor's views on fidelity and on his/her interpretation of the reason for this presence or absence.

Secondly, even when assessors agree that a particular text segment in a given translation is not "faithful", they tend to vary in the evaluation of the amplitude of the deviation. This is seen for instance when correcting or assessing translation and interpretation performance in professional and academic environments, with frequent disagreements between correctors on the severity of individual errors.

Thirdly, identifying deviations from fidelity in translation, and especially in interpretation, seems to pose cognitive problems, as explained further down.

This paper deals with the pragmatic issue of information fidelity perception in simultaneous interpreting. More specifically, it looks at inter-group, intra-group and inter-modality (visual vs. auditory) variability in fidelity perception on a case-study basis (only one source speech and one corresponding target speech are used here).

2. Sources of variability in the perception of fidelity in simultaneous interpreting

Many interpreters' professional experience includes instances where users of interpretation services were happy with the service provided while members of the team were not. This may be explained by a difference between the interpreters' expectations and the expectations of the particular group of users concerned (a growing number of investigations have been studying this issue - see a synopsis of the literature

and Kurz's own investigations in Kurz 1996, as well as a more recent survey in Moser 1997). Variability may also be inter-individual, as is often seen in the discussions between members of examination committees on the occasion of professional qualification tests in interpreter training institutions (see Namy 1978). An even more salient and possibly more fundamental difference is seen between the interpreters' fidelity norms and the norms for errors and omissions (e/o's) defined by some linguists and psychologists in their research into interpreting. Strong criticism has been levelled against Barik (1969) for the way he identified errors and omissions in his corpus (see Bros-Brann 1975, Stenzl 1983: 28, Gile 1995c:46-47, Lamberger-Felber 1998). Similar criticism has been aimed at other investigators for measuring interpretation quality by counting percentages of words correctly rendered, in systems that were generally subjective and based on the experimenters' sole intuitions about accurate translation (see Anderson 1979:6-7). Over the past few years, more sophisticated criteria have been used. For example, Tommola and Laakso (1997) have used propositional analysis, with weighting of the propositions depending on their semantic importance.

Cognitive factors may also be involved: in translation, assessing fidelity requires reading the source text and the target text and comparing the two in an operation involving little attention sharing, but sustained attention. When the reader's attention wavers, errors may be missed, which is not infrequent when revising and correcting translations. Identifying errors and omissions in interpretation seems to be more difficult, even in consecutive interpreting, where the target speech follows the source speech. In a classroom experiment on fidelity assessment in consecutive interpreting, Gile (1995b) found that student-assessors with excellent comprehension of both source language and target language did not detect all the errors, not even those which they acknowledged as flagrant when these were pointed out to them. Moreover, some participants, including the speaker, thought the interpreter had made errors which the transcript shows he had not made. The reason for this lack of sensitivity to errors in interpreting may lie in working memory limitations as regards both the amount of information it can store and the time this information can be stored before it decays. After about a second or so, the precise wording of a sentence heard is lost and only a global meaning as extracted by the listener is kept, so

that subtle semantic deviations are often overlooked (further explanations on working memory can be found in introductory textbooks on cognitive psychology). In the case of Gile's experiment, the fact that even gross deviations were lost is also likely to be associated with longer term memory involvement, since the source speech lasted 1'54" and the target speech 1'32", not counting a pause of about one minute in-between.

In simultaneous interpreting, the problem is compounded by the attention-sharing requirement: both source-text and target-text segments are simultaneously present in the listener's working memory, and comparing them is probably only possible on a sampling basis, that is by taking very short segments (of up to one or two seconds or so) in the source text and comparing them to the corresponding target-text segments, meanwhile losing track of the following incoming source-text segments (see a discussion based on the Effort models in Gile 1995a or Gile 1995c).

3. Questions and hypotheses

One recurring issue in empirical studies involving interpretation quality assessment is the legitimacy of its visual (as opposed to auditory) presentation to the assessors. In the literature, Stenzl (1983) and others have stressed that interpreting should be seen as the production of a speech to be heard and processed on the spot - as opposed to written texts, which will be perceived visually, less linearly (with the possibility of glancing again at a specific word or group of words if necessary) and over a longer period than the span of auditory memory. Besides cognitive issues, there are delivery issues; prosody is given increasing attention as an important parameter in the transmission of information (see for example Shlesinger 1994, Williams 1995, Bendik 1996, Collados Ais 1996), and conventional transcripts do not offer all the prosodic information. Some authors have been working on notation systems for prosody (e.g. Bendik 1996), but even when these become widely available, the question is whether assessors will be able to "read" them as they would read music.

In short, many interpreters' claim is that assessing interpretation on the basis of a transcript is misleading. On the other hand, if it were possible to assess fidelity on the basis of transcripts, such a method would

be very convenient due to less taxing requirements in terms of time, attention and equipment, thus providing clear advantages for research purposes. It therefore makes sense to try to see whether assessors do react differently depending on whether they listen to or read the target speech. The investigation presented here addresses the following issues:

1. Are more e/o's detected when material is presented visually, be it for cognitive or other reasons, and is this associated with a lower overall rating of fidelity ?

Hypothesis A: *More errors and omissions are identified in material presented visually than in material presented auditorily, and the overall rating of fidelity is lower.*

2. Since professional interpreters have direct intuitive knowledge of fidelity criteria and interpreting strategies, can they avoid the "traps" of visual presentation and identify the same e/o's in both presentation modes, whereas non-interpreters cannot ? Hypothesis B tests one possible effect of such knowledge.

Hypothesis B: *The difference between the number of e/o's reported in visually presented material and in auditorily presented material is smaller among professional interpreters than in other categories of assessors.*

3. Interpreters are aware of strategic decisions made in the booth on the basis of communication-driven objectives involving manipulation of information (deletion, addition, change of order) such as would be considered "unfaithful" on the basis of a surface-structure comparison. Jones (1998), an experienced European Union practitioner and teacher of interpreting, suggests that for strategic reasons, interpreters should avoid translating some of the speakers' announcements as to the next part of their speech (p.122-123), and in particular announcements about statistics that should follow immediately (p.132). They should also avoid naming the sources of famous quotations when such sources are announced in advance by the speaker (p.124). They should sometimes omit information deliberately "with a view to economy of expression, ease of listening for the audience, and maximum communication between speaker and audience" (p.139). At other times they should add information, for instance when working as a relay (an intermediary

interpreter working from a language unknown to other colleagues into a language known to them so that they can interpret the speaker into the other languages) (p.138). In all these cases, interpreters may consider that their colleague whose target speech they are assessing has done a good job in following these suggestions and view the omissions and additions as justified, while to non-interpreters not familiar with such strategies, they may appear as errors. Is this a significant difference between interpreters and non-interpreters? Hypothesis C addresses one possible consequence of this difference.

Hypothesis C: *Interpreters are more lenient in their assessment of fidelity than non-interpreters.*

4. Last but not least, since actual quality assessment is generally conducted by one, two or a handful of evaluators at most, it is important to study the extent of intra-group variability. If it is substantial, the phenomenon will have to be looked at qualitatively and methods will have to be developed in order to secure assessment reliability. The investigation presented here also looks at intra-group variability. It focuses on quantitative aspects of variability in fidelity assessment, and makes no attempt to study the assessor's norms for "errors and omissions", though the examples and discussion in Appendix D may give a few indications.

4. Material and method

4.1. General introduction to the approach

In many experimental disciplines, the most conventional way to test hypotheses is to set up an experiment, here with a group of professional interpreters (PI) and a group of non-interpreters (NI), each having to perform the same tasks under several conditions. Numerical results for each group and each 'condition' (such as visual presentation V and auditory presentation A) are then submitted to statistical tests based on probability theory. Results may show that differences in the values measured for each group and condition are 'significant' at a certain level of probability, meaning that there is a certain likelihood that they are due not to chance alone, but to the differences 'controlled' by the selection of subjects and by the design of the experimental tasks and conditions.

In the environment of CIR (Conference interpreting research), there are technical obstacles and limitations to a strict implementation of this paradigm. Possibly the most CIR-specific obstacle is the lack of availability of subjects for experiments, a recurrent theme in the literature, which has led to studies with quantitatively and qualitatively inadequate samples: too small, non-random, using non-interpreters such as students or even "bilinguals" without any training in interpreting. In such studies, not only are significance tests mathematically questionable, but the very validity of the measurements is uncertain (for a more detailed discussion of these issues, see Gile 1995c, 1998). Another major problem is the requirement for strictly controlled experimental conditions where all relevant variables are deemed to have the same (mean) value except for the specific variable(s) the effect of which is investigated. In order to meet this requirement, experimenters have sacrificed 'natural' conditions. For instance, in order to make sure subjects interpret exactly the same speech in terms of words, voice, prosody, etc. so as to avoid 'parasitic' effects from variations in these input parameters, they may ask subjects to interpret a speech from a tape-recorder rather than from a live speaker. They may even write source speeches around specific linguistic criteria necessary to achieve the required control (e.g. in Dillinger 1989) rather than collect them in the field in 'natural' situations where all relevant variables and their interactions are 'authentic'. The problem is that the precise effects of such deviations from 'natural' conditions is not known, just as the precise effect of uncontrolled variability in natural conditions is not known; what is gained in comparability may be lost in validity. Controlling experimental conditions is of course desirable, but it is not a perfect solution that does away with all the problems, including the problem of confounding variables, and it is not necessarily the best solution in all cases.

An improvement of the situation should occur as available data from empirical studies - both experimental and observational - increases in volume. The more variables are studied and the more replications are conducted, the more information will be available on relevant variables, thus providing more solid ground for experimental studies in the conventional paradigm.

Another research strategy is to use data from a series of distinct studies and make inferences by analyzing the accumulated information.

For instance, in medicine, researchers look at clinical data collected in various hospitals about patients, treatments and outcomes to make up their mind regarding the advisability of specific management strategies or regarding a causal relationship between a variable and a pathological condition (as in the case of smoking and lung cancer). Similarly, in the behavioral sciences, researchers look at existing data and analyze it so as to test the links between juvenile violence and TV violence, and in econometrics, available economic data is analyzed to study the links between government intervention on specific variables and the subsequent behavior of economic systems.

This strategy allows the use of a large amount of data, far above what could be collected in a single experimental study. This is a particularly important advantage in the CIR environment, where access to subjects is so difficult. Moreover, when studies included in the analysis are observational (i.e. when they study phenomena as they occur 'naturally', as opposed to experimental studies), there are less serious validity problems (see the discussion in Gile 1998). On the other hand, 'natural' variability in the conditions in which each study is conducted makes inferencing more difficult than in the conventional experimental paradigm.

The small series of studies presented here in a 'consolidated' form is an illustration of a middle path strategy designed to try to maximize the advantages and minimize the effects of the limitations of the two paradigms explained above. The fundamental idea is to re-use the same material in several studies so as to increase the amount of comparable data. As a study is published, other investigators can use part or all of the same input and output material for either replication or distinct research objectives, and have the benefit of an analyzable corpus far larger than what they could collect on their own (this is a growing demand in CIR circles - see Gambier et al. 1997:121).

In the present investigation, an initial endeavor to study variability in error and omission (e/o) perception was made with a first series of interpreters. At a later stage, I became interested in overall fidelity assessment and in its relation to the number of e/o's perceived. Instead of planning a completely new set-up, I decided to keep the material (the source and target speeches and their transcripts) and the task unchanged, and added a question about fidelity, which made it possible to use the previous data. Similarly, the initial interest was just in the popula-

tion of professional interpreters. At later stages, I also became interested in non-interpreters, and in particular in the sub-populations of students and teachers, mainly for two reasons:

- Students are much more available and willing to serve as subjects for research than professional interpreters. It therefore makes sense to try to see in what way their characteristics as regards research tasks (in this case, e/o identification and fidelity rating) might differ from those of professional interpreters or other non-interpreters.
- Gaining insight into the reactions of students and teachers is potentially useful in the context of translation and interpretation didactics.

The data on students and translation teachers collected here is not sufficient for much distinct inferencing at this stage (see section 5.5.4 and the discussion in section 6), but it has been integrated into the pool of data on non-interpreters; the sample can be developed at a later stage to allow further analysis of this distinct sub-population.

4.2. Description

The source speech (Appendix A) is part of a set of authentic speeches which I have collected over the years for use in the classroom and in research. Many are used several times, both for didactic and for experimental purposes. In an initial step, the speech was sent out to a handful of professional interpreters to see whether there was much variability in the way they identified e/o's. Following discussions on translation assessment with fellow translation instructors at Université Lyon 2, it was also used, with the same method, for a comparison between professional interpreters and translation instructors. The addition of fidelity ratings came in later, as well as the auditory presentation of the same material for inter-modality comparisons. More subjects were added when the opportunity arose to increase sample size, always using the same source and target speech and the same method as explained. Overall, data was collected over about 2 years.

The original speech was made in English at UNESCO at an interpreted meeting on street children by Christina Noble, who established a charitable foundation in Vietnam to take care of such children (and who kindly gave the author permission to use this material). The speech was interpreted simultaneously on site, and was made available for research

to other colleagues in various countries (it was used for example in a study on cohesion in Shlesinger 1995).

The first 3 minutes of one simultaneous interpretation by an experienced professional interpreter (>15 years of experience) working from his B language (English) into his A language (French) were used throughout the investigation:

a. Transcripts of the speech (appendix A) and its interpretation (appendix B) were shown to 13 professional conference interpreters (VPI1 to VPI13 - where "V" stands for "Visual") who were asked to underline with a pencil on the transcript of the target speech every word or speech segment they considered an error or omission. Fifteen interpreters (VFPI14 to VFPI28 - where "F" stands for "Fidelity Assessment") were also asked to rate fidelity on a scale of 1 to 5, 1 being "very poor" and 5 "very good" (see appendix C for the precise wording in French).

b. Transcripts of the speech and its interpretation were also given to 8 teachers of translation (VTT1 to VTT8) at the department of Modern languages ("Langues étrangères appliquées") at Université Lumière Lyon 2, who were asked to identify errors and omissions. None of them was a professional translator, and they use translation primarily as a language teaching exercise.

c. Transcripts of the speech and its interpretation were handed out to 12 doctoral students (VFDS1 to VFDS12) at the faculty of letters of the same university, and the same instructions were given to them. None was a translator or interpreter.

d. During a workshop at a Belgian university, participants were given a transcript of the source speech, and the first three minutes of the interpreted target speech were played sentence by sentence (to avoid the loss of errors and omissions that could have been detected but might not be remembered by raters) using a cassette player connected to a public address system. After each sentence, participants were asked to underline errors and omissions on the transcript of the source speech, and to rate fidelity on the same 1 to 5 scale. In order to keep the identity of the interpreter confidential, I had "re-interpreted" the source speech from a recording, using from a transcript the precise words that the interpreter had used. This is similar to interpretation tasks that occur occasionally in real-life, when the interpreter is asked to "interpret" a source speech using a target speech script. Listeners were informed of this "re-interpreting" procedure and of its reasons, so that they would not hesitate to

criticize the interpretation thinking that they were criticizing the experimenter who stood in front of them.

The auditory groups were 18 professional conference interpreters (AFPI1 to AFPI18 - where "A" stands for "Auditory presentation") , 26 translation and interpretation students (AFTIS1 to AFTIS26) , and 4 non-interpreting scholars (AFNIS1 to AFNIS4) from Belgian universities.

A few additional comments on the source and target speeches will help put the analysis in a proper perspective. The source speech was adlibbed and emotional. it was not informationally dense, with many names, numbers and facts, for which it would have been easy to determine which information was reproduced in the target speech and which was not, as was the case in Gile 1995b. The target speech contains no flagrantly serious error, and no e/o was singled out as such by the subjects.

5. Results

The raw data for each subject is presented in table 1 (for the visual groups) and table 2 (for the auditory groups). Table 3 presents the distribution in the number of e/o's reported ("e/o number") in each group. For each group, the number and percentage of subjects who reported 0, 1, 2, 3... e/o's are indicated. For instance, in group VFDS (doctoral students), 2 subjects (17% of the group) reported 3 e/o's each, and 3 subjects, that is 25% of the group, reported 10 e/o's each. Numbers have been rounded to the next integer.

5.1. Reported e/o's

Overall, the spread in the number of e/o's reported is wide, going from 0 to 33. Even if one disregards the highest values (17, 24 and 33) because each was only reported once in the whole series of 96 subjects, one can still find low values of 0 and 1 e/o's reported by 14 subjects, and high values of 11 e/o's and above reported by 13 subjects, which demonstrates that the spread is indeed wide.

More interestingly, the spread is wide and rather even (as opposed to a curve where most values would concentrate around one central point) not only for the whole sample, but within groups as well. This means that when asking an assessor taken at random to report all e/o's in the

visual or auditory presentation mode as was done in this investigation, there is a roughly equal likelihood of obtaining e/o numbers in a wide range of values.

The smallest spread is found in the auditory groups, with professional interpreters (0 to 7), followed by translation and interpreting students (1 to 12) and non interpreting scholars (0 to 12). In the visual groups, the spread is generally larger, with 2 to 15 for translation teachers, 1 to 17 for doctoral students, and 0 to 33 for professional interpreters. This last result is rather intriguing, especially in view of the fact that in the auditory group, professionals interpreters had the smallest spread.

The mean number of e/o's reported is smaller in the auditory groups than in the visual groups, and smaller for interpreters than for non-interpreters: in the auditory groups it is 2.89 for interpreters, 4.29 for translation and interpreting students and 4.5 for non-interpreting scholars, and in the visual groups 6.86 for interpreters, 7.29 for translation teachers and 9.25 for doctoral students.

5.2. Fidelity ratings

The spread of fidelity ratings is small and slightly larger on average in the visual groups (3 to 5 among professional interpreters and doctoral students, 2 to 5 among translation teachers) than in the auditory groups (4 to 5 among professional interpreters and non-interpreting scholars, 3 to 5 among translation and interpreting students). Differences in mean values are more clear-cut: in the visual groups, means are markedly below 4 (3.9 and 3.5), and in the auditory groups, markedly above (4.31, 4.23 and 4.5). In the visual groups, they are higher for interpreters (3.9) than for doctoral students (3.5), and in the auditory groups, they are more closely concentrated, and lie between 4.23 for translation and interpreting students and 4.5 for non-interpreting scholars, with the intermediate value of 4.31 for interpreters.

5.3. Links between fidelity ratings and the number of e/o's reported

In none of the groups does there seem to be a straightforward link between the number of e/o's reported and fidelity rating. For instance, in the visual groups (table 1), fidelity ratings of 4 are given for 0, 1 and 2

e/o's, but also for 9 and 15 e/o's, and in the auditory groups, fidelity ratings of 5 are given for 0, 1, 2, 3, 4, 5, 6 and even 12 e/o's. On the other hand, subject VFPI6 reports 3 e/o's but only gives a fidelity rating of 3. Similarly, out of 4 raters who reported 0 e/o's, two interpreters and one non-interpreting scholar only rate fidelity at 4, and one interpreter at 5.

Table 4 shows mean fidelity ratings for each e/o number. While there is a general downwards movement in the ratings as one moves from 0 e/o's to 10 e/o's and beyond, it is not very regular, insofar as fidelity ratings sometimes (counter-intuitively) go up when the e/o number increases (from 0 to 1 to 2 e/o's, then from 3 to 4 e/o's, then from 5 to 6 to 7 e/o's, then from 11 to 12 e/o's). To some extent, this can be ascribed to random variation in very small samples (see the "n" column in table 4), but this also occurs with larger samples of 9 and 8 individuals (for 1 e/o and 2 e/o's respectively).

5.4. Inter-modality differences in fidelity ratings

On the whole, fidelity ratings are higher in the auditory groups than in the visual groups. The mean fidelity rating for "auditory interpreters" is 4.31, vs. 3.9 for "visual interpreters", and the mean fidelity rating for non-interpreters in the auditory group (translation and interpreting students and non-interpreting scholars) is 4.27 vs. 3.5 in the visual group (doctoral students). When comparing the ratings for each e/o number (table 4), one also finds a difference, but it is less clear-cut. For each e/o number, the mean fidelity rating given by auditory interpreters is equal to or higher than the mean rating given by visual interpreters, but the same relation only holds true for 2 out of the three e/o numbers reported by non-interpreters: for 1 reported e/o, the mean fidelity rating for visual non-interpreters is 5, while it is only 4.33 for auditory interpreters (note, however, that only one visual non-interpreter reported 1 e/o, and out of the three auditory non-interpreters who reported 1 e/o, one also gave a fidelity rating of 5).

5.5. Synopsis of the results

5.5.1. Hypothesis A regarding inter-modal differences in e/o perception and fidelity ratings

In the groups of professional interpreters, with a mean e/o number of 2.89 in the auditory mode vs. 4.9 in the visual mode (+ 69%), there is a clear inter-modal difference in e/o perception. Similarly, with a mean fidelity rating of 4.31 in the auditory mode vs. 3.9 in the visual mode (- 0.41 points in the total range of 2 points from 3 to 5 used by all subjects), there is an inter-modal difference in their fidelity ratings.

For non-interpreters, the mean e/o number in the auditory mode is 4.33, vs. 8.4 in the visual mode (+ 94%), and the mean fidelity rating in the auditory mode is 4.27 vs. 3.5 (- 0.77 points). Again, the inter-modal difference is substantial.

These results are in line with the hypothesis that target speeches are indeed assessed more leniently when presented auditorily (vs. visual presentation).

5.5.2. Hypothesis B regarding the amplitude of inter-modal differences between professional interpreters and non-interpreters

Professional interpreters reported on average 6.86-2.89 = 3.97 (58%) less e/o's in the auditory modality than in the visual modality. Non-interpreters reported on average 8.4-4.33 = 4.07 (48%) less e/o's in the auditory modality than in the visual modality.

In terms of fidelity ratings, professional interpreters differed inter-modally by 4.31-3.9 = 0.41 (89%), while non-interpreters differed by 4.27-3.5 = 0.77 (18%).

On both parameters, the amplitude of inter-modal differences is larger among professional interpreters than among non-interpreters. This would suggest that the difference between the assessment of visually presented material and auditorily presented material in professional interpreters is not smaller than in other categories of assessors.

5.5.3. Hypothesis C regarding the relative leniency of the professional interpreter's assessment of fidelity vs. other categories of assessors

This hypothesis is borne out clearly in the visual mode, with a mean fidelity rating difference of 0.4 points (3.9 for interpreters vs. 3.5 for non-interpreters). Results are less clear in the auditory mode, with a mean rating difference of only 0.04 points (4.31 vs. 4.27), and with higher ratings for non-interpreting scholars than for interpreters (4.5 vs. 4.31).

5.5.4. Students and non-students

As explained in section 4.1, the reactions of students are of particular interest, both for research purposes and for training purposes. The students included in the auditory part of the series differ from those included in its visual part insofar as they are more familiar with and interested in translation and interpretation as future practitioners. Inter-modal comparison within a distinct category of "students" is therefore problematic. However, intra-modal comparison is possible:

It appears that in the visual mode, students report on average more e/o's than other non-interpreters (9.25 vs. 7.12). In the auditory mode, the relationship is reversed (4.29 vs. 4.72). As to fidelity ratings, at this point, data from students is only available for the auditory mode, where their mean rating is lower than the other non-interpreters' (4.23 vs. 4.5).

6. Discussion

6.1. The rather complex link between e/o numbers and fidelity ratings is made more intriguing by the fact that the target speech contained no flagrantly serious errors.

One possible explanation of this lack of a clear correlation could lie in the perception by assessors of subtle differences in information content between source speech and target speech that are not identifiable by the assessors as individual errors and omissions but which do have an impact on overall fidelity perception.

Another possibility is that assessors find it difficult to separate fidelity from other quality criteria such as the quality of the linguistic output. In the visual groups, several subjects underlined words and word

groups, and then deleted the mark because they decided that the problem was more with style than with information content (comments to that effect were made by them during the operation). Assuming that stylistic features of ad-libbed speech are more salient when looking at a transcript than when listening to a tape, this could explain at least partly the difference in overall fidelity assessment between the auditory and the visual modes of presentation. It is interesting that such differences are not smaller in professional interpreters than they are in non-interpreters. In spite of their direct intuitive knowledge of interpreting norms and strategies, such professionals seem to react like non-interpreters to transcripts (which would explain why some reported many e/o's in the visual mode while the spread was narrow in the auditory mode). Nevertheless, in absolute terms, their e/o numbers are lower than the non-interpreters' in both presentation modalities. In view of their training and the nature of their work, it is unlikely that this is due to a lower cognitive capacity in professional interpreters. In fact, in a recent study by Padilla (1995), it was found that the interpreters' working memory span was higher than that of control subjects. This suggests that the difference may lie in the subjects' norms.

The lack of a straightforward link between e/o numbers and fidelity ratings may also be explained by the latter's possibly larger dependence on expectations. Subjects whose experience with interpreting has led them to lower expectations may be more lenient in their overall assessment than 'naive' subjects (see Moser 1997 regarding the correlation between experience and expectations).

6.2. The familiarity of translation and interpreting students with interpreting processes and strategies could have been expected to generate data comparable to the professional interpreters' rather than to other non-interpreters'. One reason for the different outcome may have been psychological, as they were performing a task assigned to them in an academic environment where they may have felt some pressure to "perform well". In the visual groups, teachers (from the same university as the author) were language teachers who give translation courses and apply language-oriented criteria rather than professional-translation criteria. On the basis of the way they correct students' translations, they could have been expected to be stricter in their assessment than the students. The fact that they were not may be explained by the fact that un-

like the students, they did not feel the same pressure, and by their maturity and better grasp of real-life situations, difficulties and strategies, as opposed to the students, who may have had a strongly classroom-bound view of translation and interpreting. The findings suggest caution in the use of students as fidelity assessors.

6.3. Forty six professional interpreters were included in this study. This was only possible with the incremental sample-building strategy explained in section 4.1 and with limited time and effort requirements from each subject. In the auditory group, data collection took less than twenty minutes during a one-day workshop on interpreting research. I believe that this was perceived as acceptable as a demonstration of empirical research within that framework. However, asking participants to explain why they viewed specific segments as e/o's or to allocate relative weights would have been problematic under the circumstances, since one to two hours would have been required. In the visual group, colleague-interpreters were asked to contribute in the working place, in the midst of a busy schedule. Had they been asked to do more than underline e/o's and rate fidelity, the operation could not have been completed during the working day in the presence of the experimenter, and the response rate would probably have been very low. A compromise had to be found between a large amount of individual data obtained from a small and possibly biased sample (respondents willing to devote much time and energy to the exercise may not be representative of the population of interpreters at large), and a smaller amount of individual data from a larger sample. In this study, which focuses on variability, the choice was made in favor of the latter, with the associated limitations. On the other hand, as explained in section 4.1, using the transcripts as reproduced in this paper, it is possible and desirable to extend the study over time and/or replicate it for confirmation and fine-tunig.

7. Conclusion

Overall, the results tend to strengthen the idea that transcripts are not assessed like auditorily presented material, even by interpreters. This does not mean that transcripts should not be used, if only because they allow the identification of unchallengeable e/o's that are not detected in

an auditory presentation (Gile 1995b). It does mean, however, that they may not be a reliable tool for fidelity assessment.

Another salient finding from this study is the substantial variability in both e/o numbers and fidelity ratings, found in all groups. Its reasons remain to be investigated, but one important implication is clear: evaluators asked to identify e/o's and assess fidelity without being given more precise definitions and instructions are likely to vary substantially from each other. In an interpretation exercise involving a speech similar to the one used here, such a task is likely to produce unreliable results. This variability could probably be reduced using more precise and explicit criteria for errors and omission identification as well as more precise and explicit instructions to the assessors, but caution is necessary to avoid jeopardizing validity. As long as such methods have not been developed and validated, it is hazardous to identify errors and omissions in research and use them for comparative assessments using one or two informants only. The uncertain correlation between the number of errors and omissions reported and the corresponding fidelity ratings also suggest that the perception of fidelity may be more complex than appears, and include aspects which investigators would not necessarily think of classifying under "fidelity".

Findings from this study therefore suggest that much caution is required when assessing interpretation quality on the basis of transcripts, when using metrics such as content comparisons at the level of words or propositions, and when asking non-interpreters to act as fidelity assessors, at least in studies on the effect of manipulating input variables on interpreting performance (such as Tommola & Lindholm 1995).

References

Anderson, Linda (1979). *Simultaneous Interpretation: Contextual and Translation Aspects*. Unpublished M.A. Thesis, Concordia University, Montreal, Department of Psychology.

Barik, Henri C. (1969). *A Study of Simultaneous Interpretation*. Unpublished Ph.D. thesis, Chapel Hill. University of North Carolina.

Bendik, Jozsef (1996). 'On Suprasegmentals in Simultaneous Interpreting'. In Klaudy, Kinga, José Lambert and Aniko Sohar (eds) *Translation Studies in Hungary*. Budapest: Scholastica, 176-190.

Bros-Brann, Eliane (1975). 'Critical Comments on H.C. Barik's article: Interpreters talk a lot, among other things'. In *Babel* 21:2. 93-94.

Collados Ais, Angela (1996). *La entonación monótona como parametro de calidad en la interpretación simultánea: la evaluación de los receptores*. Tesis doctoral, Universidad de Granada, Facultad de traducción e interpretación, Departamento de lingüística Aplicada a la Traducción e Interpretación.

Dillinger, Michael (1989). *Component Processes of Simultaneous Interpreting*. Unpublished PhD dissertation, McGill University.

Gambier, Yves, Daniel Gile and Christopher Taylor (eds) (1997). *Conference Interpreting: Current Trends in Research*. Amsterdam/Philadelphia: John Benjamins.

Gile, Daniel (1995a). *Basic concepts and models for interpreter and translator training*. Amsterdam/Philadelphia: John Benjamins.

Gile, Daniel (1995b). 'Fidelity Assessment in Consecutive Interpretation: An Experiment'. In *Target* 7:1.151-164.

Gile, Daniel (1995c). *Regards sur la recherche en interprétation de conférence*. Lille: Presses Universitaires de Lille.

Gile, Daniel (1998). 'Observational Studies and Experimental Studies in the Investigation of Conference Interpreting'. In *Target* 10:1.69-93.

Jones, Roderick (1998). *Conference Interpreting Explained*. Manchester: StJerome Publishing.

Kurz, Ingrid (1996). *Simultandolmetschen als Gegenstand der Interdisziplinären Forschung*. Wien: WUV - Universitätsverlag.

Lamberger-Felber, Heike (1998). *Der Einfluss kontextueller Faktoren auf das Simultandolmetschen. Eine Fallstudie am Beispiel gelesener Reden*. Unpublished doctoral dissertation, Karl-Franzens-Universität Graz.

Moser, Peter (1997). 'Expectations of users of conference interpretation'. In *Interpreting* 1:2.145-178.

Namy, Claude (1978). 'Reflections on the training of simultaneous interpreters: a metalinguistic approach'. In Gerver, David & H. Wallace Sinaiko (eds) *Language Interpretation and Communication*. Nato Conference Series, Series III: Human Factors, New York and London: Plenum Press. 25-33.

Padilla, Presentación (1995). *Procesos de memoria y atención en la interpretación de lengua*. Tesis doctoral, Universidad de Granada, Departamento de Filología Inglesa.

Reiss, Katarina and Hans J. Vermeer. (1984/1991). *Grundlegung einer allgemeinen Translationstheorie*. Linguistische Arbeiten 147, 2nd edition, Tübingen: Niemeyer.

Shlesinger, Miriam (1994). 'Intonation in the Production and Perception of Simultaneous Interpretation'. In Lambert, Sylvie and Barbara Moser-Mercer (eds) *Bridging the Gap*. Amsterdam/Philadelphia: John Benjamins. 225-236.

Shlesinger, Miriam (1995). 'Shifts in Cohesion in Simultaneous Interpreting'. In *The Translator* 1.2:193-214.

Stenzl, Catherine (1983). *Simultaneous Interpretation: Groundwork towards a Comprehensive Model*. Unpublished M.A. thesis, University of London.

Tommola, Jorma & Tiina Laakso (1997). ‘Source Text Segmentation, Speech Rate and Language Direction: Effects on Trainee Simultaneous Interpreting’. In Klaudi, Kinga & János Kohn (eds) *Transferre Necesse Est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting, 5-7 September*, 1996, Budapest, Hungary, Budapest: Scholastica. 186-191.

Tommola, Jorma & Johan Lindholm (1995). ‘Experimental research on interpreting: which dependent variable?’. In Tommola, Jorma (ed) *Topics in Interpreting Research*. University of Turku, Centre for Translation and Interpreting.

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

Williams, Sarah (1995). ‘Observations on Anomalous Stress in Interpreting’. In *The Translator* 1:1. 47-64.

APPENDIX A - Source Speech

Hello Ladies and Gentlemen. I'd like to apologize for the the films they weren't very informative. But that wasn't my fault. It's because the Vietnamese government chopped out what I really wanted to show you. Anyway um there's one thing I'd like to correct and that is I'm down as the "Brigade Foundation". I am in fact the Christina Noble Rigade Foundation. I say I called it "Rig" because I was hoping to encourage the oil people to give a little bit back to the countries that they take the oil from. Okay. Um most of you know I think that I myself was a streetchild and indeed I lived between the streets and institutions for about six years. Can you hear me ? And uh so I know a little bit about street life and how the child feels.

Um When I went into Vietnam I went in on the pretext of doing business; it was the only way I could get in at that time and started working the same night with the street children. Since going into Vietnam I've built a medical and social center which caters for the homeless, the poor, the the street kids, the youth of the street, abandoned babies and indeed anybody who cannot afford to go to a hospital. It's open 24 hours a day and there are 29 staff split open to three shifts with a doctor on call 24 hours a day. Nobody is refused entry into the hospital. I select who comes in and who doesn't and I do that for a reason. Because there are many people in Vietnam who have got plenty of money but would still use our facilites. And I've worked too hard for those.

APPENDIX B - Target Speech

Bien Bonjour mesdames et messieurs, je voudrais m'excuser pour les films qui n'étaient pas très informatifs. Ça n'était pas vraiment ma faute. Il y a eu des parties que le gouvernement vietnamien a censurées. C'étaient justement les choses que je voulais vous montrer les parties que je voulais vous montrer.

Je voudrais simplement dire que je ne suis pas la fondation Brigade comme c'est écrit sur le papier mais je suis je représente la Fondation Rigade Christina Noble. J'ai parlé de Rig parce que j'espérais que les compagnies pétrolières rendraient quelque chose à ceux dont ils prennent le pétrole. Il y a un jeu de mots parce qu'en anglais le mot "Rig" s'applique à l'infrastructure de l'exploitation pétrolière... les machines.

Bien je pense que la plupart d'entre vous savent que j'étais moi-même une enfant de la rue et que j'ai vécu entre la rue et les institutions pendant six ans. Est-ce que vous m'entendez ? Donc je connais un petit peu la vie dans la rue je connais un petit peu les sentiments des enfants.

Quand je suis allée au Vietnam, je suis allée en prétextant vouloir y faire des affaires. C'était à ce moment là la seule manière d'y entrer. Et j'ai commencé à travailler au cours de la même nuit avec les enfants de la rue.

Depuis, j'ai mis en place un centre social et médical qui s'occupe des enfants de la rue, des bébés abandonnés, des sans-abris, des pauvres, et en fait de tous ceux qui ne peuvent se permettre d'aller à l'hopital. L'institution est

ouverte vingt-quatre heures par jour, il y a un personnel de 29 personnes trois équipes par jour et il y a un mdecin qui est de service vingt-quatre heures sur vingt-quatre.

Et on ne refuse à personne l'admission à l'hopital. C'est moi qui sélectionne ceux qui viennent ceux qui sont admis et ceux qui ne le sont pas je le fais pour une raison précise. Parce qu'il y a beaucoup de Vietnamiens qui ont beaucoup d'argent mais qui préféreraient quand-même utiliser notre centre et j'ai trop travaillé pour ces gens là.

APPENDIX C - Instructions given to participants

1. Sur le texte français, souligner les mots et groupes de mots où vous considérez qu'il y a erreur ou omission.

2. Une fois que vous avez terminé, noter la fidélité de l'interprétation par un chiffre de 1 à 5 selon l'échelle suivante:

1. Très mauvaise
2. Mauvaise
3. Moyenne
4. Bonne
5. Très bonne

APPENDIX D - Examples in the identification of individual e/o's

The following are examples of e/o's reported by subjects and some comments indicating possible directions for further investigation; some are speculative, and some reflect the subjects' own explanations and comments.

1. e/o n˚1

Addition: The interpreter started his rendering of the speech with "Bien" ("All right", "OK", "Well now", etc.) in the target speech. The numbers below show the percentage of interpreters vs. non-interpreters who reported it as an e/o in the two modes:

	PI	NI
A	0%	0%
V	14%	55%

2. e/o n˚2

Omission: In French, the interpreter did not translate the English "but" when interpreting "But that wasn't my fault".

	PI	NI
A	0%	0%
V	11%	15%

3. e/o n˚3

Addition: The interpreter added emphasis by translating the speaker's "...that wasn't my fault" by "Ça n'était pas vraiment ma faute" (It wasn't really my fault).

	PI	NI
A	0%	23%
V	21%	25%

4. e/o n˚4

Omission: The interpreter omitted the explanatory "because" in his translation of "It's because the Vietnamese government..."

	PI	NI
A	0%	0%
V	14%	30%

5. e/o n˚5

Omission: The interpreter translated "correct" in "There's one thing I'd like to correct" by "dire" (say):

	PI	NI
A	17%	20%
V	18%	30%

6. e/o n˚6

Substitution: The interpreter translated "oil people" by "les compagnies pétrolières" (the oil companies).

	PI	NI
A	6%	13%
V	14%	10%

7. e/o n˚7

Omission: The interpreter translated "countries" (in "to the countries that they take the oil from") by "ceux dont ils prennent le pétrole" (those that they take the oil from)

	PI	NI
A	17%	26%
V	21%	15%

Comments:

1. Overall, percentages are higher for non interpreters than for interpreters, except in the visual mode in the last two examples. For interpreters, percentages are higher in the visual mode than in the auditory mode, illustrating

the assessors' tendency to identify more e/o's in the former mode. For non-interpreters, they are also higher in the first 5 cases, and lower in the last two.

2. Besides possible attention fluctuations causing the subjects to "miss" e/o's, the following hypotheses, based on comments made by interpreters in the visual group discussing the various e/o's with me or amongst themselves after the experiment, are given as an initial reflection on qualitative factors which may explain the variability. Since an introspective report on decisions was not part of the study, I only picked up the comments as they came, without intervening and without trying to ascertain that these were the only explanations, or the most frequent ones.

2.1. The fact that none of the "auditory subjects" identified the word "Bien" at the beginning of the target speech as an e/o may show that they reacted to it as a starting utterance with no link to the content of the speech. For "visual subjects", either this was not clear, or else they decided that on paper, they had to identify it as an e/o, thus applying stricter criteria, though its weight in terms of fidelity rating may be nil.

2.2. The omission of "but" ("but that wasn't my fault") is clear-cut, but those respondents who did not identify it as an e/o may have considered that the opposition between the non-informativeness of the films and the speaker's statement about it not being her fault conveyed the same idea, and that "but" was redundant.

2.3. The addition of "vraiment" (really) in "...that wasn't my fault" was considered by some a "natural" utterance which did not add emphasis despite the adverb (comments made by subjects).

2.4. While it is possible that respondents missed "because" due to insufficient attention, it is more likely that at least some of them considered that the context made the causal relationship between '"it wasn't my fault" and "the Vietnamese government chopped out what I really wanted to show you" clear enough to make the word unnecessary (comment made by subjects).

2.5. Similarly, the word "correct" before a sentence correcting the name of the speaker's foundation may have been considered unnecessary in informational term by some respondents. Incidentally, the use of the word "dire" in the target speech may well have been due to the interpreter's staying too "close" to the speaker. "There's one thing I'd like to..." translates well into "Il y a une chose que je voudrais...". When the speaker then says "...to correct", a problem arises, since "Il y a une chose que je voudrais corriger" sounds clumsy in French. The interpreter may therefore have opted for "Il y a une chose que je voudrais dire" ("There's one thing I'd like to say"), leaving himself the possibility of adding the "correction" option in the next sentence if necessary. If this was the case (there is no way to ascertain it), then the deviation from linguistic correspondence was the result of a strategy to preserve good quality

of the linguistic output rather than a symptom showing miscomprehension of the source text.

2.6. As regards “oil people”, the majority of respondents may have considered that translating these words by “compagnies pétrolières” (oil companies) was legitimate and actually added value to the speech by making it more explicit through the use of a word that the speaker may have been unable to retrieve due to lexical restriction. Other assessors may have considered that interpreters have no right to take such decisions.

2.7. With respect to “countries” being translated by “ceux” (those people), the situation is opposite, with a loss of accuracy in the interpreter’s speech. While some respondents may have missed the e/o, others may have considered that the loss was not significant.

Subject	Number of e/o's	Fidelity rating
PROFESSIONAL INTERPRETERS		
VFPI1	0	4
VFPI2	1	4
VFPI3	2	4
VFPI4	2	4
VFPI5	2	5
VFPI6	3	3
VFPI7	3	4
VFPI8	5	4
VFPI9	6	4
VFPI10	6	4
VFPI11	8	3.5
VFPI12	8	4
VFPI13	8	4
VFPI14	9	4
VFPI15	11	3
For VFPI:		
Mean	4.9	3.9
Standard deviation		1.07
Range	0 to 11	3 to 5
VPI1	1	
VPI2	2	
VPI3	3	
VPI4	3	
VPI5	4	
VPI6	4	
VPI7	6	
VPI8	7	
VPI9	7	
VPI10	11	
VPI11	13	
VPI12	24	
VPI13	33	
For all interpreters:		
Mean	6.86	
Standard deviation	6.82	
Range	0 to 33	
DOCTORAL STUDENTS		
VFDS1	1	5
VFDS2	3	3
VFDS3	3	5
VFDS4	5	4
VFDS5	10	2
VFDS6	10	3
VFDS7	10	4
VFDS8	11	3
VFDS9	11	3
VFDS10	15	3
VFDS11	15	4
VFDS12	17	3
Mean	9.25	3.5
Standard deviation	5.39	1.25
Range	1 to 17	3 to 5
TRANSLATION TEACHERS		
VTT1	2	
VTT2	4	
VTT3	5	
VTT4	7	
VTT5	7	
VTT6	8	
VTT7	9	
VTT8	15	
Mean	7.12	
Standard deviation	3.65	
Range	2 to 15	
Mean for all non-interpreters	**8.4**	**3.5**

Table 1: Number of e/o's reported and fidelity ratings in the visual groups

Subject	Number of e/o's	Fidelity rating
PROFESSIONAL INTERPRETERS		
AFPI1	0	4
AFPI2	0	5
AFPI3	1	4
AFPI4	1	4
AFPI5	1	4
AFPI6	1	5
AFPI7	2	4
AFPI8	2	4.5
AFPI9	3	4
AFPI10	3	5
AFPI11	3	5
AFPI12	3	5
AFPI13	4	4
AFPI14	5	4
AFPI15	5	4
AFPI16	5	4
AFPI17	6	4
AFPI18	7	4
mean for AFPI	2.89	4.31
Standard deviation	2.02	1.08
Range for AFPI	0 to 7	4 to 5
TRANSLATION AND INTERPRETING STUDENTS		
AFTIS1	1	4
AFTIS2	1	4
AFTIS3	2	5
AFTIS4	2	5
AFTIS5	2	5
AFTIS6	3	4
AFTIS7	3	4
AFTIS8	3	4
AFTIS9	3	4
AFTIS10	3	4
AFTIS11	3	4
AFTIS12	3	5
AFTIS13	4	4
AFTIS14	4	4.5
AFTIS15	4	5
AFTIS16	5	3.5
AFTIS17	5	4
AFTIS18	5	4
AFTIS19	5	4
AFTIS20	5	5
AFTIS21	5	5
AFTIS22	6	4
AFTIS23	6	5
AFTIS24	7	4.5
AFTIS25	8	4
AFTIS26	12	3
Mean	4.29	4.23
Standard deviation	2.39	1.60
Range	1 to 12	3 to 5
NON INTERPRETING SCHOLARS		
AFNIS1	0	4
AFNIS2	1	5
AFNIS3	5	4
AFNIS4	12	5
Mean	4.5	4.5
Standard deviation	4.72	1.85
Range	0 to 12	4 to 5
Mean for all non-interpreters	4.33	4.27

Table 2: Number of e/o's reported and fidelity ratings in the auditory groups

number of e/o's reported	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	24	33
VISUAL GROUPS																				
PROFESSIONAL INTERPRETERS																				
Number	1	2	4	4	2	1	3	2	3	1		2		1					1	1
Percentage	4%	7%	14%	14%	7%	4%	11%	7%	7%	4%		7%		4%					4%	4%
DOCTORAL STUDENTS																				
Number		1		2		1					3	2				2		1		
Percentage		8%		17%		8%					25%	17%				17%		8%		
TRANSLATION TEACHERS																				
Number			1		1	1		2	1	1						1				
Percentage			12.5%		12.5%	12.5%		25%	12.5%	12.5%						12.5%				
AUDITORY GROUPS																				
PROFESSIONAL INTERPRETERS																				
Number	2	4	2	4	1	3	1	1												
Percentage	11%	22%	11%	22%	5.5%	17%	5.5%	5.5%												
TRANSLATION AND INTERPRETATION STUDENTS																				
Number		2	3	7	3	6	2	1	1				1							
Percentage		8%	11%	27%	11%	23%	8%	4%	4%				4%							
NON INTERPRETING SCHOLARS																				
Number	1	1				1							1							
Percentage	25%	25%				25%							25%							
Total number of subjects:	4	10	10	17	7	13	6	6	5	2	3	4	2	1	0	3	0	1	1	1
(out of n = 96)																				
Table 3: Number of e/o's reported in the visual and auditory groups																				

Mean fidelity rating						
Number of e/o's	**Visual interpreters**	**Auditory interpreters**	**Visual non-interpreters**	**Auditory non-interpreters**	**Mean**	**n**
	(VFPI, n=15)	(AFPI, n=18)	(VFDS, n=12)	(AFTIS, AFNIS, n=30)		
0	4	4.5		4	4.25	4
1	4	4.25	5	4.33	4.33	9
2	4.25	4.25		5	4.57	8
3	3.5	4.8	4	4.14	4.25	15
4		4		4.5	4.37	4
5	4	4	4	4.21	4.125	12
6	4	4		4.5	4.2	5
7		4		4.5	4.25	2
8	3.83			4	3.875	4
9	4				4	1
10			3.5		3.5	3
11	3		3		3	3
12			4		4	2
13						
14						
15			3.5		3.5	2
16						
17			3		3	1

Table 4: e/o numbers and fidelity ratings