PERSPECTIVES
https://doi.org/10.1080/0907676X.2022.2088296

Routledge
Taylor & Francis Group

Interpreting accuracy revisited: a refined approach to interpreting performance analysis

Anne Catherine Gieshoff and Michaela Albl-Mikasa

ZHAW – Zurich University of Applied Sciences, Winterthur, Switzerland

ABSTRACT

Interpreting accuracy is one of the most commonly used indicators of cognitive demands in experimental interpreting studies. One possibility to assess interpreting performance is to analyse interpreting accuracy based on meaning units. The methodological approaches used thus far, however, have some drawbacks: (a) they are limited to an assessment of sense consistency with no indication of the logical cohesion of the rendition, (b) they do not take into account the difference between unintended and strategic omissions or, more generally, the prioritization of source speech information as an interpreting strategy, and (c) they do not allow for the observation of fluctuations of cognitive load or effects of fatigue. In this article, we will present a refined approach to unit-based accuracy analysis that may contribute to solving the issues mentioned above. The new method will be illustrated by means of an example data set from a larger project consisting of the renditions of ten professional and ten student interpreters. It will also include relevant statistical analyses.

ARTICLE HISTORY
Received 14 December 2021
Accepted 7 June 2022

KEYWORDS
Interpreting; propositional analysis; performance assessment; expertise; accuracy

Introduction

Performance is key in conference interpreting and of very practical relevance for conference interpreters themselves (think accreditation tests, exams, professional ethos, etc.). But apart from its practical application, interpreting performance is also of interest for (quasi-) experimental and, in particular, cognitive interpreting studies, a field which has gained popularity in interpreting studies over the last thirty years (Gieshoff et al., 2022; Olalla-Soler et al., 2020). Interpreting performance has been suggested as an important pillar in assessing cognitive demands in interpreting. It is usually assumed to decrease with higher cognitive demands (Chen, 2017). One method for the evaluation of interpreting performance in experimental studies is propositional analysis or, more generally, an analysis based on meaning units (see Chen, 2017; Dillinger, 1990; Gieshoff, 2021; Hild, 2015; Jesse et al., 2000; Tommola & Lindholm, 1995). This method operationalizes interpreting performance in terms of sense consistency with the source text which, in turn, has been found to be among the most important

CONTACT Anne Catherine Gieshoff annecatherine.gieshoff@zhaw.ch
This article has been republished with minor changes. These changes do not impact the academic content of the article.

quality criteria for both interpreters (Zwischenberger, 2010) and listeners (Kurz, 2002). However, as pointed out by Setton and Motta (2007), it seems that common standards and practices for conducting unit-based accuracy analysis are lacking to date, which makes it difficult to compare different studies[1] or to apply methods that have previously been described in the literature. In the following, we will therefore present a method for conducting a unit-based accuracy analysis in simultaneous interpreting that may serve as a reference for future studies. The method will be exemplified by means of a data set that compares interpreting accuracy among two groups, (10) professional interpreters and (10) students of interpreting, all interpreting the same speech. Relevant statistical analyses will be incorporated. Although, in this case, the new method is applied to simultaneous interpreting from English to German, we suggest it would also work for consecutive interpreting and other language pairs.

Before presenting the newly developed method, we will discuss different approaches used for unit-based accuracy analysis to date.

Literature review

The body of publications suggests that unit-based accuracy analysis is a fairly common method to assess sense consistency in simultaneous interpreting (see for instance Chen, 2017, 2020; Dillinger, 1990; Gieshoff, 2021; Hild, 2015; Jesse et al., 2000; Tommola & Lindholm, 1995; Wang & Fang, 2019). The basic procedure is to divide the source speech into propositions or small meaning units and to check whether each unit is present in the target speech. The target speech is usually transcribed and subsequently compared, unit by unit, against the source speech, with a view to the extent to which each source speech unit is reflected in the target speech (for a description of the procedure see Hild, 2015; Liu et al., 2004; Tommola & Lindholm, 1995). With some exceptions (see Dillinger, 1990; Korpal & Stachowiak-Szymczak, 2018; Wang & Fang, 2019), units are simply deemed either acceptable or unacceptable. The number or percentage of correctly rendered propositions is then compared across different conditions or used as dependent variable in statistical tests. Researchers using larger scales (for instance 0, 0.5, 1) seem to add up single scores and compare the total scores of two groups (Korpal & Stachowiak-Szymczak, 2018).[2]

One advantage of unit-based accuracy analysis seems to be its reliability: Tommola and Lindholm (1995) report an inter-rater reliability with two raters of 0.98, Hild (2015) reports an agreement of 90% between the researcher and five different raters who re-evaluated a total of 25% randomly sampled passages of the renditions.[3] In Liu, Schallert and Carroll's study, inter-rater reliability was 0.79. This rather rigorous approach, however, also has some drawbacks. One of the drawbacks is that this method considers omissions for strategic purposes to an insufficient degree. As noted by Tommola and Lindholm (1995, p. 130), interpreters may omit or condense redundant or minor propositions to render the source speech in an optimal way. With simple unit-based accuracy analysis, the use of such strategies is penalized as the corresponding propositions are not incorporated. An approach which would more readily take interpreting strategies into account may be to categorize the units according to their importance. In this way, redundant or irrelevant units can be discarded without negative effects on the propositional evaluation. An example is provided by Liu et al. (2004) who divided critical

sentences into idea units and assigned them to one of two categories: essential and secondary units. The authors observed that essential units were more likely to be rendered correctly than secondary propositions. Another drawback is that unit-based accuracy analysis is limited to sense consistency irrespective of the setting in which the target speech is presented or its overall cohesion. However, logical cohesion within the target speech is an important quality criterion for conference interpreters (Zwischenberger, 2010) and users (Kurz, 2002) and should therefore be accounted for in laboratory experiments. In an attempt to take logical cohesion into account, Gieshoff, in her 2021 study, categorized meaning units into five groups: core and secondary information, repetitions, fillers and co-text information. This last category contained information about logical connectors and speech structure. This categorization allowed her to see to what extent the logical cohesion of the source speech was maintained in the target speech.

Despite the significant advances that have been made in unit-based accuracy analysis, there are still a few pitfalls that remain unaddressed. First, it seems that most assessments were made based on transcripts of the source and target speeches (Dillinger, 1990; Hild, 2015; Jesse et al., 2000; Liu et al., 2004; Tommola & Lindholm, 1995; Wang & Fang, 2019) and it is unclear whether prosodic elements were transcribed. Neglecting prosody and intonation, however, may lead to a distorted evaluation of the target text, as prosody may compensate for missing words (Kalina, 2015, p. 17). Even though Kalina (2015) does not mention specific examples, it is easy to imagine that adverbs like 'very' can be compensated for quite effectively by stressing the word in question. Indeed, studies indicate that interpreters effectively use intonation patterns to indicate sentence borders or form a meaning unit (Ahrens, 2004). With this in mind, basing this type of analysis on audio recordings of the renditions rather than transcripts may more adequately reflect the type and amount of information that is effectively transmitted or at least intelligible for listeners.

Another aspect that – to the best of the authors' knowledge – has not yet been addressed in unit-based accuracy analysis is the evolution of sense consistency over the time course of a speech. This, however, can be very interesting in showing effects of fatigue or to find specific passages within the source text that co-occur with inaccurate renditions. Existing approaches to obtaining information about sense consistency at different time points mostly seem limited to sampling interpretations at different moments throughout a conference day (see Moser-Mercer et al., 1998). Even though not impossible, it seems rather difficult to implement such an approach in a meaningful manner under (laboratory) experimental conditions where the duration of source speeches tends to be rather short compared to the usual length of booth turns in simultaneous interpreting.[4] Therefore, it might be interesting to consider a time-resolved approach to unit-based accuracy analysis, i.e. to obtain multiple observations at many different time points of a rendition.

A new approach to propositional analysis

The method we will present in this section was developed as part of the CLINT project (*Cognitive Load in Interpreting and Translation*).[5] The project primarily investigates whether non-standard language input increases the cognitive burden of interpreters.

As interpreting performance is presumed to be indicative of cognitive demands in interpreting (Chen, 2017), we decided to include this measure in our analyses. At the same time, the following aspects had to be accounted for in the performance analysis:

- The method should be suited for different types of statistical analysis, such as regression and group comparisons, and triangulation with other types of data, including the use of a 'performance score' as a predictor variable. We hypothesized that, for instance, differences in physiological response or gaze behaviour may also be related to differences in performance, the testing of which required quantitative indicators. In order to test this hypothesis, quantitative indicators were needed.
- The method should be suited for the source texts that were used and the variables of interest, i.e. non-standard language input as well as different levels of expertise (Ehrensberger-Dow et al., 2020). Differences of expertise can be addressed by holistic ratings or existing approaches of unit-based accuracy analysis since they only require group comparisons which are possible with only one observation per rendition. But non-standard language input is challenging as a variable because it affects the whole speech but not in a uniform manner. In other words, the degree to which it affects the source speech is not the same across the whole text. While some passages may be almost native-like, others may be unconventional and more difficult to understand. For this reason, it seemed crucial to take the time course of the speech into account and to obtain time-resolved data, making multiple observations at many different time points of a rendition. Existing approaches and in particular holistic ratings did not seem to offer this possibility since the rater's assessment is summarized in one single value per rendition instead of multiple observations per rendition. Such a time-resolved approach may also be suitable for the investigation of fluctuations of cognitive load more generally (see Gile, 2008).
- At the same time, the method had to take account of the oral nature of interpreting and the fact that information from the source speech may be transmitted by prosodic cues. Likewise, the use of interpreting strategies, such as restructuring, chunking, omitting redundancies or less important elements, or stalling, should not be penalized.

As outlined above, a unit-based accuracy analysis satisfies many of these requirements by proceeding unit by unit. The order of these units can be used to reflect the timeline of the source text. Similarly, the requirement of a quantitative assessment can be solved easily, especially in the case of the simple use of 'acceptable' or 'unacceptable'. In statistical terms, this corresponds to Boolean values, i.e. a distinction between True/1 and False/0, which can be conveniently used in generalized regression models. It can also be used to calculate total scores for each rendition, making it suitable as a predictor and for group comparisons. To account for prosody and to avoid penalizing the use of interpreting strategies, the use of audio recordings of the renditions for assessment as well as a weighing system that considers the role of each unit in the speech are best suited. Less important units can be weighted accordingly so that they can be omitted without negatively impacting the analysis results. With these possibilities and requirements in mind, we opted for a refined approach to unit-based accuracy analysis. The next section explains the preparatory steps for our analysis method: the preparation of the source

speech, then the categorization and weighing of each unit and, finally, the preparation of the assessment procedure itself.

Source text preparation

Preparing the source text essentially encompasses segmenting the text into units and categorizing each unit. As suggested by Tommola and Lindholm (1995), we used the guide to propositional analysis developed by Bovair and Kieras (1985) for the segmentation of the source text. The segmentation procedure was done by two researchers who discussed each entry into the segmentation table until agreement was reached. In a first step, the main verb and its arguments were extracted, because they form the core unit or the 'backbone' of the sentence. Further units were formed for:

- Modifiers like adverbial structures (especially adverbials of time and place), adjectives or adverbs
- Logical connections and conjunctions
- Meta-discourse elements and idiomatic expressions
- Filler words to account for the oral nature of the source text. Strictly speaking, these units are not propositions, but it seemed interesting from a methodological point of view to separate these units from the remaining units in order to have the possibility to investigate these units separately and to obtain insights about how interpreters prioritize information.

However, we also found some special cases not covered by the list, such as composite verb structures (want to do, try to do, can do, have been, etc.), which are common in English, French and many other languages, or multi-word units, such as technical or idiomatic expressions. These structures needed to be kept together as one unit due to differences in the morphological rules of the target language yielding different structures for correct and complete renditions. As an example: 'it can be shown that' could be rendered in German quite literally as 'es kann gezeigt werden, dass' (it can be shown that) or more freely 'es ist möglich, Folgendes zu zeigen' ('it is possible to show the following'), etc. In other languages, such as Turkish one single word may represent the whole structure ('gösterilebilir'). Therefore, it seemed more appropriate to keep these structures together. Moreover, we adopted a special procedure for numbers because generalization is a well-known strategy in dealing with such problem triggers in simultaneous interpreting, e.g. under pressure, interpreters may choose to round up the number and give only the order of magnitude instead of the exact figure. In this case, the unit would normally be deemed unacceptable in the assessment since the exact figure was not rendered. This seemed unjustified because often the order of magnitude is sufficient to preserve the propositional content and the argumentation of the source speech. For this reason, two units were used to deal with each number: the first unit included the core proposition with the approximated number or the order of magnitude, the second the exact number.

Example: 'The data set included 9752 items'.
Unit 1: The data set included [over 9000] items.
Unit 2: 9752 [items]

In this example, interpreters who rendered the exact number (unit 2) and the proposition (unit 1) correctly, could be said to have interpreted both units correctly whereas an interpreter who rendered the proposition correctly in general, but rounded the number to over 9000, could be said to have interpreted the first unit correctly, but not the second one. A similar approach can be adopted for proper names. Finally, we deleted disfluencies, false starts, and repairs of the authentic speech which served as our stimulus text during propositional unit segmentation, as these elements should be ignored by the interpreter.

Categorization and weighing of the source text

The next step consisted in assigning a category to each unit according to its role in the source text. This categorization was based on Gieshoff (2021) and then refined. Cohesive elements, i.e. logical connectors signalling conjunction, reference or lexical cohesion, and meta-discourse elements pointing to the organizational development of the source text, while summarized in Gieshoff's (2021) classification as 'co-text information', were treated as two distinct categories since logical cohesion is a highly rated quality parameter for conference interpreters (Zwischenberger, 2010). The resulting categories are described in Table 1.

Two researchers who had a very good understanding of the source text performed the categorization of the units independently of one another and subsequently discussed all of the units they had rated differently until agreement was reached. Once the

Table 1. Description and weight of categories.

Category		Description	Weight
Content	Core information	Core information is (new) content information that is essential in understanding the gist or the main ideas of the source text. It often remains relevant throughout the source text with speakers referring back to it.	10
	Secondary information	Secondary information is (new) content information that is not part of the main idea or central information. While it is not essential in understanding the main line of argumentation, it does provide new information. Typical examples are insertions or subordinate clauses (even though these can also be core information in some cases).	5
	Redundant elements	This category contains content information that is repeated or redundant.	1
Structure	Cohesive elements	Cohesive elements contain mostly logical connectors and conjunctions (*because*, *while*, *yet*), but also demonstrative pronouns (*those, this*) that reference back to previous units. In exceptional cases, logical connections were implicit, rather than verbalized. In these cases, we introduced an explicit unit for the implicit logical connection in order to be able to establish whether the logical relationship was maintained in the target text.	9
	Meta-discourse elements	Meta-discourse elements contain information about the discourse structure. Typical examples are *for example, first/second, this takes us to*, *in conclusion*. In these cases, many different renditions are possible and sometimes they need to be adapted to the target culture.	7
Miscellaneous	Modifiers	Modifiers include, in particular, modifying adverbs such as *very* or *not really*. In our study, modifiers were sometimes rendered by prosodic cues.	3
	Fillers	The category of fillers comprises all words that do not seem to carry any content or to ensure logical cohesion. Instead, they rather reflect a way of speaking (*kind of*, *well, okay* …).	0

categorization was finalized, a weighing score was assigned to each category. The weighing was meant to reflect the importance of the category in the source text. Rendering core information, for instance, is consequential for obtaining a target text which is consistent with the source text, whereas redundancies may be left out without affecting the content. Instead of a simple ranking, we decided to use a scale from 0 to 10 for the weighing. The reason for this decision was that two categories can be of similar importance and should therefore receive a similar weight whereas others may be more different in relevance which should be reflected by a larger scale interval. Weighing scores for each category were suggested by three researchers independently of one another. The ranking of the categories was exactly the same, with minor differences in the weighing score, which were resolved in joint discussion. The result of this discussion is displayed in Table 1.

Assessment procedure

In order to prepare the accuracy assessment, and based on Bovair and Kieras (1985) we represented each meaning unit in an abstract form (see Table 2 below). The purpose of this was to avoid a 'literality' bias of sorts: direct comparison of the renditions with the source text carried the risk of form-based renditions scoring better in our assessment than meaning-based ones, which did not reflect the word choice or sentence structure of the source text. For the abstract form, content information (core, secondary, redundant) was rewritten in the form of simplified sentences with time or modal markers where appropriate. Structural information (cohesive and meta-discourse elements) was represented in capitals (see examples in Table 2).

This abstract representation form was a convenient reference for the assessment of the renditions' sense consistency. It was the basis on which we listened to the recording, ticking off for each unit whether its meaning was comprehensible and consistent with the source text (1) or not (0). In cases in which the assessment was not obvious (for instance when interpreters corrected themselves) comments were entered in a comments section. Neutral additions, changes in order and chunking were not penalized as they can reflect interpreting strategies. In other words, if the participant added neutral formulations or paraphrased the unit, the rendition of the corresponding unit was still accepted

Table 2. Two example sentences from the source speech, segmented into different units.

Unit identifier	Example sentence	Representation
52	And the idea was that	INTENT
53	we combined these technologies with	INTENT: combine technologies
54	insights	INFO on combine: with insights
55	from social sciences	INFO on insights: from social sciences
56	in order to	FINAL
57	create a system	create system
58	that is effective.	INFO on system: effective
476	Unfortunately	REGRET
477	we cannot tell you	we no information
478	whether it's successful	INFO on information: success
479	or not	INFO on information: no success
480	because	CAUSAL
481	we are just at (the end of) the first step	REASON: current stage = first step
482	end of (first step)	INFO on step: end of

(1). Similarly, we did not consider disfluencies or minor language mistakes as long as the rendition was intelligible and undisrupted. Disfluent renditions were only rejected when false starts, corrections or mumbling made them unintelligible.

In the next section, we will exemplify the method described above with a data set of 20 interpretations that were recorded as part of the CLINT project.

Illustration of the method

As mentioned above, the data was collected as part of the CLINT project. Since expertise and professional experience were expected to influence the way interpreters coped with non-standard language input, study participants included both a cohort of professionals and a cohort of students. For the purpose of illustrating the new accuracy method, the data set presented here uses expertise as an independent variable. As demonstrated in multiple studies, expertise positively affects interpreting performance and, in particular, completeness and sense consistency (Dillinger, 1990; Hild, 2015; Liu et al., 2004; Rosendo & Galván, 2019). We therefore expected similar findings in our data set.

Participants and source speech

The data set compares the recordings of ten professional and ten student interpreters who simultaneously interpreted the same speech from English into German. The professional interpreters' experience ranged from 2 to 37 years ($MD = 20.5$). The students had had at least one and a half semesters of training in simultaneous interpreting. Each of them signed an informed consent form and filled in a questionnaire about their professional and linguistic background before participating in the experiment. The speech was a 725 words long conference talk on the topic of mobility. It was rather generic with only 4.0% of the words not among the 5000 most frequent words in American English (Davies, 2008). It was read out by a native speaker of General American English and recorded on video. The video was 12:05 min long and the delivery rate was 204 syllables per minute. Table 3 displays the number of units for each category in the source speech.

Procedure

Participants were tested individually in a simulated interpreting setting. They were seated in front of a computer screen and guided through the experiment with instructions in German, their native language, which appeared on the computer screen. Participants interpreted a short warm-up speech (5 min) to familiarize themselves with the

Table 3. Number of units per category in the source speech.

		Number of segments	Percentage
Content	Core information	184	37.5%
	Secondary information	138	28.5%
	Redundant elements	22	4.3%
Structure	Cohesive elements	47	9.6%
	Meta-discourse elements	50	10.5%
Miscel-laneous	Modifiers	36	7.0%
	Fillers	11	2.3%
	Total	488	100%

Table 4. Level of agreement, Cohen's kappa, *z*-value and *p*-value for inter-rater und intra-rater reliability.

	% agreement	Cohen's kappa	*z*-value	*p*-value	Interpretation
Inter-rater	88%	0.735	72.6	>0.001	Good
Intra-rater	94%	0.873	86.3	>0.001	Very good

equipment, before interpreting the source speech. Participants listened to the speech through headphones and could adapt the volume at all times.

Data preparation

Participants' interpretations were recorded and assessed according to the procedure described above by raters independently from each other. The second rater, a trained translator, did not receive any particular training for this assessment in order to see how intuitive the rating procedure is. The first rating as well as the comments section were hidden in the table to avoid biasing the second rater. Additionally, the first rater re-assessed all interpretations after three to twelve months in order to calculate not only inter-rater reliability between both raters, but also the intra-rater reliability. The level of agreement used by Hild (2015), as well as Cohen's kappa used by Liu et al. (2004) are reported in Table 4. The interpretation of these values is based on Wirtz (2021).

Statistical analyses and visualizations

The following sections present different visualizations and statistical analyses allowed for by the refined approach to propositional analysis. All plots and analyses were done in R (R Core Team, 2020) with the packages ggplot (Wickham, 2016) and tidyverse (Wickham et al., 2019). The R-script is available under: https://github.com/ac-gieshoff/interpreting-accuracy.

Total score

A first step consists in comparing the total score of the two groups, professionals and students. The total score can be calculated as follows:

1) Multiply the rating (0/1) with the weight to obtain the points for each unit.
2) Add up all points for each participant.

Figure 1 shows the total scores of students' and professionals' renditions. A Wilcoxon signed-ranks test confirms the visual impression in Figure 1 that professional interpreters obtained higher scores ($MD = 2601$) than students of interpreting ($MD = 2142$, $Z = 2.79$, $p < .01$, $r = .625$). The total scores can also be converted into ratio scores (percentages) by dividing the total score of each participant by the maximum score that could be reached ideally over the whole source text. This second option also allows for the comparison of interpreting accuracy across different source texts (for instance comparing renditions of a technical text with a more general one).

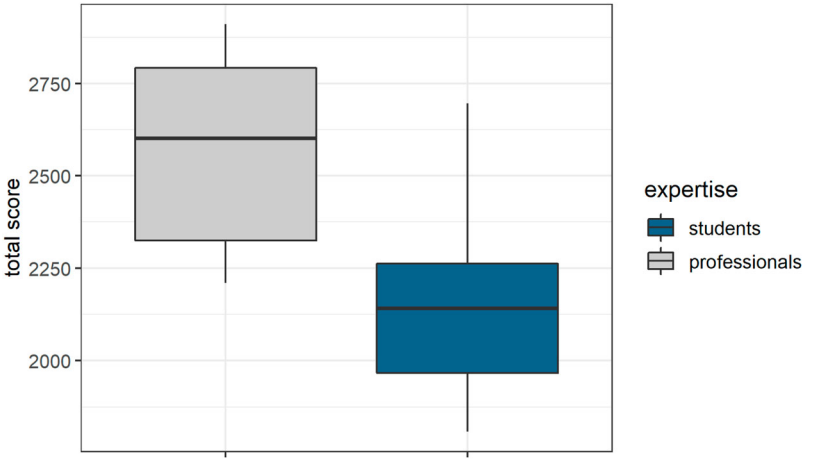

Figure 1. Total scores of professionals (grey) and students (blue).

Prioritization of information

In a second step, the method allows us to look at differences in prioritization of information. Are professionals more efficient in filtering and processing relevant units than students? To investigate this question, the number of correct renditions per category may provide useful information. If professionals are more efficient than students, we may expect that they are more successful in rendering categories of high relevance, such as core information or cohesive elements. These differences may be less pronounced in less relevant categories, such as redundant elements or fillers, as these may be omitted by professionals and students alike. Pairwise paired t-tests over each category with Bonferroni correction indicate significant differences between students' and professionals' renditions mostly in relevant categories (core and secondary information, cohesive and meta-discourse elements), whereas no significant differences are found in less relevant categories, such as redundant elements, modifiers and fillers (see Table 5 and Figure 2).

Extracting low accuracy units in the source text

For a thorough investigation of interpreting accuracy, it can be useful to check whether some units were less successfully rendered than others. Such an investigation highlights

Table 5. Statistics and effect size (Cohen's d) for pairwise paired t-tests comparing students and professionals over each category.

Category	*t*-value	Degrees of freedom	Adjusted *p*-value	Cohen's *d*
Core information	3.04	9	.014	0.96
Secondary information	3.15	9	.012	0.99
Redundant elements	1.81	9	.10	
Cohesive elements	3.62	9	.006	1.15
Meta-discourse elements	2.33	9	.045	0.73
Modifier	2.03	9	.072	
Filler	2.04	9	.072	

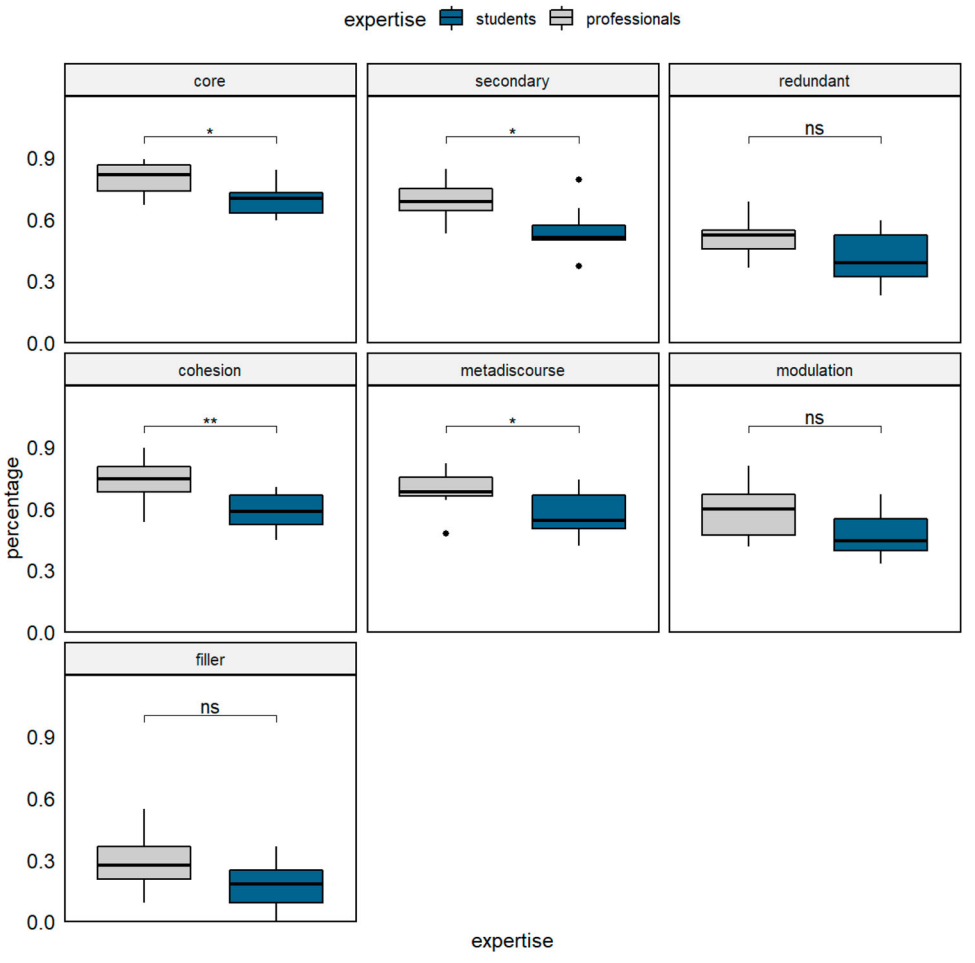

Figure 2. Comparison of differences between professionals (grey) and students (blue) in each category.

particularly challenging units and may also serve as a countercheck as to whether the categorization is meaningful. To do so, we aggregate the number of correct renditions per unit and extract outliers. In our data set, no outliers were observed. Another approach is to extract the units with an accuracy below the first quartile. In the professional group, for instance, the first quartile corresponds to 5 correct renditions per unit; units below this level may be described as 'low accuracy units'. The number of low accuracy units in the data set seems quite high at first: 122 units were not or incorrectly rendered more than 5 times. However, only 33 of them were core information. This analysis can be helpful in assessing whether the categorization was meaningful and justified or whether, retrospectively, some units should be assigned to another (less relevant) category. As for our data set, this procedure helped us to establish that the initial categorization was valid given that the first quartile of the accuracy scores is already rather high (with half the participants), so that these units do not appear to be particularly problematic.

Observing fluctuations of cognitive load and effects over time

Finally, the new approach to propositional analysis offers the possibility to look at fluctuations of cognitive load as well as effects over time, such as effects of fatigue. A relevant question to ask is whether students tire sooner than professionals. In that case, we may expect to find a higher number of incorrectly rendered units towards the end of students' renditions compared to those of professionals. This question can be addressed by conducting linear generalized regressions on the rating (0/1) as the dependent variable and using an identifying number as the time variable. The identifier is a number ranging from 1 (first unit) to the last unit of the source text (in our case 488) and reflects the order of appearance of each unit in the source text. Thus, it is not a completely accurate reflection of the timeline – this would require adding the exact time point of each unit in the source text – but it can still indicate where in the source text drops in performance tend to occur. Figure 3 displays the loss of information in terms of decreasing accuracy in students' and professionals' renditions over time. The red line corresponds to an ideal rendition in which no information is lost. The blue and grey lines correspond to one student (blue) or professional (grey) interpretation each. The loss in information is calculated by dividing the difference between the points lost so far and the optimal score that can be reached at that time point by the maximum score over the whole rendition.

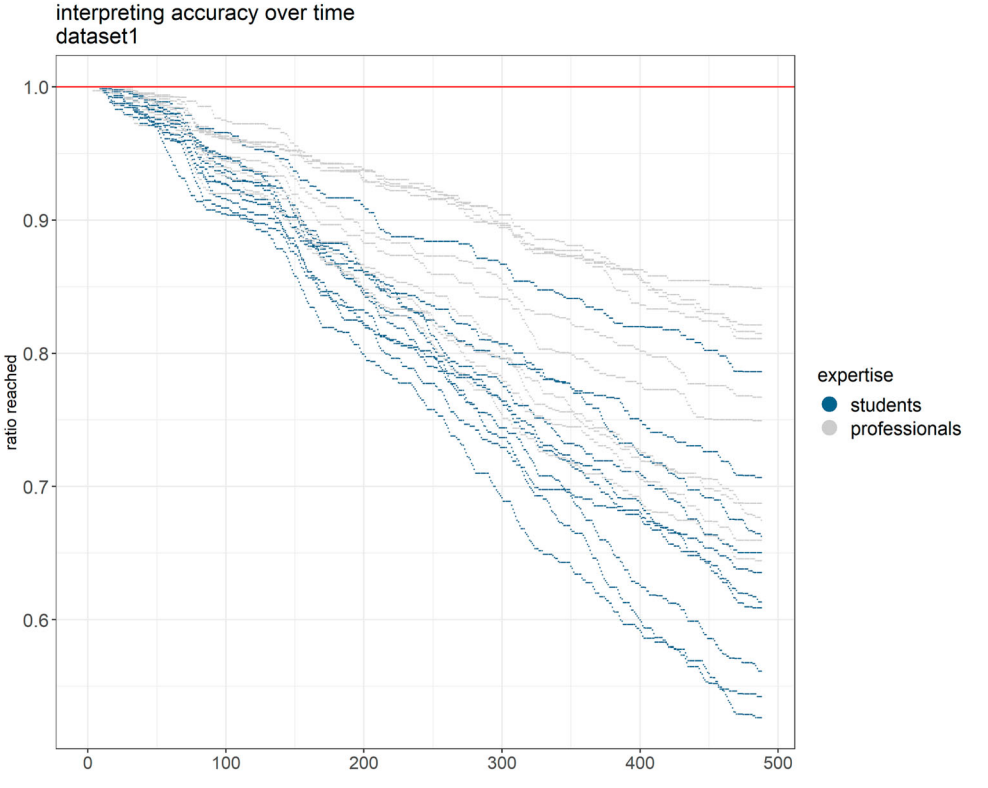

Figure 3. Loss of information over time in professionals' (grey) students' (blue) renditions.

Table 6. Statistics and effect size (Cohen's d) for pairwise paired t-tests comparing students and professionals over each passage.

Passage	*t*-value	Degrees of freedom	Adjusted *p*-value	Cohen's *d*
P1_start	4.09	9	.003	1.29
P2_idea	3.04	9	.014	.96
P3_details	2.34	9	.044	.73
P4_impact	2.27	9	.049	.71
P5_end	4.50	9	.001	1.42

In practice, it can be difficult to detect an effect of fatigue because some text passages may be easier than others, in particular, the beginning and the end of a speech may be more formulaic and generic than passages in the middle where speakers usually develop their ideas and go into more detail. This means that interpreting accuracy, i.e. the number of correct renditions, will not decrease uniformly over the time course of the speech. Hence, it may be more informative to divide the source text in different passages and to compare the score obtained for each passage. In our case, the differences between students and professionals are already well established and further analyses seem unnecessary. However, for illustrative purposes, we divided the source text into five passages with a mean length of 307 words per passage ($SD = 117$ words) and conducted pairwise paired t-tests for each passage. Unsurprisingly, this analysis revealed significant differences between students and professionals in each text passage (Table 6).

This approach has one drawback: In our case, the division into text passages is – although based on the content – to a large extent arbitrary and still rather rough. An alternative, more fine-grained approach may be to conduct generalized additive mixed regressions. Additive models can not only describe linear or exponential patterns, but also non-linear ones, i.e. 'wiggly' patterns (Wieling, 2018) and may therefore be particularly suited to modelling interpreting accuracy (for an example of additive modelling in interpreting studies see Plevoets & Defrancq, 2018).

In our case, the 0/1 rating can be used as the dependent variable and expertise as the independent variable. The time course can be modelled with the identifier of each item. The model presented below (see Figure 4) was fitted with the R-package mgcv (Wood, 2017) and plotted with itsadug (van Rij et al., 2020). It included random non-linear patterns for each participant to allow for individual variation over time, as well as random intercepts for each category because some categories, such as core information, have a higher probability of being rendered correctly than others that are less important (fillers, modifiers, redundant elements). Expertise and a non-linear pattern of expertise over time were tested as fixed effects (see Table 7). The model, however, seems to miss some of the variance, as the deviance is not very close to the residual degrees of freedom: 11165.25 and 9659.76.

A closer look at Figure 4 shows first of all that interpreting accuracy is highly variable: accuracy is higher at some points and lower at others. This information may indicate fluctuations of cognitive load: where accuracy is lower, cognitive load may be higher. Second, it shows that professionals' interpreting accuracy is consistently higher than students' (in line with previous studies), but it does not always differ significantly. Where the line is plotted within the (grey-blue) overlapping part of the confidence bands, the model does not observe significant differences. This is the case, for example, at the very beginning and the very end of the speech. Where the line is plotted outside of the overlapping confidence bands, e.g. around units 220–260, the model observes significant differences

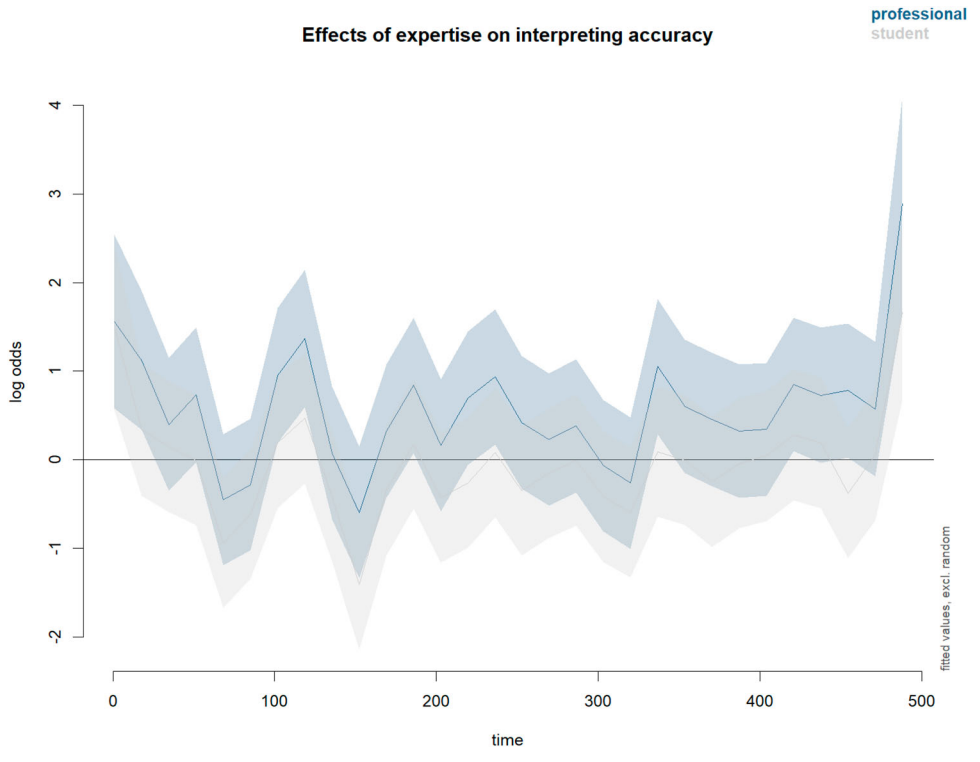

Figure 4. Additive model of rating as a function of expertise.

between professionals and students. The function plot_diff from the mgcv package outputs 'windows of significant difference' which could be used to extract those units where groups differ significantly from each other (see script available in github).

Discussion

In this article, we presented a refined approach to unit-based accuracy analysis for interpreting. The main motivations for developing this method were (1) to develop scores that are suitable for different types of statistical analysis, (2) to make it possible to observe fluctuations of cognitive load, and (3) to adopt a more ecological approach that takes into account both information prioritization by interpreters as an important strategy as well as the oral nature (and meaning conveying prosodic elements) of interpreting. The method presented here offers a binomial dependent variable that can be used to test the effects of different variables in regressions, as well as a total score that can serve either as a dependent variable for the comparison of two (or more) groups or source texts and as an independent variable for the prediction of, for instance, differences in physiological measures or gaze behaviour. An identifier for each unit that reflects the order of appearance and, hence, the time course of the speech, can be used in additive regression models to investigate fluctuations of cognitive load, but also to extract specific units, for example units with overall low accuracy, for further discussion. The categorization and weighing of units according to their role in the speech not only

Table 7. Estimates, standard error, *z*-value and *p*-value for fixed effect and fixed smooth terms.

	Estimated coefficient	Standard error	*z*-value	*p*-value
Intercept	0.504	0.338	1.491	0.136
Expertise	−0.608	0.165	−3.691	<0.0001
	Effective degrees of freedom	Reference degrees of freedom	Chi-square	*p*-value
Time*professional	28.772	33.63	251.6	<0.0001
Time*students	26.449	31.47	233.9	<0.0001

allows more insights into the prioritization of information or the maintenance of text cohesion and coherence in the target speech but, in addition, prevents an over-penalization of minor omissions or errors in the target text. Finally, the assessment was explicitly based on the audio recordings instead of a transcript in order to include oral features such as prosody (e.g. stressing a word) or pronunciation mistakes (e.g. mumbling) that make the target text unintelligible. Although a recording-based assessment seems at least in theory closer to the listener experience, it is difficult to say whether transcript-based assessments actually differ from recording-based assessments. This would require a more formal comparison of transcript-based and recording-based assessments.

The analyses of 10 professional and 10 interpreting student participants as presented in the section above confirm the effect of expertise on interpreting accuracy and are in line with existing studies (Dillinger, 1990; Hild, 2015; Liu et al., 2004; Rosendo & Galván, 2019). With a kappa-value of 0.735 (88% of agreement), the inter-rater reliability is slightly lower, but still similar to the values reported by Hild (2015, p. 90% of agreement) and Liu and colleagues who also used two raters (2004, Kappa 0.79). This is encouraging in that it suggests that the two raters reach a similar assessment when rating the same renditions even without prior training in this method. However, it should be borne in mind that inter-reliability of two raters is still a rather low number to assess the reliability of this method. For this reason, it would be advisable to use at least three different raters and to compute the inter-rater reliability before proceeding to any further analyses (see Koo & Li, 2016).

The method presented here is specifically designed for experimental purposes. It is not suited to measuring interpreting quality, as this would require incorporating further parameters, such as fluency, intonation, or use of terminology (see for instance Zwischenberger, 2010) and/or gathering listeners' experiences (Kurz, 2002). It is meant as a contribution to (quasi-)experimental research in TIS in that it may provide an interesting alternative method for researchers interested in interpreting performance measurements.

Conclusion

We have presented a new approach to unit-based accuracy analysis that differs from previous studies in that (1) it introduces categories and categorial weighing to reflect information hierarchy, (2) it offers both a total score and a binary variable suitable for regressions, and (3) it makes it possible to observe fluctuations of cognitive load thanks to an identifier for each unit. The method was tested with a data set containing renditions of the same source text by 10 professional and 10 student interpreters. The results confirm previous studies on the positive effect of expertise (Dillinger, 1990; Hild, 2015; Liu et al., 2004; Rosendo & Galván, 2019). The inter-rater reliability (kappa = 0.735, agreement: 88%) was slightly lower, but similar to previous reports on propositional analysis (Hild, 2015; Liu et al., 2004). The method was designed for (quasi-)experimental research and may contribute to enhancing common practices in TIS.

Notes

1. A similar observation has been made in TIS more generally. According to Olalla-Soler a 'frequent problem for replication indicated by the survey participants was the lack of detailed information about the research design in the original study' (2020, p. 29).

2. From the figures, it seems that Dillinger (1990) again compared the percentage of correct renditions and did not include his scale in the statistical analysis.
3. It is not clear from her description whether the inter-rater reliability was calculated between all five raters and the researcher, or in a one-by-one comparison between each rater and the researcher.
4. Gieshoff (2021) and Korpal and Stachowiak-Szymczak (2018), for instance, used speeches of four minutes' length which seems to be comparable with Dillinger's 580 words source speech (Dillinger, 1990). The stimuli used by Liu et al. (2004) were longer (between 1600 and 2100 words), but most probably still below 30 min.
5. For more information about the CLINT project see www.zhaw.ch/linguistik/iued/clint, as well as the following publications: (Albl-Mikasa et al., 2020; Ehrensberger-Dow et al., 2020).

Acknowledgements

We gratefully acknowledge the valuable input of Oleksandra Valtucheck and Katrin Andermatt during the early stages of the development process.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research has been supported by the Swiss National Science Foundation (SNSF, Sinergia grant CRSII5_173694).

Notes on contributors

Anne Catherine Gieshoff received her PhD in interpreting studies from the University of Mainz and currently holds a post-doc position at ZHAW Zurich University of Applied Sciences in the interdisciplinary SNSF Synergia project *CLINT – Cognitive load in Interpreting and Translating,* where she has been collecting first-hand experience in employing quantitative and psychophysiological methods in interpreting studies. She is a member of Translation, Research, Empiricism, Cognition (TREC), the International Association for Translation and Intercultural Studies (IATIS) and the European Society for Translation (EST). Her research focuses on cognitive load and effort, and English as a Lingua Franca in interpreting.

Michaela Albl Mikasa is Professor of Interpreting Studies at the Institute of Translation and Interpreting of ZHAW Zurich University of Applied Sciences in Switzerland, where she teaches on both the BA and MA programmes. Her research and publications focus on ITELF (interpreting, translation and English as a lingua franca), the cognitive foundations of conference and community interpreting, note taking for consecutive interpreting, the development of interpreting expertise, and medical interpreting. She is currently a general board member of the European Network of Public Service Interpreting (ENPSIT). She is also a member of the Swiss Research Centre Barrier free Communication and principal investigator of the interdisciplinary Sinergia project Cognitive Load in Interpreting and Translation (CLINT) funded by the Swiss National Science Foundation (SNSF). She is editor, together with Elisabet Tiselius, of the Routledge Handbook of Conference Interpreting.

ORCID

Anne Catherine Gieshoff http://orcid.org/0000-0002-4383-190X
Michaela Albl-Mikasa http://orcid.org/0000-0003-0933-574X

References

Ahrens, B. (2004). *Prosodie Beim Simultandolmetschen*. 41. Peter Lang.

Albl-Mikasa, M., Ehrensberger-Dow, M., Heeb, A. H., Lehr, C., Boos, M., Kobi, M., Jäncke, L., & Elmer, S. (2020). Cognitive load in relation to non-standard language input: Insights from interpreting, translation and neuropsychology. *Translation, Cognition & Behavior*, *3*(2), 263–286. https://doi.org/10.1075/tcb.00044.alb

Bovair, S., & Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text* (1st ed., pp. 315–362). Routledge.

Chen, S. (2017). The construct of cognitive load in interpreting and its measurement. *Perspectives*, *25*(4), 640–657. https://doi.org/10.1080/0907676X.2016.1278026

Chen, S. (2020). The process of note-taking in consecutive interpreting: A digital pen recording approach. *Interpreting. International Journal of Research and Practice in Interpreting*, *22*(1), 117–139. https://doi.org/10.1075/intp.00036.che

Davies, M. (2008). *Word frequency data*. The Corpus of contemporary American English (COCA). 2008. https://www.english-corpora.org/coca/

Dillinger, M. (1990). Comprehension during interpreting: What do interpreters know that bilinguals don't? *The Interpreters' Newsletter*, *3*, 41–58. http://hdl.handle.net/10077/2154

Ehrensberger-Dow, M., Albl-Mikasa, M., Andermatt, K., Heeb, A. H., & Lehr, C. (2020). Cognitive load in processing ELF: Translators, interpreters, and other multilinguals. *Journal of English as a Lingua Franca*, *9*(2), 217–238. https://doi.org/10.1515/jelf-2020-2039

Gieshoff, A. C. (2021). Does it help to see the speaker's lip movements?: An investigation of cognitive loadand mental effort in simultaneous interpreting. *Translation, Cognition & Behavior*, *4* (1), 1–25. https://doi.org/10.1075/tcb.00049.gie

Gieshoff, A. C., Lehr, C., & Heeb, A. H. (2021). Stress, cognitive, emotional and ergonomic demands in interpreting and translation: A review of physiological studies. *Cognitive Linguistic Studies*, *8*(2), 404–439. https://doi.org/10.1075/cogls.00084.gie

Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *FORUM. Revue Internationale D'interprétation et de Traduction / International Journal of Interpretation and Translation*, *6*(2), 59–77. https://doi.org/10.1075/forum.6.2.04gil

Hild, A. (2015). Discourse comprehension in simultaneous interpreting: The role of expertise and redundancy. In A. Ferreira & J. W. Schwieter (Eds.), *Benjamins translation library* (Vol. 115, pp. 67–100). John Benjamins Publishing. https://doi.org/10.1075/btl.115.04hil

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *5*(2), 95–115. https://doi.org/10.1075/intp.5.2.04jes

Kalina, S. (2015). Measure for measure – Comparing speeches with their interpreted versions. In C. Zwischenberger & M. Behr (Eds.), *Interpreting quality: A look around and ahead* (p. 20). Frank & Timme.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Korpal, P., & Stachowiak-Szymczak, K. (2018). The whole picture: Processing of numbers and their context in simultaneous interpreting. *Poznan Studies in Contemporary Linguistics*, *54* (3), 335–354. https://doi.org/10.1515/psicl-2018-0013

Kurz, I. (2002). Conference interpreting: Quality in the ears of the user. *Meta*, *46*(2), 394–409. https://doi.org/10.7202/003364ar

Liu, M., Schallert, D. L., & Carroll, P. J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*, *6*(1), 19–42. https://doi.org/10.1075/intp.6.1.04liu

Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (pilot study). *Interpreting. International*

Journal of Research and Practice in Interpreting, *3*(1), 47–64. https://doi.org/10.1075/intp.3.1.03mos

Olalla-Soler, C. (2020). Practices and attitudes toward replication in empirical translation and interpreting studies. *Target. International Journal of Translation Studies*, *32*(1), 3–36. https://doi.org/10.1075/target.18159.ola

Olalla-Soler, C., Franco Aixelá, J., & Rovira-Esteva, S. (2020). Mapping cognitive translation and interpreting studies: A bibliometric approach. *Linguistica Antverpiensia*, *19*, 25–52.

Plevoets, K., & Defrancq, B. (2018). The cognitive load of interpreters in the European parliament: A corpus-based study of predictors for the disfluency uh(m). *Interpreting. International Journal of Research and Practice in Interpreting*, *20*(1), 1–32. https://doi.org/10.1075/intp.00001.ple

R Core Team. (2020). *R: A language and environment for statistical computing (version 3.6.3)*. https://www.R-project.org

Rosendo, L. R., & Galván, M. C. (2019). Coping with speed: An experimental study on expert and novice interpreter performance in the simultaneous interpreting of scientific discourse. *Babel. Revue Internationale de La Traduction / International Journal of Translation*, *65*(1), 1–25. https://doi.org/10.1075/babel.00081.rui

Setton, R., & Motta, M. (2007). Syntacrobatics: Quality and reformulation in simultaneous-with-text. *Interpreting. International Journal of Research and Practice in Interpreting*, *9*(2), 199–230. https://doi.org/10.1075/intp.9.2.04set

Tommola, J., & Lindholm, J. (1995). Experimental research on interpreting: Which dependent variable? In J. Tommola (Ed.), *Topics in interpreting research* (pp. 121–133). University of Turku.

van Rij, J., Wieling, M., Baayen, R. H., & von Rijn, H. (2020). *Itsadug: Interpreting time series and autocorrelated data using GAMMs (version 2.4)*.

Wang, J., & Fang, J. (2019). Accuracy in telephone interpreting and on-site interpreting: A comparative study. *Interpreting. International Journal of Research and Practice in Interpreting*, *21*(1), 36–61. https://doi.org/10.1075/intp.00019.wan

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Use R!. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*(September), 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wirtz, M. (2021, June 3). Cohens Kappa. In M. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie*. Hogrefe. https://dorsch.hogrefe.com/stichwort/cohens-kappa

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.) Chapman & Hall/CRC Texts in Statistical Science. CRC Press/Taylor & Francis Group.

Zwischenberger, C. (2010). Quality criteria in simultaneous interpreting: An international vs. a national view. *The Interpreters' Newsletter*, *15*, 127–142. http://hdl.handle.net/10077/4754