Language Testing in Asia

RESEARCH Open Access

Calibrated Parsing Items Evaluation: a step towards objectifying the translation assessment

Alireza Akbari* and Mohammadtaghi Shahnazari

* Correspondence: akbari.transtech@hotmail.com
Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran

Abstract

The present research paper introduces a translation evaluation method called Calibrated Parsing Items Evaluation (CPIE hereafter). This evaluation method maximizes translators' performance through identifying the parsing items with an optimal p-docimology and d-index (item discrimination). This method checks all the possible parses (annotations) in a source text by means of the Brat Visualization Stanford CoreNLP software. CPIE takes a step towards the objectification of translation assessment by allowing evaluators to assess values (impacts) of the items in source texts via docimologically justified parsing items. For this paper, 16 evaluators were recruited to score translation drafts by means of the holistic, analytic, Preselected Items Evaluation (PIE) methods and CPIE method. For the present research paper, "F-Statistics," "Probability Plot," "Spearman rho," and "Regression Variable Plot" were applied to the evaluators' translation assessments to ensure the degree of validity and reliability of CPIE compared to the holistic, analytic, and PIE methods, respectively. The results indicated that the CPIE method was more consistent and valid in terms of docimologically justified parsing items. The limitations and the possibilities of the CPIE method in web-based platforms were also discussed.

Keywords: Calibrated parsing items evaluation, P-docimology, Item discrimination, Brat visualization Stanford CoreNLP, Validity, Reliability

Introduction

To date, research on translation evaluation has mainly concentrated on theoretical and descriptive aspects and covered issues such as criteria for good translation (Newmark, 1991), the nature of translation errors versus language errors (Gouadec, 1989), possible source of translation errors (Gouadec, 1981), linguistic and pragmatic translation quality assessment (Hatim & Mason, 1997), and different levels of translation competence (Stansfield, Scott, & Kenyon, 1992).

Today, the field of translation studies (TS) has confirmed the need for experimental and empirical evidence for the evaluation and quality of translation tests (Akbari & Segers, 2017b; Anckaert, Eyckmans, & Segers, 2008). Even though educational and professional institutions have implemented the "certifications of translation skills" (Eyckmans, Anckaert, & Segers, 2013) based on various test administrations, the reliability and validity of those test administrations remain under question. That is to say,

translation evaluation is somehow associated with the codes of practice (written rules which express how a researcher/scholar must behave in a particular situation and profession) rather than experimental-empirical research across the globe (ibid.). The term "translation evaluation" refers to the translation product (e.g., the target text), translation process (i.e., the way the translator transfers the content of a source language to the target one), translation service (e.g., invoicing, client, complaints, compliance agreement), and consequently translator competence. However, translation product, process, service, and competence of a translator cannot be assessed in the same way and require various modes of evaluation approaches.

Two factors may explain the lack of test development to evaluate translation competence. Firstly, translation tests are not valid enough to measure language ability and proficiency and this caused a certain loss of popularity during the period of Communicative Approach (CA) (Widdowson, 1978). This may be due to the fact that translation tests are not subjected to the "same psychometric scrutiny as other language testing formats" (e.g., c-test and cloze test) (Eyckmans et al., 2013). The second reason illustrates the "epistemological gap" between the hard sciences (e.g. chemistry, biology, etc.) versus human sciences such as translation and interpreting studies, language and linguistics, literature, and so forth. The presupposition that it is not possible to objectify the quality of translation while covering its very essence may be very persistent among language trainers and teachers as well as translation trainers/scholars whose "corporate culture exhibits a marked reticence towards the use of statistics" (Anckaert et al., 2008, Eyckmans, Segers, & Anckaert, 2012, 2013). With this in mind, testing and training translation and interpreting skills have been more or less in the hands of practitioners rather than of translation scholars and researchers. Due to psychometric methods, a great body of research in the field of reliability and validity of language tests has been realized. However, the field of translation and interpreting studies has been lagging behind and needs more research in this respect. As stated in Akbari and Segers (2017b), p. 4, translation assessment and evaluation research are still in their infancy.

In educational and professional contexts, translation evaluation practice can be carried out in accordance with a criterion-referenced approach (Schmitt, 2005) (an approach which assesses student performance against a fixed set of predetermined criteria). Therefore, educational and professional contexts can be assessed/evaluated in terms of some "assessment grids" (a matrix including a number of error levels and error types) to make translation evaluation more valid and reliable. Nevertheless, they are unable to diminish the degree of subjectivity of translation evaluation adequately. Also, the system of scoring which is prone to be impacted by contrast effect ("a magnification or diminishment of perception as a result of previous exposure to something of lesser or greater quality") (Gonzalez 2019) threatens the reliability of a translation test.

In the context of the above, the purpose of the present research paper is finally to introduce a model of translation evaluation called "*Calibrated Parsing Items Evaluation*" (CPIE hereafter), so as to contribute to the objectification of translation assessment. The CPIE method is characterized by a total number of parses in a source text based on translation relevance and translation norm and criterion-referenced assessments. As is the case with Preselected Items Evaluation (PIE) method, correct and incorrect solutions are listed for each parse in the source text of the test in the CPIE method. The present research aims at testing the applications of the CPIE method in

two stages: (1) calculating and recalculating scores through the CPIE translation evaluation method (a case study) and (2) measuring the degree of validity and reliability of the CPIE method compared to the holistic, analytic, and PIE methods through the proposal of two hypotheses: (a) CPIE as a method of translation assessment is more valid than holistic, analytic, and PIE methods (*the question of validity*); (b) the quality of translation can be evaluated more reliably if the method of evaluation assesses all the parsing items having good p and d docimologies (norm-referenced assessment towards criterion-referenced assessment) rather than some "specific items" (PIE method), "pre--conceived criteria" (analytic method), and "impressionistic-intuitive scoring" (holistic method) among the raters (*the question of reliability*).

State of the art

A review

Translation evaluation is largely marked by a criterion-referenced assessment (Schmitt, 2005). Based on educational and professional contexts, assessment grids are used in an attempt to make translation evaluation more objective, valid, and reliable (ibid.). Even though the utilization of the assessment grids is prompted through the grader's wish to take various dimensions of translation competence into account, one must contend that they fail in reducing the "subjectivity of translation" (Anckaert et al., 2008). Besides the subjective nature of translation sub-competences, other factors may threaten the reliability of translation administration tests. Let us start with the grader's consistency throughout the task of translation scoring during a specific period of time. Not only will the system of scoring be prone to a contrast effect, it is also necessary to provide a "sound testing practice" distinguishing good items from the bad ones. Furthermore, all scores must be docimologically (theoretically testable) justifiable and the system of scoring must discriminate the average quality of translations. Therefore, researchers from the fields of translation quality research and assessment (Akbari & Segers, 2017a, 2017b, 2017c; Conde Ruano, 2005; Kockaert & Segers, 2017) are now taking up topics such as interrater (the degree of agreement among the raters) and intrarater (the degree of agreement among repeated administration of a test through a single rater) reliability, construct (the degree to which a theoretical construct can be operationalized), and ecological (results which can be utilized within real-life context) validity in support of warrantability and "situatedness" (Muñoz Martín, 2010). The purpose of the present research paper is to free translation evaluation from the "construct-irrelevant variables", i.e., uncontrolled and extraneous variables which impact the outcome assessment, arising in analytic and holistic scoring methods (Eyckmans et al., 2013).

Current translation evaluation methods in translation quality research

Holistic method

The holistic method is deemed an objective and precise method of translation evaluation (Bahameed, 2016). Based on the corrector's appreciation/taste and the kind of translation errors which the students make, the holistic method of assessment has a confined range of objectivity. As a matter of fact, the holistic method has been applied very diversely by teachers and graders. The holistic assessment evaluates the overall quality of the end product based on a translator's intuition (Mariana, Cox, & Melby,

2015). This method is fast yet subjective, as it depends on the taste of the grader. According to Kockaert and Segers (2017), p. 149, "the value judgments of different holistic evaluators on the same translation can vary greatly." For instance, one grader considers one translation as excellent and creative, while another evaluator considers the same translation as fair or even unacceptable (Eyckmans et al., 2012). To put it briefly, the interrater reliability (intraclass correlation/interrater agreement/inter-observer reliability) is low among the evaluators for this method of assessment. Garant (2010), p. 10, has pointed out that "points-based error focused grading" (as a paradigm shift) has been replaced by the holistic method at the University of Helsinki. Translation is better evaluated with a focus of "discourse level holistic evaluation" than "grammar-like" and "analytical" evaluation (Kockaert & Segers, 2017). The holistic method concentrates chiefly on a "context sensitive evaluation" (Akbari & Segers, 2017b) and is supposed to move away from exclusive attention to grammatical errors in translation tests (Kockaert & Segers, 2017, p. 149). Waddington (2001) adapted the holistic method of assessment and designed the following paradigm (scores from 0 to 10) (Table 1).

Although the holistic method of assessment is reasonable, it does not have sufficient objectivity since the evaluators/graders are not always in a position of agreement. As Bahameed (2016), p. 144, noted, the holistic method relies partially on the "corrector's personal anticipation and appreciation." Truth be told, there are no specific criteria available while scoring a translation draft holistically.

Another disadvantage of this method is that it cannot determine the top students in a simple way as their scores "may reach one-third out of the whole translation class" (Bahameed, 2016). This makes the holistic method a lenience method since the students are not liable for minor mistakes such as lexical, grammatical, and spelling errors. These minor errors cannot be overlooked by an evaluator or the exam corrector as they constitute a matter in the quality of the holistic method of assessment which is too demanding to measure. Its leniency can reflect negatively on the quality of the end

Table 1 Holistic method of assessment (Waddington, 2001, p. 315)

Level	Accuracy of transfer of ST content	Quality of expressions in TL	Degree of task completion	Mark
Level 5	Complete transfer of ST information, only minor revision needed to reach professional standards.	Almost all the translation reads like a piece originally written in English. There may be minor lexical, grammatical, and spelling errors.	Successful	9, 10
Level 4	Almost complete transfer; there may be one or two insignificant inaccuracies; requires a certain amount of revision to reach professional standards.	Large sections read like a piece originally written in English. There are a number of lexical, grammatical, or spelling errors.	Almost completely successful	7, 8
Level 3	Transfer of general ideas but with a number of lapses in accuracy; needs considerable revision to reach professional standards.	Certain parts read like a piece originally written in English, but others read like a translation. There are a considerable number of lexical, grammatical, or spelling errors.	Adequate	5, 6
Level 2	Transfer undermined by serious inaccuracies; thorough revision required to reach professional standards.	Almost the entire text reads like a translation; there are continual lexical, grammatical, or spelling errors.	Inadequate	3, 4
Level 1	Totally inadequate transfer of ST content, the translation is not worth revising.	The candidate reveals a total lack of ability to express himself adequately in English.	Totally inadequate	1, 2

product and also the teaching process in the long run. Therefore, this method may not be sustainable and supportable in the field of translation evaluation and assessment.

Analytic method

The analytic method of assessment or assessment grids method is based on error analysis and is claimed to be more valid and reliable compared to the holistic method (Waddington, 2001, p. 136). In the analytic method, the evaluator/grader provides a grid. In doing so, the number of error types and levels can be increased; however, this must be carried out with caution. This is due to the fact that an increase in error types or levels can diminish the practical workability of analytic assessment. This method evaluates the quality of translation through scrutinizing the text segments such as paragraphs, individual words, etc., based on certain criteria. As noted by Eyckmans et al. (2013), errors associated with translation must be marked in terms of "the evaluation grid criteria". Moreover, the grader must firstly determine the types of error such as language errors or translation errors and consequently he/she provides the relevant information in the margin in accordance with the nature of the errors (Table 2).

Last but not least, the analytic method is time-consuming; however, the translator will have "a better understanding of what is right and what is wrong in translation" (Kockaert & Segers, 2017, p. 150). This method has a demerit that a grader concentrating on the small text segment of a source language does not certainly have a complete view of the target text. Besides, the analytic method is subjective and requires more time than the holistic method. Moreover, various graders/evaluators do not always concur with one another.

Preselected Items Evaluation (PIE) method

Preselected Items evaluation (PIE) method is a system which is appropriate for summative assessment (objective assessment in terms of test scores or key concepts

Table 2 Analytic Method of Assessment (Eyckmans, Anckaert, & Segers, 2009)

Meaning or Sense	Any deterioration of the denotative sense: erroneous information, nonsense, important omission…	− 1
Misinterpretation	The student misinterprets what the source text says: information is presented in a positive light whereas it is negative in the source text, confusion between the person who acts and the one who undergoes the action…	− 2
Vocabulary	Unsuited lexical choice, use of non-idiomatic collocations	− 1
Calque	Cases of a literal translation of structures, rendering the text into-French	− 1
Register	Translation that is too (in)formal or simplistic and not corresponding to the nature of the text or extract	− 0.5
Style	Awkward tone, repetition, unsuited assonances	− 0.5
Grammar	Grammatical errors in French (for example, wrong agreement of the past participle, gender confusion, wrong agreement of adjective and noun,….) + faulty comprehension of the grammar of the original text (for example, a past event rendered by a present tense,…), provided that these errors do not modify the in-depth meaning of the text	− 0.5
Omission	See sense/ meaning	− 1
Addition	Addition of information that is absent from the source text (stylistic additions are excluded from this category	− 1
Spelling	Spelling errors, provided they do not modify the meaning of the text	− 0.5
Punctuation	Omission or faulty use of punctuation. Caution: the omission of a comma leading to an interpretation that is different from the source text, is regarded as an error of meaning or sense	− 0.5

comparison) (Kockaert & Segers, 2014). As for time management and practicality, the number of preselected items in the source text is limited in the PIE method. The PIE method is a calibration and dichotomous method in which the former checks the accuracy "of the measuring instrument" and the latter inspects the distinction between correct and incorrect solutions (Kockaert & Segers, 2017, p. 150). The preselection of the items to be evaluated in a source text is selected in terms of two factors: *p value* (item difficulty) (the proportion of examinees answering an item correctly) and *d-index* (item discrimination, or candidates' differentiations on the basis of the items being measured). The calculation of the *p* value and d-index relates to "the minimum number of items needed for a desired level of score reliability or measurement accuracy" (Lei & Wu, 2007). With this in mind, the *p* value refers to the ratio of participants who answer an item correctly. According to Sabri (2013, p. 7), an ideal *p* value "should be higher than 0.20 and lower than 0.90". Therefore, the larger the population of the participants answering an item correctly, the easier and simpler the selected item will be.

In order to calculate the d-index, the PIE method applies an extreme group method through the calculation of higher group of scorers minus the lower group of scorers. Extreme group method measures the d-index with the following parameters: the top 27% candidates and the bottom 27% candidates of the entire score ranking are analyzed. Using 27% rules will maximize differences in normal distribution. The difficulty of the selected items based on *p* value and d-index is calculated after administering the test. The preselected items not responding to docimological standards (poor *p* value and d-index) will be eliminated from the translation test.

Besides stating the overall framework of this method, the validity and reliability of PIE assessment remain in question. No justification is given of why the items of the text are preselected as the most difficult or easy ones for the candidates. Which criteria determine the selection of the items and in what way(s) is this evaluation method usable in the translation classroom? Also, one has to consider the desired number of preselected items in a test. What is the ideal number of preselected items in the source text? When the translation is evaluated, what happens to other mistakes in the text? This may also raise the question of whether the PIE method is practical for every language pair.

Calibrated Parsing Items Evaluation (CPIE) method

As noted, the real significant challenge of translation evaluation methods is how to improve and increase the reliability and validity of the end product, viz., translation assessment. Therefore, proposing flexible methods of translation quality evaluation will augment the efficiency of translation quality assessment. Calibrated Parsing Items Evaluation (CPIE) will gain new perspectives to be applied in conditions such as translation service providers, universities, and companies having an advanced expertise in the evaluation of the end product. The present model is a combination of norm- and criterion-referenced assessments in which it firstly identifies the whole parses in a text (norm-referenced assessment) and then selects the docimologically justified items to be measured. The CPIE method consists of 6 stages: (1) holistic scoring by means of evaluators' intuition (the parses at this stage are docimologically unjustified), (2) the application of Brat Visualization software Stanford CoreNLP parser to distinguish every parse in a source text, (3) the calculation of p-docimology (CPIE takes up parses with

an ideal p-docimology which ranges from 0.27 to 0.79), (4) item discrimination (hereafter d-index) calculation on the basis of 21% rule instead of 27% rule of the PIE method to measure the extreme group method, (5) the extraction of the parses having a significant *p* (0.27–0.79) and d (0.30 and above), and finally (6) the recalculation of scores.

Selecting the size of the tails in a normal distribution from a distribution of a test scores is of critical importance. Traditionally, the size of the selected tails was assumed as an independent sample. However, this presupposition does not apply here. Conversely, the size of the selected tails is dependent and should contain about 21% instead of 27%. This is mainly due to the fact that the correlation between the concomitant variable [viz. *covariate*] and the test scores is not small and has correlation one (D'Agostino & Cureton, 1975), p. 49.

This norm- and criterion-referenced assessment method is a dichotomous and calibrated evaluation method (Akbari, 2017b). However, the selection of parsing items (having an acceptable *p* and d) will be different with regard to didactic translation and professional translation. In a didactic context, there should be a link between the selected parsing items (after identifying the docimologically justified parsing items) and the themes studied during the translation course such as typical characteristics of political, journalistic, and legal texts and also special terminologies covered in the classroom setting (the focus of our research paper). In a professional context, there should be a link between the selected parsing items and translators' competences (e.g., what do you expect from the translator in your translation company?).

Methods

The aim of the research

This paper first attempts to describe the full application of the CPIE method and then seeks to measure the degree of validity and reliability of this method compared to the methods such as the PIE, holistic, and analytic methods.

Description of the participants and materials

The study for the present paper took place in 2017. Forty translation students from the Bachelor of Arts in Translation Studies at the University of Isfahan, Iran, participated in this research through signing a letter of consent. The translator students were all native Persian speakers (L1) averaged age 21 years. They passed the courses associated with political translation, journalistic translation, translation of legal deeds, and literary translation through which they were exposed to various translational texts. They were asked to translate a short text (236 words) from English (L2) to Persian (L1). Although there were differences in the subjects' level of English language proficiency, the standard presupposition was that it was generally of a good standard, as the enrollment in their study programs required evidence of passing prerequisite credits such as political, economic, and journalistic translation courses.

The subjects were asked to translate a short text from "*Joint Comprehensive Plan of Action*" (The International Agreement on the Nuclear Program of Iran) among Iran and P5+1 (Germany, USA, England, Russia, France, and China) into Persian (L1). The participants were all familiar with political terminologies and structures since they passed the relevant courses associated with political and economic translations. The

length, type, register, and the difficulty of the source language were considered representative for the materials taught in the translation courses at the University of Isfahan. Finally, for the present study, five different translations made by five official translation agencies (Eshragh Translation Agency, Transnet, Mandegar, Mirpars, Iran Translation Service Agency) were used as reference translations for the evaluators, so they would have a "spectrum of correct equivalents" (Akbari & Segers, 2017b) when evaluating the translation drafts. These translation agencies have longstanding experience (nearly 10 years) in evaluating and translating all types of text particularly political texts.

Description of the procedures

The translation drafts were handed to 16 raters and they were requested to score them on the basis of the holistic (4 raters), analytic (4 raters), PIE (4 raters), and CPIE (4 raters) methods. Selecting the evaluators was also of great importance. The selection of the evaluators was carried out on the basis of their longstanding experience in translation evaluation and translation quality assessment. The evaluators were selected from the three universities, namely (1) the University of Isfahan, (2) the University of Sheikhbahaei, and (3) Azad University (Shahreza branch). They all have long-established experience (nearly 15 years) in translation quality assessment.

The analytic and holistic evaluators were requested to score the translation drafts using Eyckmans et al.'s (2013) and Waddington's (2001) frameworks, respectively. The scores were all calculated up to 20. The evaluators who applied the PIE method were asked to score the translation drafts on the basis of Kockaert and Segers' stages of PIE method by means of identifying the preselected items with good *p* values and d-indices. Finally, the evaluators using the CPIE method were requested to score the translation through docimologically justified parses having significant p and d. The evaluators were notified about the quasi-experimental design of the present study.

Type of statistical analyses

The reliability of CPIE compared to other evaluation methods was measured through "*Spearman rho*" (a non-parametric measure of rank correlation used for continuous variables) and "*Regression Variable Plots*" (which measures the correct strength of the linear and growth relationships among the variables to check for residuals or outliers) (SPSS 2017). Likewise, the validity of the CPIE method compared to other evaluation methods was calculated by means of "*F-Statistics*" (the ratio of variances measuring a degree of dispersion between the variables to understand which one is more valid and has a larger value) (SPSS 2017) and "*Probability Plots*" (a graphical technique to evaluate whether the data set is approximately normally distributed) (Minitab 2017).

The application of CPIE method: a case study

Stages of CPIE method

Holistic scoring

Forty students were asked to translate the source text (L2) into plain Persian within 90 min. Once the translation task was carried out, one expert evaluator (from the four holistic evaluators) was assigned to score the translations using a holistic (impressionistic-intuitive) system of evaluation on the basis of Waddington's (2001) framework (see

the "Holistic method" section). The scores of the participants at this stage were as follows (Table 3):

Participants [1], [8], [11], [27], [28], [29], [30], [31], [33], [36], and [37] had the lowest marks compared to the rest of the participants. As per the evaluator's comments, those participants mostly resorted to literal-word-for-word-translation, which resulted in unclear target text meanings for some parts. They often did not take up the appropriate approaches for their translations. The participants in question made critical semantic errors, which made their translations imprecise, vague, and inaccurate. They often lost the contextual function of the source text, and they relied more on a one-to-one correspondent technique (literal translation).

Brat visualization Stanford CoreNLP parser

Brat Stanford CoreNLP Software is a shared task for distinguishing how factual statements and annotations (parses) can be interpreted based on their textual contexts including a hypothesis and an experimental result. The extraction of information in general and parses (annotation) in particular is the main task of capturing information contained in text through Brat software. Brat natural language processing software was used to create Anatomical Entity Mention (AnEM), a corpus including 500 documents extracted from various databases such as PubMed and PMC with full-text extracts. AnEM was considered a reference for assessing and evaluating methods for the detection of anatomical entity motion. Also, Brat software is used to identify metaphor annotation through bottom-up identification (Stenetorp et al., 2012) which is mainly focused on the linguistic metaphors in a source text and deducing the conceptual metaphors underlying them.

Brat rapid annotation is an intuitive web-based device for parsing and annotating text based on Natural Language Processing (NLP) technology. This tool is devised for rich structured annotation in terms of various NLP tasks. The chief aim of Brat is to "support manual curation efforts and increase annotator productivity using NLP techniques" (Stenetorp et al., 2012). Modern parsing and annotation tools are technically directed to users "beyond the minimum required functionality" (Stenetorp et al., 2012). Likewise, user-friendly technologies can support—not supplant—the human decisions (judgments) which in turn can contribute to maintain the quality of annotation and cause the annotations to be more accessible to novice (non-technical) users. Brat NLP

Table 3 Holistic scoring (docimologically unjustified scores)

Par	Score (holistic)	Par	Score (holistic)	Par	Score (holistic)	Par	Score (holistic)
1	10	11	10	21	14	31	11
2	13	12	16	22	18	32	18
3	13	13	15	23	17	33	11
4	13	14	17	24	16	34	13
5	15	15	18	25	16	35	15
6	14	16	20	26	15	36	12
7	16	17	19	27	12	37	11
8	12	18	18	28	11	38	15
9	14	19	17	29	9	39	15
10	16	20	13	30	10	40	18

software is based on the stav visualizer (text annotation visualizer) and was originally designed to visualize the BioNLP'11 Shared Task (a community wide-move in bio text mining). Brat has been utilized to produce well-over 50,000 parses or annotations. Other important goals of Brat tool are the de-centralization of data making synchronization issues, simplifying the complexity of set-up annotators, and perusing the visually adjacent annotations in a text. The present study applies Brat software in order to scrutinize every annotation or parse (norm-referenced assessment) in a text and then selects the annotations or parses which are docimologically justified (criterion referenced assessment). This software supports nearly 100 various scripts in which they can be implemented in full Unicode. Also, the manuscripts can be converted to UTF-8 encoded Unicode and ASCII format. Another feature of Brat NLP parsing software includes a number of traits for "comparing multiple sets of annotations" (Brat, 2014) for the same documents involving automatic and systematic juxtaposition through distinguishing and marking the differences via side-by-side visualization. This comparison relates to assessing automatic systems and concurrence among the translation evaluators and the visualization of different parses, which are a common source of difficulty and error (docimologically justified items).

Figure 1 below is an extracted brat annotation screenshot of the source text, which categorizes every annotation into specific groups such as NNP (proper noun), POS (part of speech), NN (singular noun), VBD (past tense verb), IN (preposition), and DT (determiner). After having imported the whole text, Brat NLP tool exported 354 annotations based on the adjacent parses in a text. At this stage, the evaluator tries to extract all the parses in a text and then compares them to the participants' translation drafts. This comparison is carried out in accordance with the available translations done by various experts at translation agencies. As noted, this study provided five different translations by five translation agencies to prepare a situation for the evaluator with the spectrum of correct and acceptable equivalents of the terms of the source text when evaluating the translation drafts.

The calculation of P-Docimology

No technical merits reside in involving test exercises with a wide range of readability (difficulty), even if the test is going to distinguish the higher group of scorers from the lower group of scorers. And even if the entailment of items "with a spread difficulty"

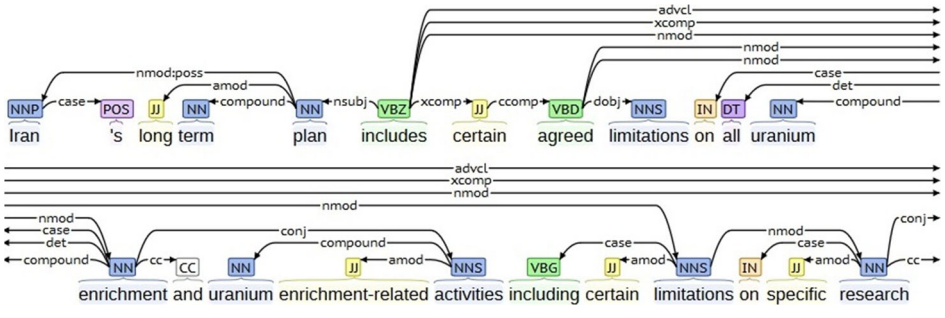

Fig. 1 Brat annotation tool. This figure showed the process of segmentation through the Brat annotation tool

may be impressionistically alluring, it is argued that the appropriate range of difficulty varies between 0.50 and 0.60, as these have a more positive effect on the outcome of the test reliability and validity (Feldt, 1993). According to Mehrens and Lehmann (1991), the items having a difficulty range close to 0.95 or 0.05 fail to distinguish and differentiate among the higher and lower marked students and therefore cannot contribute to the test reliability. In line with Mehrens and Lehmann, Tinkelman (1971) encapsulated the view of measurement authorities through expatiating that a limited and acceptable range of item difficulty in tests ranged from 0.50 to 0.65. The optimal and appropriate point of concentration of an item difficulty relies upon the probability of "guessing a correct answer" (Feldt, 1993:38). According to Feldt (1993), "the reliability advantage of a test with items concentrated near the optimal value is relatively minor". For instance, once guessing is possible, "an instrument with items' difficulties administered consistently between 0.27 and 0.79 might be exposed to have a reliability only a few hundredths lower than a test with item difficulty concentrated at the 0.50 level" (Feldt, 1993, p. 38). P-docimology is necessary since it influences all "test scores statistics" (e.g., item reliability, mean, and variance) (Tang & Logonnathan, 2016). With the above explanations, p-docimology can be described as "the ratio of the participants answering an item correctly to the total examination participants" (Akbari & Segers, 2017a):

$$\text{Pi (docimological}-p) = \frac{\text{Ri (a number of participants answer an item correctly)}}{\text{Ni (total population of the participants)}}$$

In this case, the total population of the participants equals the total population of the answers. On the basis of all the extracted answers by the evaluator, 100 items of 354 items have a difficulty range between 0.27 and 0.79 which are considered appropriate items. For instance, parsing item [150] was answered correctly by almost all participants:

$$P\text{–docimology (PI 150)} = \frac{38}{40} = 0.95$$

With this idea, the p-docimology of item [150] made this the least difficult item (the easiest) since 95% of the participants answered that item correctly. Therefore, it must be eliminated since it is out of the recommended range between 0.27 and 0.79. To take another example, parsing item [120] would be considered an appropriate p-docimological parsing item since its difficulty range was settled between 0.27 and 0.79, as nearly 73% of the participants answered/translated this item correctly.

$$P\text{–docimology (PI 120)} = \frac{29}{40} = 0.725$$

Of the 354 parsing items, 100 parsing items were considered appropriate items in terms of item difficulty. They mostly ranged between 0.35 and 0.66 so they had an appropriate p-docimology. However, one cannot label them as docimologically justified parsing items since their ranges in d-index have not been verified. Therefore, the next stage is to measure the degree of item discrimination through the help of the 21% rule to distinguish which items are regarded as justified parsing items.

The calculation of D-index (item discrimination index)

D-index is a degree to which the students "with high overall exam scores also got a particular item correct" (Exam, Understanding Your, 2017). D-index is a point biserial correlation coefficient (used for dichotomous variables) ranging from − 1 to + 1. Item discrimination can be calculated based on the following principle:

> If the test and a single item measure the same thing, one would expect people who do well on the test to answer that item correctly, and those who do poorly to answer the item incorrectly. A good item discriminates between those who do well on the test and those who do poorly. Two indices can be computed to determine the discriminating power of an item, the item discrimination index, d, and discrimination coefficients (Matlock-Hetzel, 1997).

Item discrimination is a kind of measurement to test the ability of one item to discriminate between those participants who get higher scores on the total test and those who obtain the lower scores, i.e., the proportion of higher ability students and the lower ones (Miller, Linn, & Grounlund, 2009). In order to calculate the item discrimination index, we calculate the following:

$$D\text{–index}_{(\text{extreme group method})} = \text{HG}\left(\frac{21}{100}\right) - \text{LG}\left(\frac{21}{100}\right)$$

where "HG" refers to the higher group of scores and "LG" the lower group of scorers. Twenty-one percent is used to show the maximum difference in normal distribution as maintaining enough case for analysis.

Twenty-one percent (nearly eight of the total number of participants) of the higher ability students are participants [16], [17], [15], [18], [22], [32], [40], [14], [19], and [23] while participants [1], [11], [28], [29], [30], [31], [33], and [38] are grouped as lower scorers. For instance, parsing item [120] was answered correctly by 7 students in HG and 3 students in LG. With this in mind,

$$\text{D–index}_{(\text{extreme group method})} = \text{HG} - \text{LG}$$

$$\text{D–index}_{(\text{extreme group method})=} \frac{7}{8} - \frac{3}{8} = 0.50$$

This showed that 50% of the higher and lower group of scorers answered the intended item correctly. As stated in the "Calibrated Parsing Items Evaluation (CPIE) method" section, items having a range between 0.27 and 0.79 for p-docimology and 0.40 or above for d-index can be considered docimologically justified parsing items and they must be included in the measurement since the outcome of the research depends on their docimological ranges. To take another example, parsing item [140] was answered correctly by 6 participants of HG and 5 participants of LG; therefore, its item discrimination index was as follow:

$$\text{D–index}_{(\text{extreme group method})=} \frac{6}{8} - \frac{5}{8} = 0.125$$

Statistically, this parsing item cannot be regarded as an appropriate item, since its range is below 0.40 and therefore, it must be eliminated or improved to be included in a test. Overall, items with a d-index of 0.40 and above are very good discriminators,

items with a d-index of 0.30 to 0.39 are good discriminators, items with a d-index of 0.20 to 0.29 are fairly good discriminators, and items with a d-index of 0.19 or less are poor discriminators; (Akbari & Segers, 2017a).

The extraction of Docimologically justified parsing items

On the basis of item discrimination index (d-index) and item difficulty (p-docimology) calculated in stages III and IV, 75 parsing items were grouped as docimologically justified parsing items (criterion-referenced assessment); they mostly ranged from 0.45 to 0.65 for p-docimology and 0.40 and above for d-index. The complete list of the accepted items is as follows (Table 4):

Recalculation of scores

As noted, stage I of the CPIE method tried to score the translation drafts based on holistic scoring. The last stage of the CPIE method tries to recalculate scores in terms of justified parsing items (previous section) and then compares the results with the general score calculation (stage I).

According to Table 5, the outcome of this recalculation is the most critical for participants [31] and [37], going from 11/20 (holistic scoring) to 5.375 and 8.350 (CPIE scoring) respectively. This is due to the fact that despite the overall quality of their translations, they had not been able to translate most of the docimologically justified parsing items correctly (75 parsing items) after calculating good p-docimology and d-index. However, for instance, participants [17], [24], and [40] received a higher score compared to the first calculation (19 vs. 19.8), (16 vs. 18), and (18 vs. 19.3), respectively. This was due to the fact that they have translated the justified parsing items correctly besides translating the whole parses in a text (both justified and unjustified parsing items).

Results and discussion

Verification of the first hypothesis

Hypothesis: CPIE as a method of translation assessment is more valid than the holistic, analytic, and PIE methods (*the question of validity*).

The main objective of the first hypothesis is to analyze the validity of the CPIE method compared to other evaluation methods through recruiting 16 raters to score the translation drafts using the CPIE, PIE, holistic, and analytic methods. By validity, we mean the credibility of the research. To measure CPIE's validity, we used F-statistics, i.e., the ratio of variances measuring a degree of dispersion between the variables to understand which one is more valid and has a higher value. F-statistics assess the quality of variances to account for the degree of freedom (*df*) (how many variables involved in a calculation have the freedom to vary). The main reason to use F-statistics is its flexibility in a variety of situations (Minitab 2017). Through altering the variances included in a ratio, the F-ratio (also F-statistics and F-value) becomes a very reliable and plain test. Moreover, proving the degree of validity of the CPIE method allows us to prove the

Table 4 Docimologically justified parsing items

Item	P_{doc}	d_i	Item	P_{doc}	d_i	Item	P_{doc}	d_i
5	0.550	0.450	55	0.455	0.525	95	0.742	0.455
11	0.650	0.650	66	0.455	0.525	104	0.790	0.555
18	0.752	0.550	67	0.355	0.425	113	0.740	0.625
20	0.752	0.550	69	0.352	0.555	120	0.725	0.500
33	0.552	0.450	72	0.552	0.625	135	0.650	0.525
38	0.355	0.450	77	0.630	0.850	139	0.650	0.525
41	0.655	0.550	80	0.655	0.950	141	0.630	0.455
47	0.452	0.650	84	0.752	0.525	143	0.790	0.625
49	0.455	0.725	88	0.745	0.455	148	0.530	0.425
53	0.750	0.625	91	0.742	0.655	152	0.630	0.455
155	0.755	0.525	191	0.555	0.755	251	0.752	0.555
160	0.755	0.630	195	0.655	0.825	259	0.752	0.555
167	0.655	0.650	199	0.790	0.855	260	0.630	0.425
173	0.652	0.435	206	0.755	0.630	267	0.355	0.465
177	0.645	0.425	208	0.745	0.625	277	0.352	0.435
178	0.523	0.435	220	0.355	0.550	279	0.655	0.765
185	0.425	0.525	228	0.465	0.635	282	0.630	0.665
188	0.455	0.555	233	0.455	0.725	296	0.455	0.565
189	0.750	0.525	234	0.635	0.435	298	0.452	0.525
190	0.552	0.455	245	0.535	0.635	307	0.465	0.665
309	0.275	0.455	322	0.365	0.455	337	0.755	0.455
311	0.275	0.455	325	0.455	0.625	341	0.725	0.455
315	0.335	0.550	327	0.425	0.675	346	0.755	0.425
318	0.352	0.525	330	0.555	0.655	349	0.655	0.825
320	0.352	0.525	333	0.525	0.625	352	0.630	0.425

reliability of the data as well because validity implies reliability (McKenna & Dougherty Stahl, 2015).

Figure 2 visualizes the differences between the degrees of validity among the four methods by means of a probability plot (Minitab 2017), which measures whether or not the variable (data) set is approximately normally distributed. The validity of translation quality assessment methods used to rate a translation draft based on the seriousness of detected errors has been under question since they disregard the macrotextual features and also the degree of dispersion of variables of the target text and the fact that an end product along with more linguistic or language errors may nonetheless be of better overall quality. By applying criterion-referenced assessment, this study aims to select parses objectively (having good p-docimology and d-index). The results with regard to the first hypothesis are as follows (Table 6):

By and large, the results of the present paper showed a significant difference between the CPIE method and the PIE, holistic, and analytic methods in terms of F-statistics. Therefore, according to the obtained results, the null hypothesis was rejected in favor of the CPIE method. The F-ratio of CPIE method indicates that the differences which can be observed between the results are not only due to the fact that the translation method (CPIE) is different, but also to differences

Table 5 CPIE recalculation of scores (Par, participant)

Par	Score (holistic)	CPIE	Par	Score (holistic)	CPIE	Par	Score (holistic)	CPIE	Par	Score (holistic)	CPIE
1	10	13.385	11	10	10.380	21	14	15.750	31	11	5.375
2	13	14.112	12	16	14.955	22	18	18.725	32	18	18.375
3	13	13.380	13	15	16.575	23	17	18.585	33	11	12.425
4	13	15.255	14	17	17.535	24	16	18.025	34	13	14.450
5	15	15.755	15	18	18.555	25	16	17.055	35	15	15.625
6	14	15.252	16	20	19.975	26	15	15.955	36	12	13.752
7	16	15.375	17	19	19.772	27	12	13.032	37	11	8.350
8	12	13.332	18	18	18.375	28	11	13.530	38	15	17.025
9	14	14.852	19	17	18.025	29	9	10.225	39	15	16.470
10	16	15.225	20	13	14.455	30	10	11.220	40	18	19.255

among the evaluators in their application of each evaluation method. Also, the visual inspection of the four evaluation methods indicates that the CPIE method is more valid compared to the holistic, analytic, and PIE methods. The *p* value of the figures showed a significant difference in favor of the CPIE method.

The four plots in Fig. 2 were compared in terms of *p* value. Generally, a *p* value of 0.05 works well as "a significant level of 0.05 indicates that the risk of concluding the data do not follow the distribution when, actually, the data do follow the distribution is 5%" (Minitab 2017). There are two options with regard to the *p* value in a probability plot: (1) *p* value ≤ *α* (0.05), this shows the data do not follow the distribution which indicates that the null hypothesis must be rejected

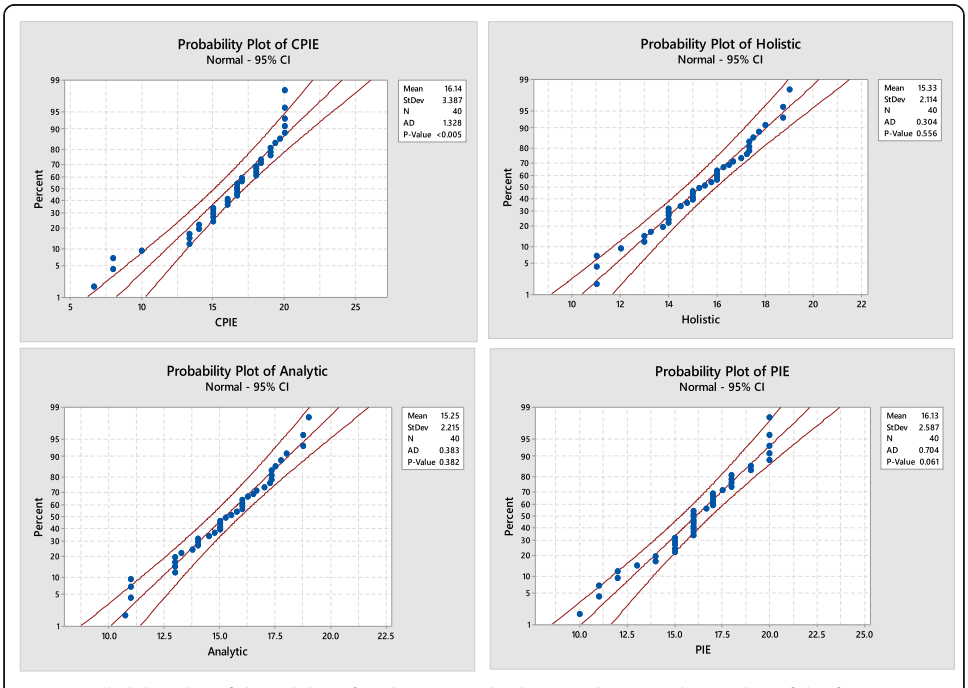

Fig. 2 Probability plot of the validity of evaluation methods (Minitab 2017). The *p* value of the figures showed a significant difference in favor of the CPIE method. The *p* values for the PIE, analytic, and holistic methods were 0.061, 0.382, and 0.556, respectively, which were greater than 0.05

Table 6 Validity of the Four Methods (SPSS 2017) (α level: 0.05)

		Sum of squares	df	Mean square	*F*	Sig[a] (*p* value)
Validity (CPIE)						
Between people		1320.322	39	33.854		
Within people	Between items	13.147	3	4.382	*3.364*	*.021*
	Residual	152.426	117	1.303		
	Total	165.573	120	1.380		
Total		1485.895	159	9.345		
Grand mean = 15.7518						
Validity (holistic)						
Between people		761.154	39	19.517		
Within people	Between items	6.327	3	2.109	*.475*	*.700*
	Residual	518.999	117	4.436		
	Total	525.326	120	4.378		
Total		1286.480	159	8.091		
Grand mean = 15.5214						
Validity (analytic)						
Between people		738.569	39	18.938		
Within people	Between items	9.418	3	3.139	*.868*	*.460*
	Residual	423.041	117	3.616		
	Total	432.460	120	3.604		
Total		1171.028	159	7.365		
Grand mean = 15.6574						
Validity (PIE)						
Between people		627.058	39	16.078		
Within people	Between items	5.900	3	1.967	*1.031*	*.382*
	Residual	223.153	117	1.907		
	Total	229.052	120	1.909		
Total		856.111	159	5.384		
Grand mean = 15.9516						

(Probability Plot of CPIE) and (2) p value ≥ α (0.05), this shows that there is no sufficient evidence to conclude that the data do not follow the distribution, and as a result the decision is to reject the null hypothesis. Therefore, the null hypotheses with regard to the PIE, analytic, and holistic probability plots state that the data follow a normal distribution. However, the p value for the PIE, analytic, and holistic methods are 0.061, 0.382, and 0.556, respectively, which is greater than 0.05. This indicates that the null hypothesis should be rejected. On the basis of the plots, the validity set of the four methods is as follows:

CPIE>>>PIE>>Analytic>Holistic.

Verification of the second hypothesis

Hypothesis: The quality of a translation can be evaluated more reliably if the method of evaluation assesses all the parsing items having good p and d (norm-referenced assessment towards criterion-referenced assessment) rather than some "specific items" (PIE

method), "pre-conceived criteria" (analytic method), and "impressionistic-intuitive scoring" (holistic method) among the raters (*the question of reliability*).

The main objective of the second hypothesis is to measure the degree of reliability of the four methods to analyze which of the evaluation methods is more consistent and produces the same results when applied repeatedly "to the same population under the same conditions" (Williams, 2013). In this respect, translation quality assessment is reliable when the decisions made by the evaluators are consistent and stable. To measure their degree of reliability, this study used Spearman's rank correlation coefficient (for continuous variables). The reason to select Spearman rho is that it assesses the relationship between the variables through applying a monotonic function.

The results of Spearman's rank correlation coefficient were used to analyze the interrater reliability among the evaluators who used the four methods of translation evaluation. The results of the interrater reliability illustrate the superiority of CPIE evaluators in terms of docimologically justified parsing items (0.806, 0.857, 0.896, 0.911, 0.920, and 0.898). The results indicated that the CPIE method is more consistent (as highlighted in Table 7—see appendix 2) compared to the PIE, holistic, and analytic methods. According to Morales (2000, cited in Waddington 2004, p. 33),

> The adequate level of reliability depends above all on the use that is going to be made of the marks obtained. If the marks are going to be used as a basis for decision taking, then Morales recommends that the reliability coefficient should be at least 0.85.

Also, a regression variable plot (for continuous variables) was applied to predict the value of the variable on the basis of the relationship among the evaluators, as can be seen in Fig. 3. The regression plots are as follows:

As we may see in Fig. 3, all figures display some outliers. An outlier is an observed data point having a different value from the predicted value through the regression equations (Williams, 2016). In this respect, the more outliers in a translation evaluation method, the larger the residuals will be (Williams, 2016, p. 3). The outliers generally have a negative effect on the regression analysis, decreasing the fit of the regression equation. As can be seen, there are few outliers among the CPIE evaluators, which clearly shows that CPIE evaluators are more consistent with one another when scoring the translation drafts. By contrast, for the three other evaluation methods (PIE, holistic, and analytic), a great number of outliers were observed. This indicates that the scoring systems and the evaluation systems for these three methods are not consistent enough and have negative effects on both the outcome of the test and the fit of the regression analysis. Therefore, evaluating translations by means of the holistic, analytic, and PIE methods must be carried out with caution since the reliability of the results may be exposed to adverse effects.

Discussion

Why brat Stanford CoreNLP software?

Brat parsing software is based on the concept of "*what you see is what you get*" (Brat, 2014), in which all aspects in a text are represented visually on an intuitive

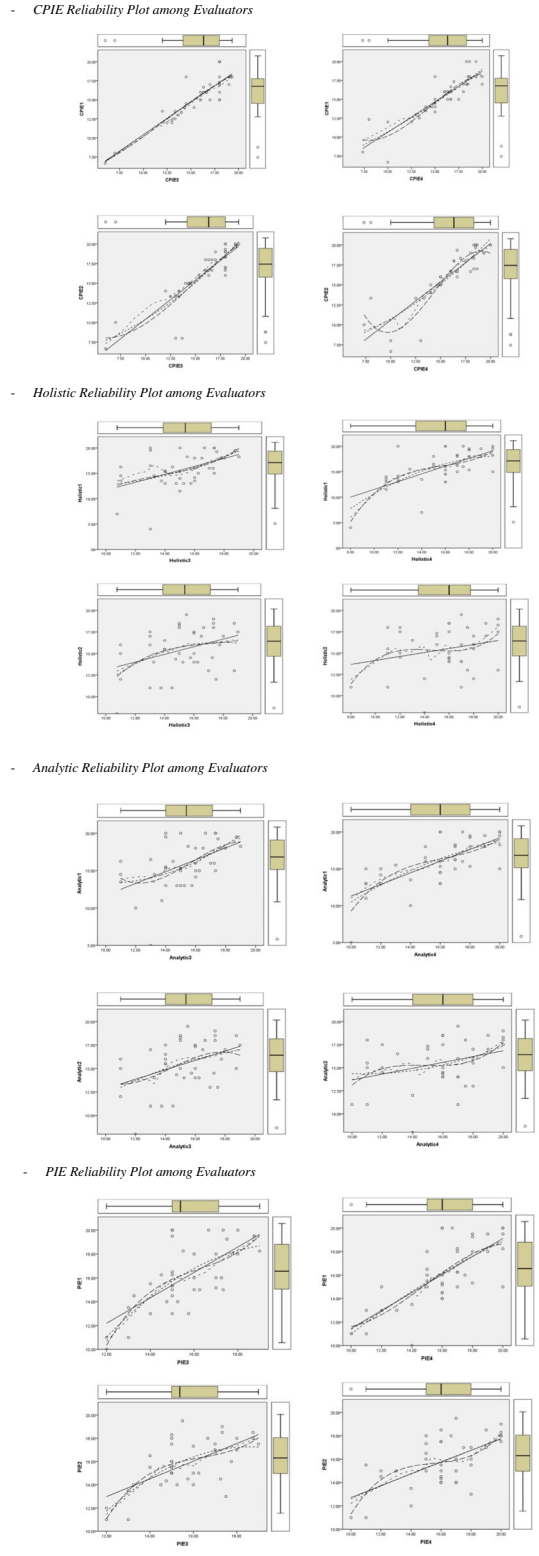

Fig 3. Regression Variable Plot of CPIE, Holistic, Analytic and PIE Methods

Fig. 3 Regression variable plot of CPIE, holistic, analytic, and PIE methods. A regression variable plot was applied to predict the value of the variable on the basis of the relationship among the evaluators. There were few outliers among the CPIE evaluators, which clearly showed that CPIE evaluators were more consistent with one another when scoring the translation drafts

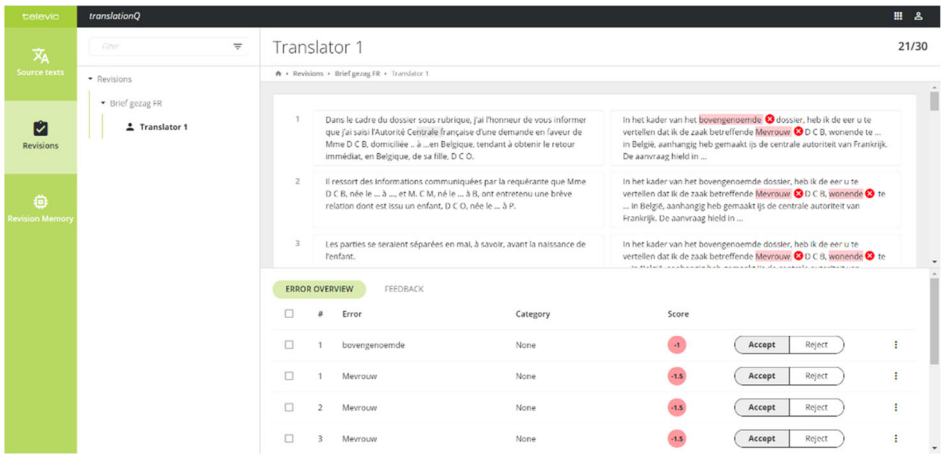

Fig. 4 Illustration of Brat software analysis. In this figure, the term "development" was analyzed by its related segments

basis. For instance, the extract in Fig. 4 is represented visually through Brat parsing software. Brat NLP software connects annotations, for instance, through adding a relation/connection between dichotomous parses. As illustrated in Fig. 4, every parse in a text is represented visually through different colors. Also, Brat software identifies the relation between the distinguished chunks in a text so that the evaluator can easily find the correspondent parsing items in a target language to check whether the identified parsing item is translated correctly. One of the most important features of the CPIE method is to evaluate all chunks which are docimologically unjustified parsing items (norm-referenced assessment) and then select the chunks or parses which are docimologically justified parsing items (criterion-referenced assessment) in a text.

For instance, the noun "development" (NN) in the source text corresponds to the term "activities" (*NN-compound-NNS*). In this respect, the evaluator must look for the corresponding translation of the terms "development" and "development activities" in the Persian language. The corresponding Persian translations were "*towse^{c}e*" (NN) and "*Gostæreše fæ^{c}ālijæt'hā*" (*NN-compound-NNS*), which were agreed upon by the evaluators as correct translations. To take another example, the term "stage" (NN) has relations with the terms "purposes" (*N-MOD*), "activities" (*N-MOD*), on the right side and "followed" (*N-MOD*), "pace" (*N-MOD*), "stage" (*CASE*), "stage" (*DET*), and "next" (*A-MOD*) on the left side. Also, the corresponding Persian translations were "*mærhæle*" (in general) (NN), "*æhdafe mærāhel*" (N-MOD), "*mærāhele fæ^{c}ālijæt'hā*" (right side), "*mærāhele pišerou*" (N-MOD), "*soræte pišræfte mærāhel*" (N-MOD), "*dæstjābi be mærāhel*" (CASE), "*mærhæleje xās*" (DET), and "*Mærhæleje bæ^{c}d*" (left side), respectively. CPIE evaluators checked the corresponding translations of the source terms in the Persian language and measured the acceptability of the translations (the degree of p-docimology and index discrimination) so as to label the source terms as docimologically justified parsing items or not.

With this idea, the evaluator must inspect the corresponding translations in the target language. These one-to-one correspondences, two-to-two correspondences,

one-to-many correspondences, and many-to-one correspondences pave the way for the evaluator to scrutinize the impact and values of the source language terms on the reciprocal language. Brat software inspects the impact (value) of all extracted chunks in a text to check the relation among them. Brat NLP software supports for normalization and different traits for connecting parses accompanied by data in external databases such as lexical and ontological resources (e.g., Freebase, Wikipedia, and Open Biomedical Ontologies).

Brat software integrates with other automatic parsing tools accessible as web-services such as CoNLL+MUC Model (a model used to identify general anaphoric co-references such as high coverage verbs, noun propositions, partial verbs, and noun word senses) (CoNLL, 2012) and Genia Model (a model used for a larger size of a training corpus and it is a combination of Treebank) (Bunt, Merlo, & Nivre, 2010) supported by Stanford NER and NERtagger respectively to feature lucid integrations with advanced methods such as sentence splitting and tokenization. Consequently, Brat NLP parsing software maintains a rich set of annotation primitives such as entity annotations, dichotomous relations, equivalence classes, *n*-ary associations (relationship among three or more classes), and attributes which can be utilized in any annotation or parsing task.

CPIE: norm or criterion referenced assessment method?

The tenseness between norm and criterion (outcome-based approach) assessment methods is probed in the domain of translation evaluation. The core principle of the criterion referenced assessment method is to what extent the values or the criteria selected are implicitly norm referenced. It is vague that neither assessment method is acceptable in extreme scenarios (Lok, McNaught, & Young, 2016). Most evaluators and researchers have confessed to a "pragmatic hybrid" respecting the convention of grade evaluation. Lok et al. (2016) have pointed out that there are differences and similarities between criterion and norm-referenced assessment methods; however, the distinction is blurred in practice.

In recent years, the criteria used in obviously criterion-referenced assessment methods are often latently based on norms derived from a group. In other words, one evaluator must look empirically at the ability and performance of the cohort in order to decide whether one criterion is acceptable. When the need for such analysis is conceded, then the evaluator must accept the possibility of a mismatch between criterion- and norm-referenced assessment methods and also the resultant need to deal with the disparity between these referenced methods. Not only is the meaning of criterion-referenced assessment "often norm-referenced, but also its interpretation has to be made in the group context" (Lok et al., 2016). The definition of criterion-referenced assessment methods in translation studies (1) has to be explicit through the active engagement of the translation students and the translation trainers/evaluators in interpreting their understanding (O'Donovan, Price, & Rust, 2004; Shay, 2008), (2) needs to be situated in a specific context (Sadler, 2005), and (3) requires the monitoring of norm-based distributions. According to Lok et al. (2016, p. 458), "norm referencing, as a result, becomes a strategy for checking on decisions made in a criterion referenced fashion". On the basis of the above

explanations regarding criterion- and norm-referenced assessment methods, the CPIE method benefits from the synthesis of norm- and criterion-referenced assessment methods through a feedback loop, unlike other translation evaluation methods, including a norm-referenced assessment method, criterion-referenced assessment (rubric), and the actual evaluating. This loop can be repeated many times. First, the participants' scores based on the holistic method are utilized to derive a set of docimologically unjustified parsing items in a source text which are incorporated into criterion-referenced rubrics. Second, after the first score calculation, the justified parsing items with acceptable p(s) and d(s) are derived (criterion-referenced assessment), the evaluator arrives at a set of scores (CPIE run). Finally, after the complete evaluation of the translation drafts via the CPIE method, the participants' performance is monitored to analyze the differences between their first score calculation and score recalculation. The use of this feedback loop has a number of benefits such as (1) both norm- and criterion-referenced assessment methods are both present in the CPIE method containing the degree of flexibility, (2) these two referenced assessment methods in a loop pave the way for the participants to receive beneficial feedback and summative information when the item they are translating is considered a docimologically justified parsing item, and (3) this feedback loop guards against the inflation of scores through the simultaneous use of both norm- and criterion-referenced assessment methods.

Conclusion

Limitations of the research

First, among the limitations of the present study are the proportionately small number of participants at the BA level and the fact that the translation assignment was carried out with paper and pencil. In a replication of the research paper with a larger number of participants, care must therefore be taken to provide a situation mimicking a real and professional environment by allowing the participants to perform the translation assignment on a computer. Second, the CPIE method is a time-consuming activity. A computerized platform is needed to control and check the answers in the imported translation drafts, and a list of correct and incorrect solutions of the parsing items needs to be prepared.

Implication of the research

Calibrated Parsing Items Evaluation has the potential to be applied in translation quality platform such as translationQ to measure the quality of the end product. TranslationQ is an advanced web-based platform that automates the objective revision of translations using a unique error, correction, and feedback memory through identifying the appropriate and acceptable docimologically justified items in the source language (Fig. 5). TranslationQ allows the translator to revise translations in an efficient and objective way, which is also the aim of the CPIE method. Also, the reviser can add new errors accompanied by the appropriate corrections and feedback in the course of the revision stage. TranslationQ will then automatically detect the same error in other translations and allow the reviser to apply the corrections and feedback. This process saves the reviser a

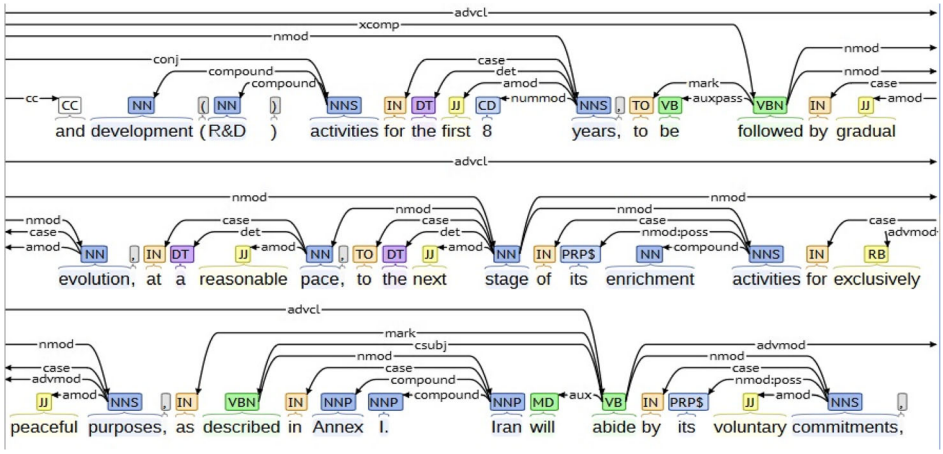

Fig. 5 TranslationQ platform. TranslationQ was considered a revision memory and likewise CPIE had the potential to be applied in the TranslationQ platform

significant amount of time and supports him/her in being objective: all translation drafts are corrected using the exact same criteria.

The core of translationQ is a revision memory; it recognizes errors in new translations and suggests corrections and feedback automatically. In this respect, an evaluator can still accept or reject the suggestions. TranslationQ allows the evaluators to exchange and merge revision memories, and reuse them with new texts and with other translations. Consequently, at the end of the revision stage, translationQ sends a detailed feedback report including the source text, the translation, a model answer, and all the corrections and feedback that apply to the translation. Every translator receives a personal report with only the remarks relevant for him/her. The CPIE method can be applied to the translationQ platform, since it can be operated in multiple domains such as legal, technical, medical, cultural, and political texts (Akbari, 2017a). Moreover, the CPIE method can be automated, as it has the potential to add options during the Brat process, update all existing corrections constantly, and more and more parses will be recognized. Furthermore, this method has the potential to be operated via feedback memory.

To sum, this research paper introduced a translation method called Calibrated Parsing Items Evaluation (CPIE) method seeking to objectify translation evaluation. This method tried to distinguish competent translators through six stages as stated in "The application of CPIE method: a case study" section. This norm- and criterion-referenced assessment method applied Brat Visualization Stanford CoreNLP parser to identify all annotations in a source text (norm-referenced assessment) and then determine the docimologically justified parsing items (criterion-referenced assessment). To corroborate the objectivity of this assessment method, interrater reliability (intraclass correlation) was conducted to analyze the significant differences among the holistic, analytic, and PIE methods. The results indicated that CPIE method complemented and solved the question of validity and reliability between the scores obtained by the CPIE evaluators and the scores obtained and evaluated by the holistic, analytic, and PIE evaluators.

Appendix

Table 7 Reliability of the CPIE, Holistic, Analytic, and PIE Method

Correlations (Evaluators)																		
			CPIE1	CPIE2	CPIE3	CPIE4	Holistic1	Holistic2	Holistic3	Holistic4	Analytic1	Analytic2	Analytic3	Analytic4	PIE1	PIE2	PIE3	PIE4
Spearman's rho	CPIE1	Correlation Coefficient	1.000	**.806****	**.857****	**.896****	.804**	.475**	.393*	.638**	.798**	.505**	.430**	.675**	.803**	.518**	.488**	.661**
		Sig. (2-tailed)	.	.000	.000	.000	.000	.002	.012	.000	.000	.001	.006	.000	.000	.001	.001	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	CPIE2	Correlation Coefficient	**.806****	1.000	**.911****	**.920****	.881**	.458**	.545**	.672**	.876**	.501**	.574**	.763**	.879**	.591**	.642**	.740**
		Sig. (2-tailed)	.000	.	.000	.000	.000	.003	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	CPIE3	Correlation Coefficient	**.857****	**.911****	1.000	**.898****	.866**	.505**	.452**	.591**	.855**	.540**	.480**	.683**	.855**	.615**	.563**	.671**
		Sig. (2-tailed)	.000	.000	.	.000	.000	.001	.003	.000	.000	.000	.002	.000	.000	.000	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	CPIE4	Correlation Coefficient	**.896****	**.920****	**.898****	1.000	.893**	.413**	.432**	.638**	.883**	.444**	.463**	.736**	.883**	.508**	.560**	.707**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.008	.005	.000	.000	.004	.003	.000	.000	.001	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	Holistic1	Correlation Coefficient	.804**	.881**	.866**	.893**	1.000	**.502****	**.460****	**.625****	.990**	.541**	.493**	.716**	.987**	.619**	.572**	.694**
		Sig. (2-tailed)	.000	.000	.000	.000	.	.001	.003	.000	.000	.000	.001	.000	.000	.000	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40

Table 7 Reliability of the CPIE, Holistic, Analytic, and PIE Method *(Continued)*

Correlations (Evaluators)																	
		CPIE1	CPIE2	CPIE3	CPIE4	Holistic1	Holistic2	Holistic3	Holistic4	Analytic1	Analytic2	Analytic3	Analytic4	PIE1	PIE2	PIE3	PIE4
Holistic2	Correlation Coefficient	.475**	.458**	.505**	.413**	**.502****	1.000	**.314***	**.222**	.523**	.991**	.328*	.294	.532**	.851**	.398*	.336*
	Sig. (2-tailed)	.002	.003	.001	.008	.001	.	.048	.169	.001	.000	.039	.065	.000	.000	.011	.034
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Holistic3	Correlation Coefficient	.393*	.545**	.452**	.432**	**.460****	**.314***	1.000	**.431****	.477**	.348*	.994**	.503**	.490**	.485**	.927**	.516**
	Sig. (2-tailed)	.012	.000	.003	.005	.003	.048	.	.006	.002	.028	.000	.001	.001	.002	.000	.001
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Holistic4	Correlation Coefficient	.638**	.672**	.591**	.638**	**.625****	**.222**	**.431****	1.000	.609**	.258	.451**	.909**	.604**	.395*	.501**	.881**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.169	.006	.	.000	.108	.004	.000	.000	.012	.001	.000
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Analytic1	Correlation Coefficient	.798**	.876**	.855**	.883**	.990**	.523**	.477**	.609**	1.000	**.561****	**.513****	**.700****	.994**	.638**	.589**	.678**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.001	.002	.000	.	.000	.001	.000	.000	.000	.000	.000
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Analytic2	Correlation Coefficient	.505**	.501**	.540**	.444**	.541**	.991**	.348*	.258	**.561****	1.000	**.362***	**.331***	.571**	.886**	.434**	.374*
	Sig. (2-tailed)	.001	.001	.000	.004	.000	.000	.028	.108	.000	.	.022	.037	.000	.000	.005	.017
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
Analytic3	Correlation Coefficient	.430**	.574**	.480**	.463**	.493**	.328*	.994**	.451**	**.513****	**.362***	1.000	**.523****	.526**	.505**	.935**	.534**
	Sig. (2-tailed)	.006	.000	.002	.003	.001	.039	.000	.004	.001	.022	.	.001	.000	.001	.000	.000
	N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40

Table 7 Reliability of the CPIE, Holistic, Analytic, and PIE Method *(Continued)*

Correlations (Evaluators)																	
			CPIE1	CPIE2	CPIE3	CPIE4	Holistic1	Holistic2	Holistic3	Holistic4	Analytic1	Analytic2	Analytic3	Analytic4	PIE1	PIE2	PIE3	PIE4
	Analytic4	Correlation Coefficient	.675**	.763**	.683**	.736**	.716**	.294	.503**	.909**	**.700****	**.331***	**.523****	1.000	.696**	.471**	.590**	.974**
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.065	.001	.000	.000	.037	.001	.	.000	.002	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	PIE1	Correlation Coefficient	.803**	.879**	.855**	.883**	.987**	.532**	.490**	.604**	.994**	.571**	.526**	.696**	1.000	**.646****	**.605****	**.687****
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.	.000	.000	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	PIE2	Correlation Coefficient	.518**	.591**	.615**	.508**	.619**	.851**	.485**	.395*	.638**	.886**	.505**	.471**	**.646****	1.000	**.560****	**.525****
		Sig. (2-tailed)	.001	.000	.000	.001	.000	.000	.002	.012	.000	.000	.001	.002	.000	.	.000	.001
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	PIE3	Correlation Coefficient	.488**	.642**	.563**	.560**	.572**	.398*	.927**	.501**	.589**	.434**	.935**	.590**	**.605****	**.560****	1.000	**.607****
		Sig. (2-tailed)	.001	.000	.000	.000	.000	.011	.000	.001	.000	.005	.000	.000	.000	.000	.	.000
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40
	PIE4	Correlation Coefficient	.661**	.740**	.671**	.707**	.694**	.336*	.516**	.881**	.678**	.374*	.534**	.974**	**.687****	**.525****	**.607****	1.000
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.034	.001	.000	.000	.017	.000	.000	.000	.001	.000	.
		N	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40	40

Abbreviations
CDI: Calibration of dichotomous items; CoNLL: Computational Natural Language Learning; CoreNLP: Core-Natural Language Processing; CPIE: Calibrated Parsing Items Evaluation; CTIC: Council of Translators and Interpreters in Canada; df: Degree of freedom; d-index: Item discrimination; DT: Determiner; HG: Higher Group; IN: Preposition; ITR: International Translation Resources; LG: Lower group; LISA QA: Localization Industry Standards Association Quality Approach; MQM: Multidimensional Quality Metrics; NAATI: National Accreditation Authority for Translators and Interpreters; NB: Nota Bene; NN: Singular noun; NNP: Proper noun; p-docimology: Probability-docimology; PIE: Preselected Items Evaluation; POS: Part of speech; SICAL: Canadian Language Quality Measurement System (English Translation); ST: Source text; TL: Target language; TranslationQ: Translation quality; TS: Translation studies; TTX: Tradostags; VBD: Past tense verb; XLIFF: XML-based localization interchange file format

Funding
The authors received no funding for this article.

Availability of data and materials
The dataset analyzed during the current study are not publicly available because they will be used in a PhD dissertation but are available from the corresponding author on reasonable request.

Authors' contributions
All authors made a contribution to this manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 21 January 2019 Accepted: 7 April 2019
Published online: 31 May 2019

References
Akbari, A. (2017a). *The software tool TranslationQ (Televic-KU Leuven) and the translations from English to Persian*. CIUTI Forum 2017. Geneva: United Nations.
Akbari, A. (2017b). Docimologically Justified Parsing Items: Introducing a new method of translation evaluation." Translation and interpreting in transition 3, Ghent University, 13 and 14 July.
Akbari, A., & Segers, W. (2017a). Translation difficulty: How to measure and what to measure. *Lebende Sprachen, 62*(1), 3–29.
Akbari, A., & Segers, W. (2017b). Translation evaluation methods and the end-product: Which one paves the way for a more reliable and objective assessment? *Skase Journal of Translation and Interpretation, 11*(1), 2–24.
Akbari, A., & Segers, W. (2017c). Evaluation of translation through the proposal of error typology: An explanatory attempt. *Lebende Sprachen, 62*(2), 408–430.
Anckaert, P., Eyckmans, J., & Segers, W. (2008). Pour Une Évaluation Normative De La Compétence De Traduction. *ITL - International Journal of Applied Linguistics, 155*(1), 53–76. https://doi.org/10.2143/ITL.155.0.2032361.
Bahameed, A. S. (2016). Applying assessment holistic method to the translation exam in Yemen. *Babel, 62*(1), 135–149. https://doi.org/10.1075/babel.62.1.08bah.
Brat. (2014). Brat features http://brat.nlplab.org/features.html.
Bunt, H., Merlo, P., & Nivre, J. (2010). *Trends in parsing technology: Dependency parsing, domain adaptation, and deep parsing*. London/New York: Springer.
Conde Ruano, T. (2005). No Me Parece Mal. Comportamiento y Resultados de Estudiantes al Evaluartraducciones. University of Granada: Unpublished doctoral dissertation.
CoNLL. (2012). Conference on computational natural language learning.
D'Agostino, R., & Cureton, E. (1975). The 27 percent rule revisited. *Educational and Psychological Measurement, 35*, 47–50.
Exam, Understanding Your. (2017). Understanding your exam analysis report. PennSatate. https://www.schreyerinstitute.psu.edu/scanning/UnderstandingExamAnalysisReport.
Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice* (pp. 73–93). Amsterdam/Philadelphia: John Benjamins.
Eyckmans, J., Anckaert, P. & Segers, W. (2013). Assessing translation competence. Actualizaciones en Comunicación Social, Centro de Lingüística Aplicada, Santiago de Cuba (2), 513–515.
Eyckmans, J., Segers, W., & Anckaert, P. (2012). Translation assessment methodology and the prospects of European collaboration. In D. Tsagari & I. Csépes (Eds.), *Collaboration in language testing and assessment* (pp. 171–184). Bruxelles: Peter Lang.
Feldt, L. S. (1993). The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education, 6*(1), 37–48. https://doi.org/10.1207/s15324818ame0601_3.
Garant, M. (2010). A case for holistic translation assessment. *AFinLA-e: Soveltavan kielitieteen tutkimuksia*:5–17%N 1.
Gonzalez, K. (2019). Contrast Effect: Definition & Example. https://study.com/academy/lesson/contrast-effect-definition-example.html.
Gouadec, D. (1981). Paramètres de l'évaluation des traductions. *Meta, 26*(2), 99–116.
Gouadec, D. (1989). Comprendre, évaluer, prévenir : Pratique, enseignement et recherche face à l'erreur et à la faute en traduction. *TTR, 2*(2), 35–54.
Hatim, B., & Mason, I. (1997). *The translator as communicator*. London/New York: Routledge.

Kockaert, H., & Segers, W. (2014). Evaluation de la Traduction: La Méthode PIE (preselected items evaluation). *Turjuman, 23*(2), 232–250.
Kockaert, H., & Segers, W. (2017). Evaluation of legal translations: PIE method (preselected items evaluation). *Journal of Specialized Translation*, (27), 148–163 https://www.jostrans.org/issue27/art_kockaert.php.
Lei, P., & Wu, Q. (2007). CTTITEM: SAS macro and SPSS syntax for classical item analysis. *Behavior Research Methods, 39*(3), 527–530. https://doi.org/10.3758/bf03193021.
Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136.
Mariana, V., Cox, T., & Melby, A. (2015). The Multidimensional Quality Metrics (MQM) framework: A new framework for translation quality assessment. *Journal of Specialized Translation*, (23), 137–161 https://www.jostrans.org/issue23/art_melby.php.
Matlock-Hetzel, S. (1997). *Basic concepts in item and test analysis*. Austin: Annual meeting of the southwest educational research association 23–25 January.
McKenna, M. C., & Dougherty Stahl, K. A. (2015). *Assessment for Reading instruction* (Third ed.). New York: The Guilford Press.
Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart and Winston.
Miller, M. D., Linn, R. L., & Grounlund, N. E. (2009). *Measurement and assessment in teaching*. Upper Saddle River: Pearson Education.
Minitab (2017). https://www.minitab.com/en-us/products/minitab/.
Morales, P. (2000). *Medición de Actitudes en Psicología y Educación*. Madrid: Universidad Pontificia Comillas.
Muñoz Martín, R. (2010). On paradigms and cognitive Translatology. In G. Schreve & E. Angelone (Eds.), *Translation and cognition* (pp. 169–187). Amsterdam and Philadelphia: John Benhamins.
Newmark, P. (1991). *About translation*. Clevedon: Multilingual Matters.
O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education, 9*(3), 325–335. https://doi.org/10.1080/1356251042000216642.
Sabri, S. (2013). Item analysis of student comprehensive test for researchin teaching beginner string ensemble using model based teaching among MusicStudents in public universities. *International Journal of Education and Research, 1*(12), 91–104.
Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175–194. https://doi.org/10.1080/0260293042000264262.
Schmitt, P. A. (2005). *Qualitätsbeurteilung von Fachübersetzungen in der Übersetzerausbildung. Probleme und Methoden, paper presented at Vertaaldagen Hoger Instituut voor Vertalers en Tolken, the Netherlands, 16-17 March 2005*.
Shay, S. (2008). Beyond social constructivist perspectives on assessment: The centring of knowledge. *Teaching in Higher Education, 13*(5), 595–605. https://doi.org/10.1080/13562510802334970.
SPSS. (2017). https://www.ibm.com/analytics/spss-statistics-software.
Stansfield, C. W., Scott, M. L., & Kenyon, D. M. (1992). The measurement of translation ability. *The Modern Language Journal, 76*(4), 455–467. https://doi.org/10.2307/330046.
Stenetorp, P., Pyysalo, S., Topi, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). *BRAT: A web-based tool for NLP-assisted text annotation*. Avignon: Proceedings of the demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics.
Tang, S. F., & Logonnathan, L. (2016). *Assessment for learning within and beyond the classroom: Taylor's 8th teaching and learning conference 2015 proceedings*. Singapore: Springer.
Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 46–80). Washington, DC: American Council on Education.
Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta, 46*(2), 311–325.
Waddington, C. (2004). Should Translations be Assessed Holistically or through Error Analysis? *Lebende Sprachen, 49*(1), 28–35.
Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.
Williams, M. (2013). A holistic-componential model for assessing translation student performance and competency. *Mutatis Mutandis, 6*(2), 419–443.
Williams, R. (2016). Outliers. www3.nd.edu/~rwilliam/stats2/l24.pdf.