

A systematic review of experimental research in audiovisual translation 1992–2020

Abstract: This study systematically reviewed 61 experimental studies on audiovisual translation (AVT), published in academic journals between 1992 and 2020. The review examined five aspects of these studies: publication trends, research scopes, research designs, statistical procedures, and reporting practices. Major findings include: (a) there has been a slight concentration of publication outlets for experimental AVT research; (b) the focal studies could be categorized into three themes (product, process, and pedagogy), with the product theme being the most popular and subtitling the dominant AVT modality; (c) the inclusion of a comparison group was the most common design feature, and questionnaires and tests were the most popular research instruments; (d) inferential statistical analysis was favored over descriptive statistical analysis; (e) data normality information and effect sizes were not regularly reported. Based on the systematic review, suggestions are made for the future development of experimental AVT research.

Keywords: systematic review, audiovisual translation, experiment, research synthesis

1. Introduction

Audiovisual translation (AVT) research can be dated back to Laks (1957), who wrote about the techniques and aesthetics of subtitling films. More than 60 years later, AVT research has developed into “a legitimate and independent field” in translation studies (Chaume 2018, 41). The vibrancy of this field is indicated by an increasing number of thematic journal issues, edited volumes, handbooks, monographs, and collaborative projects joined by international scholars and/or industry partners (Pérez-González 2019). For instance, a search in the Bibliography of Interpreting and Translation (BITRA) showed that the growth of AVT publications was “exponential” in the 2010s, with AVT publications already taking up 9.8% of the total in the period (2011 to mid-2016) as compared to 6.7% in 2001-2010 (Orero et al 2018, 117). The sustained development of the field has been driven by new research tools, methods, and paradigms. Chaume (2018) outlined four paradigm shifts in the six-decade AVT research, or what he called *descriptive turn*, *cultural turn*, *social turn*, and *cognitive and empirical turn*. Chaume (2018) also pointed out that the “cognitive and empirical turn...is gaining ground” (42). Typically, studies in this paradigm employ an experimental design with a view to understanding the translator’s cognitive process and the audience’s reception of AVT products. As a sizable number of experimental AVT studies have been produced, it is necessary to review and reflect on research practices for future directions. As such, this paper reports on a systematic review of experimental AVT studies to understand the publication trends, research scopes, experimental designs, statistical procedures, and reporting practices in these studies. In the following sections, we will first make a case for a systematic review (Section 2) and then we will introduce the parameters of the review (Section 3). In Section 4, we will report the findings and relate them to the discussion of observed patterns and suggestions for experimental AVT research in the future (Section 5).

2. Making a case for a systematic review

There is no lack of reviews in the AVT literature, with two most recent ones being Orrego-Carmona (2019) and Di Giovanni (2020). These reviews take stock of what AVT research has achieved and offer suggestions to move the field forward. For instance, Orrego-Carmona (2019) made observation of a possible rise of scholarly interest in other AVT modalities besides subtitling. Di Giovanni (2020) called for a move away from the “exclusive recourse to university students” as participants in AVT studies (408). While these reviews and similar others provide valuable insights, they are by nature *narrative reviews*, in which the selection of studies is “left to the tacit good sense of the expert reviewer” (Norris & Ortega 2007, 807). Additionally, narrative reviews tend to summarize the results of individual studies, but tend not to answer such questions as “what is the mean sample size of the focal studies?” and “how many studies have used eye-tracking as a research tool?” To obtain these insights, *systematic reviews* are needed.

A systematic review is different from a narrative review in three aspects (Norris & Ortega 2007).

First, sampling of studies is guided by inclusion and exclusion criteria established before the review process. Second, primary data reported in the studies are aggregated, scrutinized, and reanalyzed. Third, a coding protocol is developed to distill and discern features in individual studies and synthesize them into patterns. An example of a systematic review is Kruger & Doherty (2018), focusing on AVT studies that have used eye-tracking. The review compared information about eye-tracking measures (e.g., dwell time and fixation duration), other measures (e.g., questionnaires on attitudes and self-rated cognitive load) and whether the eye-tracking data were triangulated.

Clearly, reviews of this kind focus less on the commentary narration of (groups of) individual studies, but more on developing a synthetic understanding of a research domain. In light of this benefit, the present study systematically examined experimental AVT research from a pre-defined list of features, informed by narrative reviews in the AVT literature (e.g., Orrego-Carmona 2019; Di Giovanni 2020) and systematic reviews in the neighboring discipline (e.g., Avgousti 2018; Plonsky 2014, in applied linguistics). Specifically, our review attempts to address the following questions:

1. What have been the publication trends of experimental AVT research?
2. What have been the research scopes of experimental AVT research?
3. How have AVT experiments been designed?
4. What statistical procedures have been used in experimental AVT studies?
5. What reporting practices have been employed in experimental AVT studies?

3. Methodology

3.1 Selection criteria

As explained in Section 2, inclusion and exclusion criteria are essential for a systematic review. To formulate inclusion criteria, we deliberated four aspects: publication type, publication language, research design, and research scope. First, only journal articles were considered in this study, because journals have been the primary channel for research dissemination (Grbić & Pöllabauer 2008). It would have been ideal to include other publication types, such as conference proceedings, book chapters, monographs, and dissertations. However, balancing the resources available/accessible to us and the objectives of the present review, we believe that journal articles provide a fairly representative picture of experimental AVT research. Second, articles written in English were included, because over the past decades English has been the dominant language in international publications (Hamel 2007) and in translation and interpreting studies (Yan, Pan, & Wang 2018). Third, both experimental and quasi-experimental studies were included in this review. An experimental study is characterized by two elements: (a) controlling and manipulating the independent variable(s); and (b) randomly assigning participants into treatment conditions (Mellinger & Hanson 2017). Relatedly, a quasi-experimental design does not control the independent variable(s) or allow for random assignment (Mellinger & Hanson 2017). It should be noted that (quasi-)experimental studies are not necessarily quantitative (Gile 2016). Hence, (quasi-)experimental, qualitative studies were also included in this review. Fourth, the review covered studies on interlingual AVT, because interlingual translation “has been the traditional focus of translation studies” (Munday 2016, 10) and AVT has been “mainly concerned with the transfer of multimodal and multimedia speech...into *another* language/culture” (Gambier 2013, 45, our emphasis).

Based on the previous deliberation, the inclusion criteria were formulated as follows. Studies were included if they were (a) published in journals, (b) written in English, (c) (quasi-)experimental, and (d) about interlingual AVT.

Two databases were used for the systematic review. The first one was BITRA, a monthly-updated open database currently comprising more than 81,000 entries (Franco Aixelá 2001-2020). The second one was Translation Studies Bibliography, another regularly updated database, with more than 30,000 references (Gambier & van Doorslaer n.d.). These two databases have been frequently used to locate and review translation and interpreting studies (e.g., Olalla-Soler 2020; Rogers 2018; Roser 2016).

Two groups of keywords were used to locate AVT experimental studies. The first group contained seven words, the first of which was *audiovisual*, the most straightforward word for the

search, and the other six were major AVT modalities, i.e., *subtitling*, *dubbing*, *audio description*, *voice-over*, *respeaking*, and *surtitling* (Chaume 2018).¹ The second group contained signal words indicative of experimental studies, i.e., *experiment*, *participant*, and *reception*. In fact, we also piloted the keywords *study* and *subject*. However, due to a large number of false hits (with *study* as a catch-all term and *subject* a polysemy), we decided not to use these two keywords. To perform searches in the databases, abstracts were searched for any hit of the 8 (*audiovisual*, *subtitl**, *dub**, *audio descri**, *voice?over*, *respeak**, *respok**, and *surtitl**) × 3 (*experiment**, *participant**, and *reception*) combination. Two wild cards were used to cover possible variations, with the asterisk (*) standing for zero or more characters and the question mark (?) representing zero or one character. The publication language was set as “English” and no date restrictions for publication were set. A total of 1073 hits were returned (as of October 25, 2020).

To exclude irrelevant data, the 1073 hits were subjected to the exclusion process shown in Figure 1. First, 668 duplicates were discarded, leaving 405 studies for further screening. Then, the title and abstract of each remaining study were screened based on the following exclusion criteria: (a) the study did not focus on the context of interlingual AVT; (b) the study did not employ a (quasi-)experimental method; (c) the study was not published in a journal. Based on these criteria, 311 studies were excluded, with 94 studies remained for full-text examination. One article’s full text was not accessible and thus excluded. After the full texts were examined, 32 more studies were removed because they were not (quasi-)experimental or did not focus on interlingual AVT. Finally, a total of 61 studies (reporting 69 experiments) were included for the systematic review, with the earliest article published in 1992 and the latest ones in 2020.

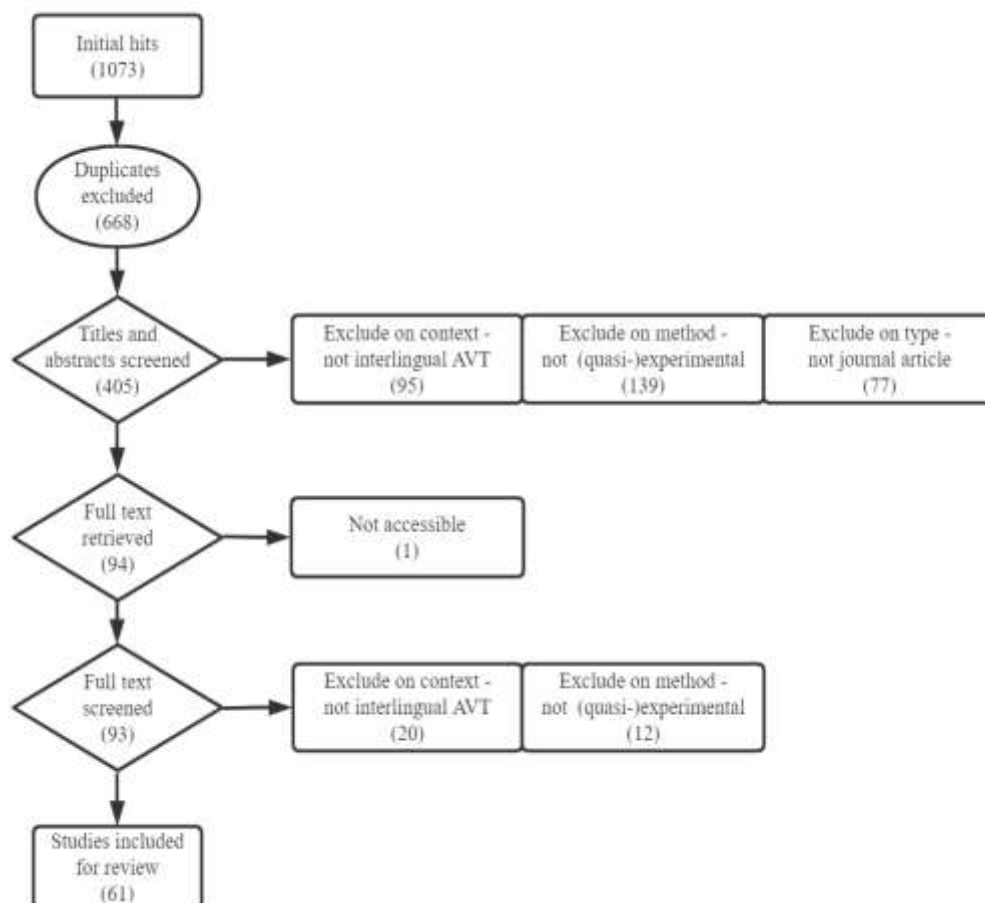


Figure 1. Data selection process

¹ It should be noted that although audio description (AD) has been traditionally conceptualized as intersemiotic translation, a few AD studies have focused on interlingual transfer (e.g., Walczak & Fryer 2018) and were thus included in this review.

3.2 Coding protocol

To address the research questions, a coding protocol was adapted from Avgousti (2018) and Plonsky (2014). The protocol comprised five major categories: (a) publication trends, (b) research scopes, (c) research designs, (d) statistical procedures and (e) reporting practices. Subcategories and example items (if any) are shown in Table 1.

Table 1. Coding protocol for the focal studies

Category	Subcategory	Item
Publication trends	Journals	
	Publication year	
Research scopes	Themes	Pedagogy, product, process
	AVT modalities	Subtitling, dubbing, audio description, voice-over, respeaking, surtitling
Research designs	Participant characteristics	Gender, age, educational/working status
	Sample size	
	Number of groups	
	Design features	Comparison group, pretest, posttest, delayed posttest, random assignment
Statistical procedures	Research instruments	Questionnaire, test, interview, eye-tracking, observation, EEG, Inputlog, etc.
	Descriptive statistical procedures	
Reporting practices	Inferential statistical procedures	ANOVA, ANCOVA, t-test, correlation, chi-square, regression, etc.
	Descriptive statistics reported	Sample size, frequency, percentage, mean, standard deviation, effect size
	Inferential statistics reported	<i>p</i> value

The coding protocol was first piloted on two articles (reporting four experiments) by the second author. Then, a calibration session was held between the authors to ensure that the coding protocol was valid to capture the information relevant to the research questions. Next, the second author coded all the 61 studies, which were checked by the first author. On a few occasions, coding disagreements were derived from different labels used by the researchers of the studies. For example, *testing* items were administered as part of a questionnaire and were labeled as *questionnaire* items. We resolved these coding disagreements through discussion.

4. Results

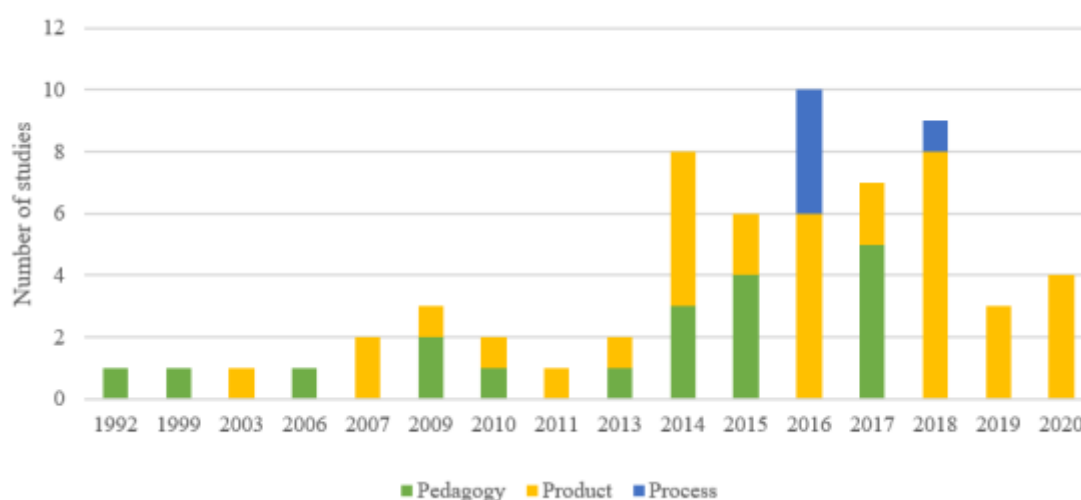
4.1 What have been the publication trends of experimental AVT research?

Orero et al (2018) observed that “some translation journals tend to host AVT papers more often than others” (112). Our systematic review further showed that some journals tended to publish more experimental AVT studies than others. As can be seen in Table 2, *The Journal of Specialised Translation* (also known as *JoSTrans*) published the largest number ($k=9$) of studies in our sample, followed by other translation journals, such as *Across Languages and Cultures*, *Babel*, and *Linguistica Antverpiensia, New Series – Themes in Translation Studies*. In addition, interdisciplinary journals (e.g., *Media Psychology* and *PLoS ONE*) were possible outlets. It is interesting to note that 34% of the journals (13 out of 38) have published at least two focal studies, accounting for about 60% of the total sample (36 out of 61). This seems to suggest that while publication outlets are diverse, experimental AVT studies are slightly concentrated in a dozen of journals, but not highly concentrated (e.g., hypothetically, 80% of articles in 20% of journals).

Table 2. Publication outlets for experimental AVT studies ($k=61$)

Journal title	k
<i>The Journal of Specialised Translation</i>	9
<i>Across Languages and Cultures</i>	3
<i>Babel</i>	3
<i>Linguistica Antverpiensia, New Series – Themes in Translation Studies</i>	3
<i>inTRAlinea</i>	2
<i>Lingue e Linguaggi</i>	2
<i>Media Psychology</i>	2
<i>Perspectives: Studies in Translation Theory and Practice</i>	2
<i>PLoS ONE</i>	2
<i>SKASE Journal of translation and interpretation</i>	2
<i>The Interpreter and Translator Trainer</i>	2
<i>The Translator</i>	2
<i>Translation, Cognition & Behavior</i>	2
Other journals	25

From a diachronic perspective, the 61 focal studies were published between 1992 and 2020. As shown in Figure 2, a surge of publications was observed in 2014. It is clearly a watershed year for the three-decade period, because the number of publications from 2014 to 2020 takes up 77% of the total sample (47 out of 61). It is fair to say that the past decade has witnessed increased scholarly interest in experimental interlingual AVT research.

**Figure 2.** Number of publications from 1992 to 2020

4.2 What have been the research scopes of experimental AVT research?

The focal studies were coded on three themes (see also Table 3). We used the role of the participants as the point of differentiation. The pedagogy theme focused on the extent to which exposure to AVT products developed learners' (foreign) language skills and translation skills. The product theme examined audiences' reception and comprehension of AVT products. The process theme explored how professional translators and trainees created AVT products. The process and product themes corresponded to what Di Giovanni & Gambier (2018, x) called "the *before* and *after* of audiovisual texts" (emphasis in the original).

According to this tripartite categorization, studies on the product theme take up about 61% of the total sample (37 out of 61), followed by studies on the pedagogy theme (31%). Studies on the process theme account for a mere 8%. Figure 2 further shows that the pedagogy theme and the product theme have attracted sustained scholarly attention, while the process theme emerged in the second half of 2010s. Echoing the findings reported in the previous subsection, 2014 is the

watershed year for the product theme, the growth of which is remarkable.

Table 3. Research themes in experimental AVT research

Themes	Roles of participants	Example research questions	Example studies
Pedagogy	Learners of language skills or translation skills	To what extent do standard subtitled videos and reverse subtitled videos enhance vocabulary learning? To what extent can reverse subtitling develop translation and writing skills?	Danan (1992) Talaván & Rodríguez-Arancón (2014) Božović (2019)
Product	Audience of AVT products	Do viewers have different expectations and preferences of translation strategies for subtitled and dubbed products? How do speech rates affect audiences' reading patterns of subtitles and film comprehension?	Szarkowska & Bogucka (2019)
Process	Creators of AVT products	How do professional subtitlers and subtitling trainees differ in the subtitling process? How do interlingual respeaking and intralingual respeaking differ in cognitive load?	Orrego-Carmona, Dutka & Szarkowska (2018) Szarkowska, Krejtz, Dutka & Pilipczuk (2016)

With respect to AVT modalities, Table 4 shows that dominant attention has been given to subtitling (72%, 44 out of 61 studies), followed by dubbing (34%). The remaining modalities (i.e., audio description, voice-over, and respeaking) each account for less than 10%. Surtitling, although searched in the databases, was not examined in experimental studies. In terms of the number of AVT modalities examined within a study, 45 studies focused on one single modality, while 16 focused on two.

Table 4. AVT modalities examined in experimental studies

AVT modalities	<i>k</i>	%*
Subtitling	44	72
Dubbing	21	34
Voice-over	5	8
Audio description	4	7
Respeaking	3	5
Surtitling	0	0

* *Note.* 16 studies investigated two AVT modalities in the same experiments, so the total percentage exceeds 100%.

4.3 How have AVT experiments been designed?

4.3.1 Participant characteristics

Of the 69 experiments reported by the focal studies, 36 provided information about the mean age of the participants, while 17 experiments provided an age range. Overall, the age of the participants varied, with the youngest being 8 years old and the oldest 85. The mean age of the 36 experiments was 27 (SD=8.2). Among the 47 experiments that reported gender information, a total of 986 males and 1,785 females were involved. In other words, females were more represented than males.

53 experiments reported the participants' educational and/or working status. The majority

(87%, $k=46$) of experiments focused on university students, along with a small number of experiments examining translation professionals/trainers (15%, $k=8$) and other working professionals ($k=7$). There were occasional cases in which primary school students ($k=4$) and middle school students ($k=3$) were involved.

4.3.2 *Sample size*

A wide array of sample sizes have been reported, from the smallest size of 4 to as large as 312. The mean sample size of the 69 experiments was 60.3. However, when 53 experiments with a comparison group were considered, the mean subgroup sample size was 20.4. If the experiments were divided into two periods (before and after 2010s), the mean sample size was 97.6 and the mean subgroup sample size was 16.4 in the period of 1992-2010. Comparatively, the mean sample size was 50.7 and the mean subgroup sample size was 21.7 in the period of 2011-2020. This seems to suggest that while the total sample size within a study decreased, the subgroup sample size increased. It should also be noted that in the 20 experiments that reported initial and final sample sizes, the average attrition rate was 16.3%.

4.3.3 *Design features*

According to Plonsky (2013, 667) and Plonsky (2014, 452), four design features are important to enhance the quality of an experiment: assigning participants randomly, including a comparison group, conducting a pretest, and conducting a delayed posttest. In light of this, five design features were coded in our sample: (a) comparison group, (b) random assignment, (c) pretest, (d) posttest, and (e) delayed posttest. As it turned out, 77% of the experiments (53 out of 69) included a comparison group. The next most frequent feature was posttest (64%, $k=44$), while random assignment was adopted by one third of the experiments ($k=23$). Pretests ($k=8$) and delayed posttests ($k=3$) were rarely conducted. Although all the three delayed posttests were found in the experiments with a pedagogy theme, we were somewhat surprised by the small number of use. Delayed posttests would have been useful to examine the extent to which a treatment effect lasted in relation to the acquisition of a language/translation skill after the participants had been exposed to AVT contents.

In terms of the combination of design features within one experiment, it is interesting to note that a comparison group and a posttest were the most frequent combination ($k=40$). This might be the most popular research design: comparing how two or more groups of participants differed (e.g., in the comprehension of different versions of a video) immediately after the treatment (e.g., watching a subtitled video vs. a dubbed video).

4.3.4 *Research instruments*

Among the 69 experiments, the top two research instruments were questionnaires (86%, $k=59$) and tests (62%, $k=43$), while eye-tracking was a distant third (26%, $k=18$). EEG and Inputlog (a research tool to log keyboard, mouse, and speech input data, Leijten & Van Waes 2013) were also added to the researchers' repertoire, although much less popular (2.9% each). In terms of the number of research instruments within an experiment (a sign pointing to data triangulation), 68% used more than one instrument ($k=47$), while about one third used one instrument ($k=22$).

4.4 *What statistical procedures have been used in experimental AVT studies?*

One experiment did not resort to any statistical procedure. 19 experiments (28%) relied on descriptive statistical procedures only, while 49 experiments (71%) used at least one inferential statistical procedure. Of these 49 experiments, more than half of them used ANOVAs ($k=28$), making it the most popular choice. The next most frequently used inferential statistical procedures were correlation and t-tests ($k=18$ each). There were occasional uses of regression ($k=7$), chi-square ($k=5$), and ANCOVA ($k=4$). Only four studies used nonparametric tests, a potential issue we will circle back to.

4.5 *What reporting practices have been employed in experimental AVT studies?*

All 69 experiments reported their sample sizes and the majority of them ($k=60$) used percentages to represent the data. Additionally, 58 experiments reported mean values and 39 reported standard deviations. It should be noted that p values were reported in 48 experiments, as compared to effect sizes reported in 21 experiments. Interestingly, more than half of studies that reported effect sizes ($k=12$) were published from 2018 onwards. There seems to be a recent trend to move beyond the

traditional focus on *p* values to include effect sizes in experimental AVT studies.

Of the 42 experiments that used ANOVAs, t-tests, or ANCOVAs, 36 did not report whether the data were normally distributed. Comparatively, all the four experiments that used nonparametric tests reported the normality information.

5. Discussion

5.1 Publication trends

We found a variety of journals that published experimental AVT studies over the past three decades. More specifically, about 60% of the articles were published in 34% of the journals, indicating a slight preference and concentration of publication outlets. *The Journal of Specialised Translation* came out as the top host. Some long-standing journals (e.g., *Babel*) and recently launched ones (e.g., *Translation, Cognition & Behavior*) have also contributed to the dissemination of experimental AVT research. The diverse mix of journals as publication outlets points to a burgeoning trend of experimental research in AVT (Díaz Cintas & Szarkowska 2020).

As Snell-Hornby (2006) observed, researchers called for more empirical studies in the field of translation and interpreting studies, thus giving rise to the “empirical turn” in the 1990s. It so happened that the earliest experimental AVT study in our sample was published in 1992. However, it was not until 2014 (two decades later) that the number of journal articles on experimental AVT research started to gain momentum. Although the numbers in 2019 and 2020 seem to dip (see Figure 2), we posit that this might be caused by the time lapse between the publication of journal articles and their inclusion in the two databases (as the sources of the present review). We suggest waiting for a couple of years to see whether the momentum will trail off or sustain.

5.2 Research scopes

Based on the three-theme categorization, our review found that the product theme (in which the participants were audiences of AVT contents) was the most popular, while the process theme (in which the participants were creators of AVT contents) was the least. More studies on examining the (cognitive) processes of creating AVT contents are warranted so that we will have an in-depth understanding of the “black box” of audiovisual practitioners. For instance, Ortiz-Boix & Matamala (2016) used Inputlog to examine the cognitive effort in post-editing machine-translated texts and translating source texts for documentary films. Studies of this kind can shed important light on how the audiovisual translation process unfolds and how audiovisual translators expend energy on cognitive and metacognitive decision-making. Therefore, we call for more research on the process theme.

With respect to AVT modalities, it was found that experimental research largely focused on subtitling throughout the three decades. Dubbing, albeit ranked the second, received only half of the attention given to subtitling (21 studies vs. 44 studies). This lends empirical support to Di Giovanni's (2020) observation that dubbing is “the child of a lesser god” in AVT reception studies (405). Audio description, voice-over, and respoken have received scarce attention, and surtitling none at all. This discrepancy might be explained by four factors: supply, cost, demand, and ecological validity. First, preexisting subtitled and dubbed products are easier to obtain, which provide a variety of source materials for researchers to manipulate in experiments. The variety and number of AVT products in other modalities are relatively smaller, thus limiting the choices of experimental materials. Second, relatedly, the production of audiovisual contents in alternative modalities is expensive (e.g., audio description, Szarkowska 2011). If audiovisual materials in alternative modalities are not available, it is costly to produce them for the purposes of experiments. Third, the market demands of these alternative AVT modalities might be comparatively small (Szarkowska 2011), which makes it difficult for researchers to recruit participants for studies focusing on AVT products. Fourth, surtitling has not been examined in experiments probably because its context of use is difficult to be simulated. As surtitles are usually used in theatre and opera performances (Gambier 2013), it is challenging to simulate the use experience in a lab or a normal classroom. Notwithstanding the difficulties previously outlined, more experimental studies on the AVT modalities other than subtitling and dubbing are needed.

5.3 Research designs

Our review showed that the participants involved in the AVT experiments tended to be female and in their 20s. When the educational/working status was considered, the majority of the studies recruited university students. This choice is understandable since most researchers work at university and convenience sampling is a potential option. However, Di Giovanni (2018) has cautioned us “to avoid recourse to the researcher’s students as targets,” because “their participation to experiments will hardly ever be deprived of bias and could not thus claim to reliability in results” (174). This unwitting sampling bias should be addressed if we want to develop a comprehensive understanding about audience reception affected by a diverse set of demographic variables, such as age, literacy level, and occupation (Li 2019).

Another important issue concerning participants is the sample size. Although our review showed that the mean subgroup sample size improved from 16.4 before 2010 to 21.7 after 2010, it was not optimal. As pointed out by Orero et al (2018), “[s]ample sizes of lower than 25 per group are unlikely to yield statistical power” (110). When data normality is factored, Mellinger & Hanson (2017) suggests “a minimum sample size of 30 or 40” (64). A small sample size might affect normal distribution of data, which is an important assumption to be met before statistical analyses (e.g., ANOVAs and t-tests) are conducted. Additionally, based on the experiments that reported initial and final sample sizes, the average attrition rate of the participants was 16.3%. Taking these factors into consideration, researchers can budget at least 36 participants per group in an AVT experiment.

With respect to the design features that could enhance study quality (Plonsky 2013, 2014), more than three quarters of the experiments included a comparison group and two thirds a posttest. The combination of these two features was adopted by more than half of the experiments. However, more attention should be paid to random assignment (adopted by one third of the experiments) and pretests (conducted by eight experiments). Random assignment can control for confounding variables (Mellinger & Hanson 2017), while a pretest can “test or control for pretreatment differences” (Plonsky 2013, 670), thus ensuring the comparability of two or more groups. Studies focusing on pedagogies are also advised to include a delayed posttest to examine whether the treatment/training effect is sustainable (although a delayed posttest is less relevant for studies focusing on AVT products and processes).

Regarding research instruments, although questionnaires and tests have been the top two choices, eye-tracking has been deployed to collect physiological data about participants’ cognitive processes. Our review also found that about one third of the experiments relied on one research instrument. As the production and reception of AVT contents are complex in nature, researchers are suggested to use multiple research instruments to collect data and triangulate data interpretation (Di Giovanni, 2020; Orero et al 2018).

A final word about research design is a call for replication studies (Díaz Cintas & Szarkowska 2020). Given its controlled nature, experimental research can and should be replicated to verify and/or extend previous findings. Replication studies are still not common in translation and interpreting studies. Olalla-Soler (2020) managed to locate 23 replication studies (as of November 2018), 15 of which were published as journal articles. In our AVT sample, we found one self-labeled replication study (i.e., Flis, Sikorski & Szarkowska 2020). Given the sizable number of experimental AVT publications, more replication studies can be conducted to generate more (nuanced) insights. Some possible ideas for AVT replication studies include (Marsden et al 2018): (a) trying to replicate the research results of a previous study on a larger sample size; (b) collecting and analyzing new data in conjunction with previous data (as in Flis et al 2020); and (c) examining the extent to which previous findings can be replicated on a new sample with different characteristics (e.g., language background, viewing habits, and age).

5.4 Statistical procedures and reporting practices

About 30% of the experiments relied on descriptive statistical procedures only. Researchers are reminded that at times descriptive statistical analysis should be complemented with inferential statistical analysis (Saldanha & O’Brien 2014). For instance, when two groups differ by a small amount (e.g., 5%) in a video comprehension test, inferential statistical procedures are needed to ascertain whether the difference is by chance or indeed caused by the treatment condition.

In the experiments that used inferential statistical procedures, ANOVAs, t-tests and correlation were the top three choices. Although ANOVAs, t-tests, and ANCOVAs require data points to meet the assumption of normal distribution, only six experiments reported the normality information. In

the experiments that did not provide the normality information, it might be possible that the data were indeed normally distributed, so the absence of the information did not necessarily mean the absence of normal distribution. However, given the small sub-group sample size in AVT research, it might be difficult to expect normality (see paragraph 2 in Section 5.3). As such, researchers are suggested to check this assumption before conducting ANOVAs, t-tests, and ANCOVAs (Orero et al 2018). If this assumption is not met, researchers can (a) transform the data so that the normality assumption is met, (b) run nonparametric tests, or (c) bootstrap the data (see Larson-Hall 2016 for a detailed treatment).

Our review found that the number of experiments reporting p values was twice the number of experiments reporting effect sizes. In the field of applied linguistics, researchers have been suggested to include effect size estimates along with significance values (Norris & Ortega 2000; Plonsky 2013). In the field of AVT, researchers are also suggested to move away from the sole focus on the p values to greater inclusion of effect sizes (Díaz Cintas & Szarkowska 2020). This would change the mindset of being fixated on the presence of an effect (e.g., $p < 0.05$) to deliberating the magnitude of an effect.

6. Conclusion

Recent development of AVT research has been characterized as “a thrust beyond descriptivism” (Di Giovanni & Gambier 2018, x), which “has led to an unprecedented boom in experimental research” (Díaz Cintas & Szarkowska 2020, 4). As the experimental approach has become a central driver in the field, it is timely to develop a synthetic account of experimental AVT research. To this end, we systematically reviewed 61 journal articles published between 1992 and 2020. We charted the publication trends, identified the research scopes, examined research designs, and took stock of statistical procedures and reporting practices. We acknowledge that our findings are potentially limited by (a) the choice of databases and (b) the inclusion of journal articles explicitly using our search terms in the abstracts (see also Olalla-Soler 2020), thus leaving some experimental AVT studies unexamined. Despite these limitations, we hope that our study has demonstrated the heuristic values of systematic reviews and similar reviews of this kind will be conducted to include more databases (e.g., Scopus and Web of Science) and more diverse publication types (e.g., dissertations, monographs and book chapters). The patterns distilled from the present and future systematic reviews can and will generate insights to inform the AVT research community and beyond.

References

- Avgousti, Maria Iosifina. 2018. “Intercultural Communicative Competence and Online Exchanges: A Systematic Review.” *Computer Assisted Language Learning* 31 (8): 819-853.
- Božović, Petar. 2019. “How Should Culture Be Rendered in Subtitling and Dubbing? A Reception Study on Preferences and Attitudes of End-Users.” *Babel* 65 (1): 81-95.
- Chaume, Frederic. 2018. “An Overview of Audiovisual Translation: Four Methodological Turns in a Mature Discipline.” *Journal of Audiovisual Translation* 1 (1): 40-63.
- Danan, Martine. 1992. “Reversed Subtitling and Dual Coding Theory: New Directions for Foreign Language Instruction.” *Language Learning* 42 (4): 497-527.
- Di Giovanni, Elena. 2018. “Dubbing, Perception and Reception.” In *Reception Studies and Audiovisual Translation*, edited by Elena Di Giovanni and Yves Gambier, 159-177. John Benjamins.
- Di Giovanni, Elena. 2020. “Reception Studies and Audiovisual Translation.” In *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*, edited by Łukasz Bogucki and Mikołaj Deckert, 397-413. Palgrave Macmillan.
- Di Giovanni, Elena, and Yves Gambier. 2018. “Introduction.” In *Reception Studies and Audiovisual Translation*, edited by Elena Di Giovanni and Yves Gambier, vii-xii. John Benjamins.
- Díaz Cintas, Jorge, and Agnieszka Szarkowska. 2020. “Introduction: Experimental Research in Audiovisual Translation—Cognition, Reception, Production.” *The Journal of Specialised Translation* 33: 3-16.
- Flis, Gabriela, Adam Sikorski, and Agnieszka Szarkowska. 2020. “Does the Dubbing Effect

- Apply to Voice-over: A Conceptual Replication Study on Visual Attention and Immersion.” *The Journal of Specialised Translation* 33: 41-69.
- Franco Aixelá, Javier. 2001–2019. BITRA (Bibliography of Interpreting and Translation). Accessed October 15, 2020. <https://dti.ua.es/en/bitra/introduction.html>
- Gambier, Yves, and Luc van Doorslaer. (n.d). Translation studies bibliography. Accessed October 15, 2020. <http://benjamins.com/online/tsb/>
- Gambier, Yves. 2013. “The Position of Audiovisual Translation Studies.” In *The Routledge Handbook of Translation Studies*, edited by Carmen Millán and Francesca Bartrina, 45-59. Routledge.
- Gile, Daniel. 2016. “Experimental Research.” In *Researching Translation and Interpreting*, edited by Claudia V. Angelelli and Brian James Baer, 220-228. Routledge.
- Grbić, Nadja, and Sonja Pöllabauer. 2008. “Counting What Counts: Research on Community Interpreting In German-Speaking Countries—A Scientometric Study.” *Target* 20 (2): 297-332.
- Larson-Hall, Jenifer. 2016. *A Guide to Doing Statistics in Second Language Research Using SPSS and R* (2nd ed.). Routledge.
- Hamel, Rainer Enrique. 2007. “The Dominance of English in the International Scientific Periodical Literature and the Future of Language Use in Science.” *AILA Review* 20(1): 53-71.
- Kruger, Jan-Louis, and Stephen Doherty. 2018. “Triangulation of Online and Offline Measures of Processing and Reception in AVT.” In *Reception Studies and Audiovisual Translation*, edited by Elena Di Giovanni and Yves Gambier, 91-109. John Benjamins.
- Laks, Simon. 1957. Le sous-titrage des films. Sa technique. Son esthétique [Subtitling films. Technique. Aesthetics] (Unpublished manuscript).
- Leijten, Mariëlle, and Luuk Van Waes. 2013. “Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes.” *Written Communication* 30 (3): 358-392.
- Li, Dang. 2019. “Ethnographic Research in Audiovisual Translation.” In *The Routledge Handbook of Audiovisual Translation*, edited by Luis Pérez-González, 383-398. Routledge.
- Marsden, Emma. Kara Morgan-Short, Sophie Thompson, and David Abugaber. 2018. “Replication in Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field.” *Language Learning*, 68(2): 321-391.
- Mellinger, Christopher D., and Thomas A. Hanson. 2017. *Quantitative Research Methods in Translation and Interpreting Studies*. Routledge.
- Munday, Jeremy. 2016. *Introducing Translation Studies: Theories and Applications*. 4th ed. Routledge.
- Norris, John M. and Lourdes Ortega. 2000. “Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis.” *Language Learning*, 50: 417-528.
- Norris, John M. and Lourdes Ortega. 2007. “The Future of Research Synthesis in Applied Linguistics: Beyond Art or Science.” *TESOL Quarterly* 41 (4): 805-815.
- Olalla-Soler, Christian. 2020. “Practices and Attitudes toward Replication in Empirical Translation and Interpreting Studies.” *Target* 32 (1): 3-36.
- Orero, Pilar, Stephen Doherty, Jan-Louis Kruger, Anna Matamala, Jan Pedersen, Elisa Perego, Pablo Romero-Fresco, Sara Rovira-Esteva, Olga Soler-Vilageliu, and Agnieszka Szarkowska. 2018. “Conducting Experimental Research in Audiovisual Translation (AVT): A Position Paper.” *The Journal of Specialised Translation* 30: 105-126.
- Orrego-Carmona, David. 2019. “Audiovisual translation and audience reception.” In *The Routledge Handbook of Audiovisual Translation*, edited by Luis Pérez-González, 367-382. Routledge.
- Orrego-Carmona, David, Łukasz Dutka, and Agnieszka Szarkowska. 2018. “Using Translation Process Research to Explore the Creation of Subtitles: An Eye-Tracking Study Comparing Professional and Trainee Subtitlers.” *The Journal of Specialised Translation* 30: 150-180.
- Ortiz-Boix, Carla, and Anna Matamala. 2016. “Post-Editing Wildlife Documentary Films: A New Possible Scenario.” *The Journal of Specialised Translation* 26: 187-210.
- Pérez-González, Luis. 2019. “Rewiring the Circuitry of Audiovisual Translation: Introduction.” In *The Routledge Handbook of Audiovisual Translation*, edited by Luis Pérez-González, 1-12. Routledge.
- Plonsky, Luke. 2013. “Study Quality in SLA: An Assessment of Designs, Analyses, and

- Reporting Practices in Quantitative L2 Research.” *Studies in Second Language Acquisition* 35(4): 655-687.
- Plonsky, Luke. 2014. “Study Quality in Quantitative L2 Research (1990–2010): A Methodological Synthesis and Call for Reform.” *The Modern Language Journal* 98(1): 450-470.
- Rogers, Margaret. 2018. “Specialised Translation Today: A View from the *JoSTrans* Bridge.” *The Journal of Specialised Translation* (30): 3-22.
- Rosa, Alexandra Assis. 2016. “Descriptive translation studies of audiovisual translation: 21st-century issues, challenges and opportunities.” *Target* 28 (2): 192-205.
- Saldanha, Gabriela, and Sharon O’Brien. 2014. *Research Methodologies in Translation Studies*. Routledge.
- Snell-Hornby, Mary. 2006. *The Turns of Translation Studies: New Paradigms or Shifting Viewpoints?*. John Benjamins.
- Szarkowska, Agnieszka. 2011. “Text-To-Speech Audio Description: Towards Wider Availability of AD.” *The Journal of Specialised Translation* (15): 142-162.
- Szarkowska, Agnieszka, Krzysztof Krejtz, Łukasz Dutka, and Olga Pilipczuk. 2016. “Cognitive Load in Intralingual and Interlingual Respeaking – a Preliminary Study.” *Poznan Studies in Contemporary Linguistics* 52 (2): 209-233.
- Szarkowska, Agnieszka, and Lidia Bogucka. 2019. “Six-Second Rule Revisited: An Eye-Tracking Study on the Impact of Speech Rate and Language Proficiency on Subtitle Reading.” *Translation, Cognition & Behavior* 2 (1): 101-124.
- Talaván, Noa, and Pilar Rodríguez-Arancón. 2014. “The Use of Reverse Subtitling as an Online Collaborative Language Learning Tool.” *The Interpreter and Translator Trainer* 8 (1): 84-101.
- Walczak, Agnieszka, and Louise Fryer. 2018. “Vocal Delivery of Audio Description by Genre: Measuring Users’ Presence.” *Perspectives* 26 (1): 69–83.
<https://doi.org/10.1080/0907676X.2017.1298634>.
- Yan, Jackie Xiu, Jun Pan, and Honghua Wang. 2018. *Research on Translator and Interpreter Training: A Collective Volume of Bibliometric Reviews and Empirical Studies on Learners*. Springer.