Design and Statistics in Quantitative Translation (Process) Research

Balling, Laura Winther; Hvelplund, Kristian Tangsgaard

Published in:
Translation as a cognitive activity

Publication date:
2015

Citation for published version (APA):
Balling, L. W., & Hvelplund, K. T. (2015). Design and Statistics in Quantitative Translation (Process) Research. In F. Alves, A. Hurtado, I. Lacruz, & R. Muñoz Martín (Eds.), *Translation as a cognitive activity* (pp. 170–187). Bloomsbury Academic. Translation Spaces

Design and Statistics in Quantitative Translation (Process) Research

Laura Winther Balling
Department of International Business Communication, Copenhagen Business School
Dalgas Have 15
2000 Frederiksberg
DENMARK
lwb.ibc@cbs.dk

Kristian Tangsgaard Hvelplund
Department of English, Germanic and Romance Studies, University of Copenhagen
Njalsgade 128
2300 København S
DENMARK
bnm486@hum.ku.dk

Abstract

Traditionally, translation research has been qualitative, but quantitative research is becoming increasingly important, especially in translation process research but also in other areas of translation studies. This poses problems to many translation scholars since this way of thinking is unfamiliar. In this article, we attempt to mitigate these problems by outlining our approach to good quantitative research, all the way from research questions and study design to data preparation and statistics. We concentrate especially on the nature of the variables involved, both in terms of their scale and their role in the design; this has implications for both design and choice of statistics. Although we focus on quantitative research, we also argue that such research should be supplemented with qualitative analyses and considerations of the translation product.

Keywords

Quantitative research; experimental design; variables; statistics; translation processes

1. Introduction

At its core, translation process research (TPR) is about cognition: the cognitive processes associated with the complex task of translating. As such, the field should be understood as part of the cognitive sciences, and its methods are indeed inspired by the experimental approaches of these disciplines, especially psycholinguistics. However, this inspiration from the cognitive sciences could be applied more systematically in TPR and more widely in translation studies (TS). We often see students as well as more experienced researchers struggling with what is, to many translation scholars, a new way of thinking. We cannot in a single article do away with that struggle, but our hope is to contribute to making it easier, by highlighting some keys to good quantitative method, design and analysis. Although the examples in this article stem from TPR, many of the issues are also relevant to TS more broadly.

Two key distinctions in research design give rise to two crucial choices that must be addressed early in the design process. The most obvious one is the decision whether a certain research question is most appropriately addressed using qualitative or quantitative methods (or a combination). In TS generally, qualitative methods are dominant, while quantitative methods dominate in TPR. In this article, our main focus is on quantitative research, but we argue in section 5 that qualitative analyses of product and process are also important.

A second key distinction is between naturalistic and experimental approaches: in naturalistic studies, observations are made in the natural environment of whatever is observed, while experiments create an environment and/or task to elicit the relevant behaviour or other object of study. This distinction aligns to some extent with the distinction between qualitative and quantitative methods, with naturalistic designs often being analysed qualitatively and experimental designs often quantitatively. However, such an alignment is not logically necessary: quantitative research may fall anywhere on a continuum from strictly experimental to fully naturalistic, and it is useful in general to see these two paradigms as ends of a continuum, rather than a dichotomy. An experiment in the strict sense requires control of all variables except those under investigation, but in fields involving the study of human behaviour, this is often not possible and the term 'experiment' is therefore often used in the less strict sense of the task and/or environment being created for the purpose of the study, but with a, greater or smaller, naturalistic element. In that respect, TS tends to fall further towards the naturalistic end of the continuum than for instance the related discipline of psycholinguistics, even in the case of experimental research. This makes modern statistical methods that allow for statistical control in (quasi-)naturalistic set-ups all the more relevant in TS.

2. Quantitative research design in translation studies

Two practical aspects are crucial to the success of a quantitative study all the way from idea to result. Firstly, every part of the study – the whole research design – should be thought through before starting any data collection. Secondly, the design should be made as explicit as possible. Importantly, a thorough and explicit design substantially facilitates the choice and execution of statistical analyses. In this section, we present a number of steps that are important when designing a quantitative study. These must be addressed in a certain order, and there are indeed some steps which should precede others, but it is also important to see the design process as iterative: certain steps may need to be worked through several times.

2.1 Questions and hypotheses

The starting point for any study is what we could call an 'I-wonder'-question, which may be based on previous research or on personal observation and experiences. A simple example could be: '*I wonder whether student and professional translators differ in their translation processes*'. This is a

good question, and has indeed been addressed in many ways in the literature. However, this formulation does not (yet) constitute a research question. A research question needs to work as a foundation for the design, and for this to be possible, it needs to be more precise and explicit – in this case, the most obvious aspect to specify is what we mean by 'translation process', which part of the process we are interested in, and how we propose to measure it, but we also need to be clearer on what exactly is meant by students and professional translators.

One valid way of reformulating the original question is to stick with the question format but make it more explicit and ask: *Do translation students enrolled in an MA-programme differ from professional translators with more than ten years full-time experience in terms of their gaze time on the source text (ST)*? Now, the question is clear about the groups compared and about the process studied, and also indicates the measuring method, namely eye-tracking. The initial 'I-wonder'-question has been substantially narrowed down to focus on specific groups and a specific part of the process. This narrowing down is by no means a trivial operation, but it is necessarily very specific to specific questions, and therefore it cannot be in focus here; importantly, the researcher must be clear and explicit on why the groups and process chosen are important.

Alternatively, the original question may be reformulated as a hypothesis. Based on previous literature or experience, it may be reasonable to assume that there is a difference between students and professionals in terms of gaze time on ST, leading to a formulation as a hypothesis rather than a question: *There is a difference between translation students enrolled in an MA-programme and professional translators with more than ten years full-time experience in terms of their gaze time on the ST*. We might even be so sure about the difference that we formulate a directional hypothesis: *There is a difference between translation students enrolled in an MA-programme and professional translators with more than ten years full-time experience, such that the students gaze longer at the ST*. With the directional hypothesis, we predict that translations students will gaze longer at the ST whereas the non-directional hypothesis predicted that the difference could go either way. This distinction between non-directional and directional hypotheses is important in terms of clarifying one's design and expectations, but it also has ramifications for the statistical analysis: If the hypothesis is clearly directional, it may be valid to use so-called *one-tailed* tests, where the significance of the test is based on the assumption that the result can only go in *one* direction. However, one-tailed tests are often not accepted in high-quality publications and indeed one must be very sure of the direction of the hypothesis in order to choose a one-tailed test.

A further consideration in relation to research questions and hypotheses is whether they focus on differences between groups or relations between variables. Within this classification, there are three types of design:

1) In *independent groups designs*, groups of independent entities are compared in terms of the same parameter. Our example research question falls into this category: we are comparing two *different* groups of translators in terms of the same behaviour.

2) In *repeated measures designs*, several measures are examined for the *same* group of entities or individuals, for instance a group of translators working under different conditions or the same group of students doing the same or similar tests before or after a pedagogical intervention.

3) *Functional relations designs* consider relations instead of differences, for instance whether there is a relation between gaze time and typing speed. This is possible with variables that are numerical rather than categorical (see section 2.3.2).

Categorising a research design into one (or more) of these three groups helps clarify the design. The more technical reason for doing this is that it co-determines the choice of statistics, which we return to in section 4.

2.2 Populations and samples

The next important consideration has to do with the population we are investigating. The term population indicates a focus on people, and indeed in TPR, the population under investigation consists of people, typically translators. It is most often not possible to test the entire population that one is interested in: if we are interested in how professional translators perform a certain task, we cannot test the entire theoretical population of professional translators, past, present and future across the world. Instead, it is necessary to study an appropriate sample of the population of interest and use inferential statistics (see section 4.1) to evaluate whether any patterns observed for that sample are likely to generalise to the wider population.

A fundamental assumption of such inferential statistics is that the sample is *random*: that every individual in the population of interest has had an equal probability of being selected for the sample and that this selection of individuals has been random. In many fields, strictly random sampling is impossible, and we therefore have to strike a balance between randomness and possibility. However, it remains important that statistical tests do assume random sampling, and it may be necessary to reconsider what we define as the population of interest based on the sample we

can access: if we can only sample a subpopulation, then we need to acknowledge that we can only attempt to generalise to that subpopulation.

For example, we may be interested in the behaviour of professional translators, but we cannot sample randomly from such a broad population. Rather, translation scholars usually have access to translators with a restricted combination of working languages, based in a certain area, and should therefore consider the possibility that the conclusions drawn may not apply to all professional translators. Once we begin to see studies with different samples from different subpopulations showing similar patterns, we may also begin to conclude that those patterns apply generally.

A related question that often comes up is how large a sample is necessary for a given research purpose. There is no hard and fast rule for this but a number of relevant considerations. Firstly, of course, the bigger the sample the better: other things being equal, a bigger sample is likely to be more representative and gives superior statistical power. Superior statistical power means that smaller differences may be detected in the statistical tests, and we can be more certain of our estimates. However, power calculations are not straightforward, so we need rules of thumb. In psycholinguistics, experiments tend to include at least 20 participants, often more. In TS, even that number is often more than is feasible, but given the variability in translator behaviour we would recommend at least ten participants, and more in an independent groups design (though we have been guilty of going below that number ourselves, cf. Jensen *et al.* 2009; Balling *et al.* 2014).

The number of participants is not independent of the number of items, since the number of observations that the analysis is based on is the product of the number of participants and the number of items. If the number of different items – e.g. words, sentences or areas of interest in eye-tracking – is relatively large, it may be possible to make do with ten participants. By contrast, very small numbers of items or observations, in an extreme case for instance one manual translation production time and one post-editing production time per participant, would require a higher number of participants. This does not mean that observations for a large number of participants and a small number of items is equal to observations for a small number of participants and a large number of items: in the former case, we can generalise with more confidence to the population of participants, in our case translators, and in the latter case with more confidence to the population of items, e.g. texts or words.

Finally, the number of participants and the number of items necessarily depend on the complexity of the design. Generally, the more effects we are investigating, the more observations

are necessary, making it necessary to include more participants and more items to detect effects in a complex design. Here another rule of thumb (Harrell 2001, 61), specifically for regression modelling (see section 4.3), is that it is necessary to have at least 10-20 observations per parameter in the analysis, where parameters can roughly be described as effects of variables such as expertise or task, or interactions between them. All in all, the sample size must be based on a balance between the number of participants and items, the expected effect sizes and the complexity of the design (see Field *et al.* 2012, 273f).

2.3 Variables

Consideration of the variables of a study is arguably *the* central step of the design process, and certainly crucial for the choice of statistical approach. Two questions must be addressed: Firstly, what are the roles of different variables in the design? Secondly, what is the nature of each variable in the study?

2.3.1 Dependent, explanatory and control variables.

Starting with the first question, variables may be categorised as dependent, explanatory or control variables. The dependent and explanatory variables are reflected in the well-formulated research question or hypothesis: The dependent variable is the behaviour or other phenomenon measured during the study, while the explanatory variable is whatever variable potentially explains variation in the dependent variable. In other words, the dependent variable depends on the explanatory variable(s), while the explanatory variable(s) explain the dependent variable. Alternative terms are sometimes used: dependent variables may be called outcome or response variables, while explanatory variables are often called independent variables or, in the case of regression analyses, predictors. Importantly, the role of a given variable is not fixed: it may be a dependent variable in one analysis and an explanatory or control variable in another.

In the case of our example question from above – *do translation students enrolled in an MA-programme differ from professional translators with more than ten years full-time experience in terms of their gaze time on the ST?* – the dependent variable is the behaviour measured, namely gaze time on ST, while the underlying explanatory variable is expertise, operationalized as the difference between MA-students and professional translators with ten years of translation experience.

While the dependent and explanatory variables are in many ways the most important ones, they only become important if the third type of variable has been considered, namely control variables. Control variables are the variables that are necessary to control in order to know that whatever effect of the explanatory variable is observed is not an artefact of something else. If we want to find out whether gaze time on ST varies as a function of translator expertise, for instance, we need to make sure that expertise is not confounded, i.e. mixed up, with other variables. Ideally, we want to make sure that our group of students and our group of professionals are matched except that one group consists of students and one group of professionals. This is sometimes not possible, but the most important control variables at least need to be considered. In our example, it may be difficult to match a group of students and a group of professionals in terms of age (the professionals tend to be older) and L2 proficiency (the professionals will tend to be more proficient), but we should do what we can to take these issues into account. Other variables may be easier to control, for instance, it should not be too difficult to construct groups with comparable distributions of males and females. If we consider a comparison of texts rather than groups, the same logic applies: we need to make sure that whatever aspect of the text our hypothesis is about, the central explanatory variable of our design, is not confounded with other variables. This should be established through careful study of relevant previous literature, consideration of the set-up of the study, as well as personal experience.

The necessary control can be done by matching beforehand, so-called experimental control, or by measuring the variables and including them in the analysis, so-called statistical control. This difference is to some extent aligned with the distinction between experimental and naturalistic designs. In practice, some variables are easy to control experimentally and should be handled that way, while others are difficult or impossible to control beforehand and should be dealt with statistically. We return to this issue in section 4.3.

Control variables are almost by necessity plural. Dependent variables tend to be fixed in number, namely typically one per analysis, resulting in several analyses if a study yields several dependent variables. This leaves the explanatory variables, the number of which may vary: the simplest and most manageable designs manipulate only one explanatory variable, as in our expertise example. Even in more complicated setups, there is a limit to the number of variables that should be included if the analysis is to be practicable. One of the reasons for that is that the explanatory variables may enter into interactions, such that the effect of one explanatory variable differs

depending on another explanatory variable. With increasing numbers of explanatory variables, this line of thinking becomes increasingly unmanageable.

2.3.2 Numerical and categorical variables.

We need to consider not only the role of each variable in our design but also its scale. The most important distinction is between numerical and categorical variables: those phenomena of interest that may be meaningfully conceptualised as numbers vs. those that are inherently categories.

The simplest example of a categorical variable is one that has two unordered categories, such as sex/gender with the categories male and female. However, categorical variables also come in more complex flavours: there may be more than two categories, such as the class of nouns, or the categories may be ordered, as they are in our example where the group of professionals have more of something – namely expertise – than the group of students. A classic example of an ordered categorical variable with multiple levels is social class (used in sociolinguistics) with the levels for instance being working, lower middle, upper middle and upper class. These are clearly ordered, with the higher levels having more of something – socio-economic status – than the lower levels, but the intervals between them are not (necessarily) the same: the distance between working and lower middle class is not the same as the distance between lower middle and upper middle class. In terms of the statistical theory, the important consideration is whether the categories are ordered or not; in more practical terms, the statistical analysis becomes increasingly complicated with increasing number of category levels.

Turning to numerical variables, these are also of two kinds: discrete or continuous. Discrete variables are made up of integers (whole numbers), while continuous variables consist of real numbers, including an, in principle, infinite number of decimals. An example of a discrete numerical variable is counts of words in a corpus, where a word may occur exactly 3 times, never for instance 3.45 times (though means across several words would of course often be decimal numbers). By contrast, time measurements, such as the gaze time variable in our example, are continuous variables that may be subdivided depending on the measurement equipment, so we may talk about a measurement being 584 ms, but with coarser equipment, we would have a measurement of 0.6 seconds while the value with a finer measurement could be 584.4 ms.

Sometimes the classification of a variable is inherent – the gender of most people is inherently either female or male, and it is difficult to conceptualise differently. However, sometimes the scale of a variable can be a matter of conceptualisation, depending also on the available data. If

we look at the explanatory variable expertise, it is classified as a categorical variable, and a binary one where the ordering does not matter. However, it may also – just – be conceptualised as falling into any of the other three categories: We could compare three ordered categories, for instance beginning students, advanced students and professional translators, where we have a clear ordering, but the intervals between the levels are not the same. We could also conceptualise expertise in terms of year of study, comparing first, second and third year students. In that case, we assume that the interval between first and second year students is the same as the interval between second and third year students and therefore understand expertise as discrete numerical variable. Finally, we may see expertise as a continuous variable by measuring the time a participant has worked as a translator. Which conceptualisation to choose to some extent depends on the available data: if the sample we can get consists of fairly distinct groups, then we should see the variable as categorical. If, on the other hand, we can get participants with a range of different times of working as translators, this gives us more information and more powerful statistics and there is no reason to force the variable into categories.

In general, the most powerful statistics is available for continuous variables and the least powerful for unordered categories. This is partly because there is less information in the categorical variables and more in continuous ones – for instance, we know more about our participants' expertise/experience if we place them on a spectrum of 0 to 400 months of full-time experience than if we have them in two groups with some more or less arbitrary cut-off. This is illustrated in Figure 1, where the left panel shows the simple difference that can be uncovered by comparing two groups, and the right panel shows some different possible relations between the dependent and explanatory variables if we work with a continuous version of the latter.

INSERT FIG 1 HERE

Figure 1: Illustrations of possible outcomes of a hypothetical experiment with a categorical explanatory variable (left) and a continuous one (right).

3. Collecting and preparing data for statistical analysis

An important step in the analysis process is to make sure that the collected data are as representative as possible of the specific phenomenon under investigation. With respect to our professional translators and students, we want to be sure that the data are reasonably representative of professionals' and students' actual processes before we perform statistical analyses. Anomalies may be caused by instrument insensitivity to behavioural changes, instrument error, or human error in the recording process, and they are not uncommon.

A first thing to do is to consider the exclusion of outliers. Outliers are observations that are numerically distant from the rest of the sample. An outlier can be naturally distant from the sample. For instance, a fixation may be unusually long compared to the rest of the fixations in a recording, because the participant is in fact looking at the same locale for a very long time. Although natural, this observation may not reflect the processes that we are interested in.

Outliers may also be the result of measurement error. If we consider pupil dilation, which is sometimes used as an index of cognitive effort, it is very unlikely that a recorded pupil diameter of >9 mm reflects the actual pupil size of a participant since diameters greater than 9 mm are anatomically impossible. Similarly, extremely short fixations of <40 milliseconds are not likely to occur in reading during translation, and, while not anatomically impossible, they should still be considered as possible outliers since they deviate drastically from what we would normally expect.

Outlier exclusion is difficult since there will always be some element of subjective assessment involved. Outliers can be excluded manually but this is potentially very time-consuming, especially with large data sets. Outliers can also be removed 'automatically' by excluding those observations that deviate from the rest of the sample in a certain way, frequently those that are more than two or three *standard deviations* from the person's or item's mean. The standard deviation is a number which tells us how much observations are spread out from the mean of a sample, where a low number indicates that the observations cluster around the mean and a high number indicates the opposite.

In process-oriented studies, another important consideration in connection with data preparation has to do with the quality of eye-tracking data. Problems with eye-tracking data involve, for instance, abnormally short fixations, conspicuously few fixations in a recording or disagreement between registered fixation location and actual fixation location. Eye-tracking data quality is often not considered systematically and there is a strong risk that analyses are based on eye-tracking data that do not reflect actual cognitive processing. Thorough quality analysis is thus crucial (cf. Hvelplund 2014).

4. Statistical analysis

4.1 Description and inference

Once the data has been prepared, we turn to the statistical analysis with two distinct types being relevant for different purposes: descriptive statistics to describe the sample and inferential statistics to make inferences to a wider population.

An appropriate use of descriptive statistics is calculating and presenting for instance means or medians (the middle value of the dataset with half the observations being above and half below), standard deviations and absolute or relative frequencies as a description of the sample under investigation. For data addressing our research question from above, it would make good sense to consider various descriptive measures: for instance, what is the mean and median gaze time on ST for each of the groups? What is the standard deviation around the mean? If the standard deviation is large relative to the mean, it indicates that there is much variation within the group which is an interesting fact about the sample.

The answers to all these questions are informative and relevant and, along with good graphs of the raw data, give a necessary sense of the sample. However, the descriptive statistics do not provide more than a description of the sample. To make inferences beyond the sample – which is usually what researchers aim for – we need inferential statistics. Large differences in group means often co-occur with significance on inferential tests, and indeed inferential statistics use means to derive test statistics, but a range of other issues play in, including variance and sample size. We need inferential tests to draw any firm conclusions about how likely it is that patterns observed in the sample generalise to a wider population. Unfortunately, in TS, appropriate inferential statistics are often not used. Instead, general conclusions have been drawn based purely on descriptive statistics. Some of these conclusions are undoubtedly correct, but appropriate tests are necessary to give us an idea of how much confidence we can have that they generalise.

4.2 The mechanics of the statistical test

4.2.1 Hypothesis testing and significance

Statistical tests distinguish between systematic effects (across the population) and random variation (in the sample). They are based on two contrasting hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is that there is no difference, no relation or no effect on the dependent variable of whatever explanatory variable we are investigating; in our example from

above, the null hypothesis would be that there is *no difference* between a group of students and a group of professional translators in terms of gaze time on ST. The null hypothesis is a dummy hypothesis that is formulated because this is what the statistics can test, but most researchers are interested in the alternative hypothesis, in our case that there *is* a difference between students and professionals in terms of gaze time on ST. However, this alternative, scientific hypothesis cannot be directly tested, so it is done indirectly through a test of the null hypothesis. This holds irrespective of whether the design is based on a question or (directional) hypothesis.

In testing the null hypothesis, the statistical test provides the probability of observing more extreme values than those in our sample, if the null hypothesis is true. This is expressed in the p-value associated with the test statistic that we have calculated (for instance *t*-values for *t*-tests, χ^2 for χ^2-tests). A more informal understanding of the p-value is that it expresses the probability that an effect observed in the sample would arise if there was no difference in the population (i.e. if the null hypothesis was true), or, even more simply put, the probability that the effect observed *does not* generalise to the wider population. If that probability is low, as expressed in a p-value below a certain threshold, usually 0.05 or 5%, we assume that the null hypothesis is not true, leading us to reject the null hypothesis and accept the alternative hypothesis. The way this is usually phrased is in terms of the effect being significant, i.e. in terms of the question or hypothesis that is really of interest, but it is important to bear in mind that all the statistical test does is reject or not reject the null hypothesis. The alternative hypothesis cannot be directly accepted or rejected, and the null hypothesis cannot be accepted. This means that if we get a non-significant result, with a p-value above our threshold of 0.05, we can only say that we cannot reject the null hypothesis, but we cannot say that the null hypothesis is true. It could for instance be that with a larger sample or better measurements, the effect would be significant.

The threshold of 0.05 is an arbitrary one (but widely used across many scientific fields) which essentially means that we accept that 5% of the times we interpret a result as significant, supporting the alternative hypothesis, we are in fact wrong, and the null hypothesis is true. This kind of error is known as type I error. Importantly, if we run multiple significance tests on the same data, the risk of a type I error is inflated; if for instance we run comparisons of three groups using three *t*-tests, the type I error rate becomes almost 15%. This problem needs to be addressed, for instance using the Bonferroni correction (described in most statistics books), or the less conservative sequential Bonferroni correction (Holm 1979; a more approachable presentation is Rice 1989).

4.2.2 Assumptions

Before running a statistical test, there are certain assumptions that should be considered. One very important assumption is that the observations in the sample are independent of each other (except for pairs of observations in repeated measures designs, for instance before- and after-tests for the same person). If the observations are not independent, for instance with multiple observations from the same person or multiple responses to the same item, the dependencies must be either modelled or eliminated; this will be further discussed in section 4.3.

Another assumption that holds for the most powerful, so-called parametric statistical tests is that the sample comes from a distribution of a certain type: the classical example here is that *t*-tests – which are used with continuous dependent variables such as response times – are based on the assumption that the data come from a normal distribution. We cannot go into details of individual distributions here, but refer the reader to any of the statistics books cited in section 4.3.

Importantly, if the assumptions of parametric tests are not met by the data in the sample, the test results may not be reliable – or informally speaking: garbage in, garbage out. This means that if we want to run a test (e.g. a *t*-test) that assumes that the data come from a normal distribution, it is necessary to assess whether this is the case. There are specific tests of normality, e.g. the Kolmogorov-Smirnov, but these are problematic because for very small samples, where normality is very important, the normality tests almost always conclude that the data are normally distributed, while for large samples, where deviation from normality is less of a problem, the normality tests are hypersensitive and almost always reject that the distribution is normal. The normality test may be supplemented with a plot of the sample distribution, but this requires experience to judge, so for small samples it may be the safer choice to go – if possible – for non-parametric tests that do not assume specific underlying distributions, though this may lead to a loss of statistical power.

In some cases, it may also be possible to transform the relevant variable so that the normality assumptions of a test are met, using for instance a logarithmic transformation (see e.g. Bordens and Abbott 2005, 418-419). As long as the transformation is a systematic and commonly used one, this is completely accepted practice. For the logarithmic transformation, the result of the transformation is that large differences at the higher end of the scale, e.g. differences between response times of 1,000 and 10,000 milliseconds, become equal to smaller differences at the lower end of the scale, e.g. between 100 and 1000 milliseconds. This is not trivial to the interpretation of the variable, but may in fact reflect cognitive reality (e.g. Smith and Levy 2013).

4.3 Statistics for naturalistic and experimental research

4.3.1 Factorial and regression designs

Several of the issues discussed in sections 1 and 2 are crucial when it comes to the choice and execution of statistical tests. Most prominent among these are the balance between naturalistic and experimental aspects of the design and the categorisation of variables in terms of their role in the design (dependent, explanatory and control variables) and their scale (numerical or categorical).

The difference between experimental and naturalistic research is relevant because the two types may lead to fundamentally different approaches to design and statistics. So-called factorial designs and analyses may be useful for experimental research, while regression designs and analyses are more appropriate for naturalistic or quasi-experimental research.

In a factorial design, items or participants are chosen so that the explanatory variable(s) may be conceived as factor(s), i.e. categorical variables, while all relevant control variables are matched between the levels of the factor. Our example research question from above could be the foundation of a factorial design, if we were to work with two closely matched groups of participants, and if all relevant control variables were controlled beforehand. For factorial designs, the appropriate statistics are factorial statistics: *t*-tests when comparing two groups, analyses of variance (ANOVAs) when working with more than two groups or more than one explanatory variable.

However, very often in TS and other fields strict experimental control is not feasible; if we want to work with authentic texts and (quasi-)naturalistic setups – i.e. a more naturalistic approach – we cannot control all relevant variables beforehand. We need another approach to control, namely statistical control. Statistical control may be implemented by measuring a range of relevant variables and including them in the statistical analysis. In that case, we need some kind of regression model, which may include numerical variables only, or a mixture of numerical and categorical variables. Importantly, the statistical approach to control also allows us to investigate more variables, instead of controlling them beforehand, providing more information to the researcher, which is particularly relevant in a relatively new field like TPR. The fact that regression designs may include both categorical and numerical variables leads us to a second strong argument in favour of using regression models in quantitative translation research: it makes it possible to accept the nature of the variable, instead of forcing variables into categories, as is necessary for factorial designs and analyses. This means, as discussed above, that much more information is obtained and may aid our understanding. As seen in Figure 1, different shapes of an effect may

become apparent with a numerical explanatory variable, but not with a categorical one. Numerical explanatory variables are also associated with more statistical power, making it easier to detect effects.

The factorial approach has been dominant for decades, for the technical reason that *t*-tests and ANOVAs can, with relative ease, be done by hand. However, with the computing power available today, this is no longer necessary, and regression approaches are gaining a footing in psycholinguistics. We think that translation research should take inspiration from this development, and use possibilities for statistical control and more naturalistic setups and the increased information available in regression approaches.

4.3.2 Mixed models

More specifically, we recommend a type of regression model called mixed models, sometimes referred to as linear mixed-effects regression (LMER). The key characteristic of this kind of analysis is that it includes both so-called fixed and random effects, hence the name mixed-effects. The fixed effects are typically the explanatory and control variables that we investigate in the analysis, while the random effects represent those variables that arise as a consequence of our sample being (in principle) random. For instance, the effect of our explanatory variable group (students vs. professionals) would be a fixed effect, while the differences between the individual participants would be a random effect. The key point is that while we are interested in certain characteristics of the participants, here their professional status, we are not interested in the individual participants themselves, since the participants are – in principle – randomly chosen. Similarly, we could imagine a situation where we are interested in certain characteristics of our items, e.g. words or sentences, but not in the words or sentences themselves, making item a random effect.

Including random effects of participant and item means that the model takes the random variation between individual participants and between individual items into account. This makes intuitive sense in the cognitive sciences where there is typically quite large variation in people's performance. It also has a very important technical reason, namely that it is a way of dealing with the fact that observations are not typically independent of each other, because we typically have multiple observations for each participant and/or each item. Since the statistical model assumes independence between observations, we need to deal with cases of non-independence, and including random effects is a good way of doing that. The alternative, more traditional way of achieving

independence is to run statistical analyses on participant and/or item means, but this has a number of disadvantages: means are problematic if we do not have the same number of observations contributing to each mean; taking means removes variation in the observations which may be meaningful; and we typically need two analyses, one on item means and one on participant means, whereas a single mixed model suffices because both kinds of dependencies can be modelled at the same time.

These advantages of mixed models do not come without drawbacks, most seriously that these are complex analyses which take time to master and may pose interpretational challenges, even to experienced users. Also, the use of mixed models, like all but the most basic statistics, requires a certain level of familiarity with statistical software, such as R, SPSS or SAS. We tend to use and recommend R (www.r-project.org) because it is a flexible and powerful tool that is free and open-source. In order to learn to use R and mixed models, we recommend Gries (2013) and Baayen (2008); other books introducing R and statistics for the language sciences are Johnson (2008) and Vasishth and Broe (2011). More information specifically about mixed models may be found in Baayen et al. (2008) and, more informally, in Balling (2008). More domain-general introductions also abound, e.g. Field, Miles and Field (2012).

Although mixed models are more complicated than some of the alternative approaches, the difficulty of learning to use them is not fundamentally different from the difficulties associated with learning any kind of just moderately advanced statistical technique. Crucially, the key advantages – statistical control, investigation of multiple variables on different scales and modelling non-independence and participant-specific effects – far outweigh the steep learning curve.

5. Supplementing quantitative process data

Although this article is primarily concerned with quantitative research, a main point is in fact that such research should be supplemented by other approaches in at least two ways. Firstly, on the most general level, it is often useful to let quantitative and qualitative approaches complement each other. In the case of our example hypothesis, we would interpret quantitative differences in eye-movements as evidence of differences between the groups, with longer fixations indicating increased difficulty. However, the quantitative data do not tell us *which* difficulties arise and *why* and should therefore be supplemented with qualitative analyses, which may be based on the translation product, interviews, questionnaires or cued retrospection.

Secondly, and more specifically for TPR, the translation product is often automatically assumed to be acceptable, but we want to argue that the quantitative analysis of the process data should be supplemented with considerations of the outcome of that process, the translation product. Although we may reasonably assume that professional translators produce acceptable translations, it is not meaningful to assume that there are no variations in quality, which is essentially what is done when the product is disregarded. This disregard for variations in the product and its quality risks undermining our attempt at a general description of the translation process.

There are at least two main reasons that product and quality have not received much focus in process studies. Firstly, product analysis is often very labour intensive; secondly, and more seriously, the concept of translation quality is ill-defined. To take both process and product into account would demand a more well-defined idea of what translation quality is and how to measure it.

One approach to defining translation quality is to have panels assess the quality of translated texts. Panels in this context could consist of professional translators, teachers of translation and language, or end-users of the translated texts. However, this is potentially very time-consuming and expensive. Another approach to quality assessment comes from the field of machine translation. Carl and Buch-Kromann (2010) demonstrate how human translations can be assessed using BLEU (Bilingual Evaluation Understudy, Papineni *et al*. 2002), a method for automatically evaluating the quality of machine translation using probabilistic estimation. Carl and Buch-Kromann found that the automated BLEU score was useful for assessing the quality of human translated texts. This tool provides a cost-efficient and quickly implementable way of assessing quality and could therefore be useful, but more research is needed on this and other ways of assessing translation quality.

6. Conclusions

Although this paper outlines many important aspects of quantitative research, the crucial message is a very simple one, namely that quantitative studies should be based on completely explicit research designs that have been thought through at the beginning of the investigation. If our designs are explicit and systematic in terms of research questions and hypotheses, sample and population and the role and scale of variables, the statistical analyses become – at least comparatively – simple, especially if the design process has also included considerations of what constitutes confirmation or rejection of hypotheses and of which other patterns of results could be interesting and interpretable. Our approach does of course – and luckily – not eliminate the challenges and surprises of doing

quantitative research, but hopefully helps us derive more and more solid information from our studies, increasing our understanding of the fascinating process of translation.

7. References

Baayen, R.Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. "Mixed-effects modeling with crossed random effects for subjects and items." *Journal of Memory and Language* 59: 390-412.

Balling, Laura W. 2008. "A brief introduction to regression designs and mixed-effects modelling by a recent convert." *Copenhagen Studies in Language* 36: 175-192.

Balling, Laura W., Kristian T. Hvelplund, and Annette C. Sjørup. 2014. "Evidence of parallel processing during translation." *Meta* 59: 234-259.

Bordens, Kenneth S., and Bruce B. Abbott. 2005. *Research and Design Methods. A Process Approach*. 6th ed. Boston: McGraw Hill.

Carl, Michael, and Matthias Buch-Kromann. 2010. "Correlating Translation Product and Translation Process Data of Professional and Student Translators." In *Proceedings of EAMT*, Saint-Raphaël, France.

Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. London: Sage.

Gries, Stefan T. 2013. *Statistics for Linguistics with R. A Practical Introduction*. 2nd edition. Berlin: Mouton de Gruyter.

Harrell, Frank E. Jr. 2001. *Regression Modelling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Holm, S. 1979. "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* 6: 65-70.

Hvelplund, Kristian T. 2014. "Eye tracking and the translation process: reflections on the analysis and interpretation of eye tracking data." In *MonTI Special 1: Minding Translation, Con la traducción en mente*, ed. by Ricardo Muñoz Martín. 201-223.

Jensen, Kristian T. Hvelplund, Annette C. Sjørup, and Laura W. Balling. 2009. "Effects of L1 syntax on L2 translation." *Copenhagen Studies in Language* 38: 319-336.

Johnson, Keith. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “BLEU: A method for automatic evaluation of machine translation.” In *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*: 311-318.

Rice, William R. 1989. Analyzing Tables of Statistical Tests. *Evolution* 43: 223-225.

Smith, Nathaniel J., and Roger Levy. 2013. “The effect of word predictability on reading time is logarithmic.” *Cognition* 128: 302–319.

Vasishth, Shravan, and Michael Broe. 2011. *The Foundations of Statistics: A Simulation-based Approach*. Heidelberg: Springer.