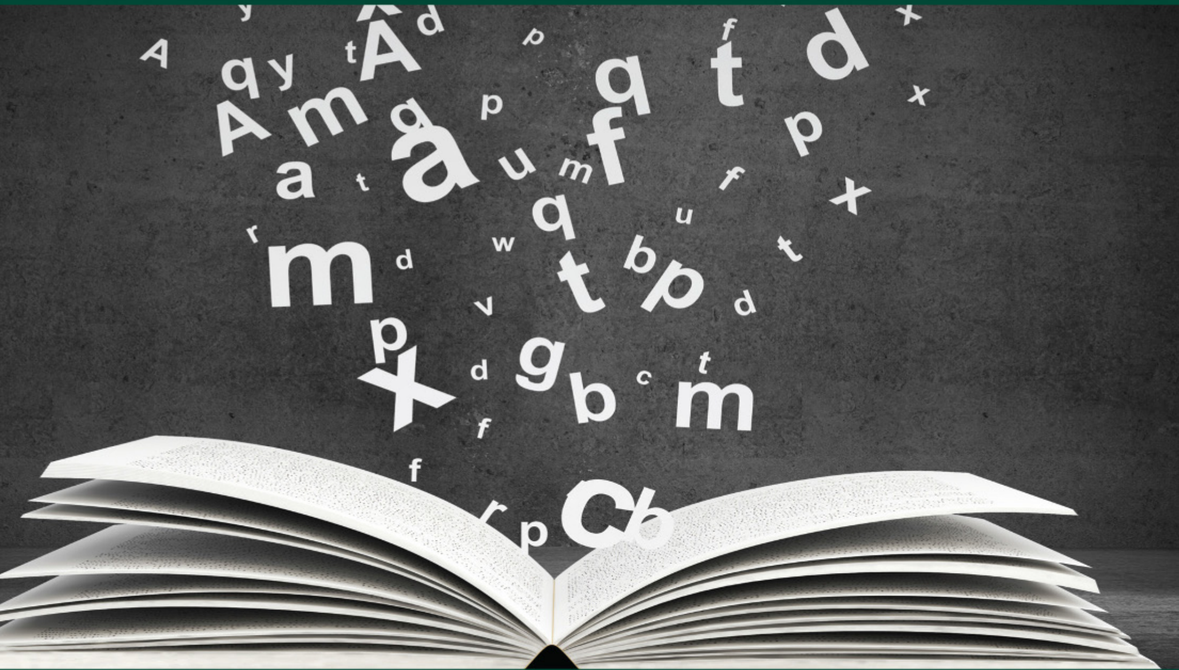
COGNITIVE SCIENCE SERIES

Introduction to Experimental Linguistics

Christelle Gillioz
Sandrine Zufferey



iSTE

WILEY

Introduction to Experimental Lingustics

Introduction to Experimental Linguistics

Christelle Gillioz
Sandrine Zufferey

iSTE

WILEY

First published 2020 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2020943938

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-418-6

Contents

Preface

This book aims to present the theoretical and methodological principles of experimental linguistics in an accessible manner. It intends to offer an overall vision of the field, so as to help the non-initiated audience to become familiar with the necessary concepts for carrying out linguistic experiments. The elements discussed in this book can particularly serve as a basis for a critical understanding of the results published in the scientific literature and as a starting point for carrying out experiments.

Since the field of experimental linguistics is rich and varied, both in terms of the phenomena studied and of the methods employed, it is impossible to offer an exhaustive presentation. The choice of aspects introduced in this book aims to provide an overview of the different possibilities available to those wishing to carry out an experimental study about language. For every aspect developed in the chapters of the book, there exist specific works which, due to their complexity and the prerequisites they demand, are often reserved for an expert audience. This is why we have deliberately chosen to select the information we deem essential for building a knowledge base that will later enable readers to explore the scientific literature and other works on this topic. Therefore, the emphasis will be placed on understanding the scientific approach and the methodological principles underlying the construction of experiments, and on analyzing the data which results from these experiments. In regards to research methods, we chose to make a presentation of the most accessible methods for linguists. In order to illustrate the many possibilities for applying such methods, we have provided examples drawn from different fields in linguistics. Finally, a list of more specific resources and available tools is provided at the end of each chapter, in order to encourage the interested reader to deepen and put into practice the knowledge acquired in this book.

This book begins with an introductory chapter, offering a general overview of the principles underlying experimental methodology, as well as the key concepts which will be developed in the rest of the chapters.

Chapter 2 goes through the various points the researcher should comply with in order to conduct a valid and reliable experiment, thus making it possible to infer solid conclusions. First, we will define the concepts of validity and reliability and then discuss the notion of variables, as well as present different options for measuring such variables. We will pay special attention to the stages involved in the transformation of the research question into an experimentally testable hypothesis.

Chapters 3–5 are dedicated to the different methods used for studying language production (Chapter 3) and language comprehension, focusing not only on the results of the comprehension process (Chapter 4), but also on the process itself (Chapter 5).

Chapter 6 presents the main practical aspects associated with the construction of an experiment, such as the various possibilities offered by different types of experimental designs, the criteria for choosing the experimental material, the stages involved in an experiment, the aspects related to data collection, as well as the ethical principles that should be observed while carrying out research with human participants.

Finally, Chapter 7 offers an introduction to the analysis of quantitative data, aiming to summarize the key elements for understanding descriptive and inferential statistics, as found in the scientific literature devoted to experimental linguistics. This chapter will also emphasize the peculiarities of the data acquired through linguistic experiments, namely the interdependence of observations. Then, we will introduce mixed linear models that can be used to analyze such types of data.

Christelle GILLIOZ
Sandrine ZUFFEREY
August 2020

1

Experimental Linguistics: General Principles

We start this chapter by outlining the foundations of the experimental methodology and its main features. Then, we discuss the advantages and disadvantages of this type of methodology, as well as the main arguments in favor of its use in the field of linguistics. Last, we present a series of resources offering access to research in experimental linguistics.

1.1. The scientific process

The experimental methodology in linguistics is part of a scientific approach for studying language. It aims to observe language facts from an *objective* and *quantitative* point of view. The general idea behind this approach is that it is impossible to rely on one's own intuitions in order to understand the world. Quite the contrary, it is necessary to observe objective data reflecting reality. For example, by simply observing the world around us, and relying solely on our own intuition, we might believe that the Earth is flat. This is why the scientific approach, used in fields such as psychology or physics, is based on specific principles and stages, instead of relying on the intuition of scientists. Let us briefly go through these stages:

The first stage in the scientific process involves the observation of concrete phenomena and the subsequent generalization of observations, in order to build a scientific fact: *a fact which does not depend on a specific place, time, object or person*. At this first stage, it is also possible to trace

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

certain regularities concerning the emergence of a phenomenon, and to try to define the conditions in which such phenomenon generally appears. So, let us illustrate this process by reviewing the stages involved in the discovery of gravitation. This finding is usually attributed to Isaac Newton, who is said to have had a revelation after seeing several apples fall from a tree. As he watched the apples fall, Newton wondered why the apples always fell in a perpendicular direction from the apple tree to the ground, never to the side or upwards.

During the second stage, all of the scientific facts concerning the same phenomenon may prompt the development of a law or theory aimed at explaining such facts. A *theory* synthesizes knowledge about a phenomenon at a given moment and is therefore provisional, insofar as it can evolve according to new knowledge. We should make it clear that the notion of theory in science is rather distant from the meaning of the word *theory* as we use it in everyday language. While this word can be used to refer to personal ideas or reasoning mechanisms, its use in the scientific field only applies to coherent and well-established principles or explanations. Going back to our example, in Newton's time, two models coexisted for describing the movement of bodies: one followed Galileo's law and was devoted to terrestrial bodies, whereas the other was oriented by Kepler's law and made reference to celestial bodies. On the basis of this knowledge and his own observations, Newton suggested the existence of a force which made objects attract one another and which could explain the movements of both celestial and terrestrial bodies.

At the third stage, a theory is capable of predicting the emergence of observable facts, or to put it differently, to formulate precise *hypotheses* which can be put to the test. In order to test these hypotheses, it is necessary to collect a large amount of data and check whether they support the initial theory. In this way, it is possible to know to what extent we can rely on our theory. The more the predictions made on the basis of the theory are fulfilled, that is, the more the data collected corresponds to what might be expected according to the theory, the higher the confidence level will be. Otherwise, if the predictions did not come true, the theory should be put into question and re-examined. Newton's law of universal gravitation has made it possible to predict and explain the movement of the tides thanks to the moon's gravitational pull on the Earth, the elliptical movement of celestial bodies or the equatorial bulge.

In summary, the scientific approach is a circular and dynamic process, originating in the reality of the facts, abstracting itself from them in an attempt to explain them, and then approaching them again to check the validity of the explanation.

1.1.1. *Qualitative and quantitative approaches*

It is possible to investigate a research question in different ways and from different perspectives. Let us imagine that you wish to study second language acquisition within the context of linguistic immersion. The first way of doing this could be to contact students attending your university for a language stay and to interview them. These interviews can later be viewed to analyze the opinions of students regarding their experience during their stay, their feelings on its advantages and disadvantages, or their opinion on the impact of such a stay on their linguistic competences. By doing this, you would be carrying out what is called *qualitative* research.

The qualitative approach helps us to explore and understand a phenomenon by studying it in detail and trying to take hold of it in a holistic manner, based on the meanings that people assign to the phenomenon. This type of research takes a long time when conducting interviews and interpreting the results; hence, only a small number of individuals can be questioned. Due to this characteristic, the results of a qualitative study are strongly anchored to the context in which the study was carried out, and cannot be generalized to other people or to other contexts. This is not a problem insofar, as qualitative studies do not aim to make such a generalization. The subjectivity of the individuals involved in the study is acknowledged as an integral part of qualitative research. This methodology is built on the principles of a constructivist vision of knowledge, according to which there is not only one, but many realities construed by people's interpretations and the meanings they attribute to events or things, on the basis of their own experience.

When reading this first proposal for investigating second language acquisition within a context of linguistic immersion, you might think that although it may be interesting to know learners' opinions about their experience in a language stay, you also desire to know more about the benefits of such a stay on the evolution of their linguistic competences. The conclusions drawn based on the opinions of a few interviewees may not reflect the reality of all learners. It is possible that the interviewees could subjectively overestimate or underestimate the evolution of their skills, or that these particular cases do not mirror the typical experience learners have during a language stay. One possibility, to obtain more objective data on the advantages of a language stay for improving linguistic competences, could be to take into account the experience of more people and to measure their linguistic competences at the start and end of the stay, for example, with an assessment test. By comparing the results before and after the stay with the help of a *statistical test*, you could determine whether the students' linguistic skills have evolved and in what aspect. If you chose this second option, your research would follow a *quantitative* methodology, in the sense that your conclusions would be drawn from the analysis of numerical data pertaining to a large number of people, and objectively assessed through a test. Your results would depend little on the respondents, their subjective perceptions or your interpretation of their declarations. If learners have really benefited from their language stay, this should be reflected in their results to the test, probably higher at the end than at the beginning of the stay, and this is what you would measure directly.

This example illustrates to what extent quantitative research differs from qualitative research, in that it aims to observe quantifiable elements and to *measure* a phenomenon. The techniques used for measuring a phenomenon can be extremely varied, depending on how the phenomenon is defined. Going back to our previous example, it is possible to measure language proficiency using a general language test (such as the placement tests used in language schools). Another way of doing this would be to count the number of mistakes students make in a grammar test or to measure the size of their second language lexicon. Choosing the proper measures for undertaking research is a big question in itself. We will return to this in Chapter 2, where we will discuss the different stages of choosing the measures involved in an experiment.

Quantitative research also differs from qualitative research in terms of the type of reasoning on which it is based. We have seen that in qualitative research, we draw upon data in order to outline a structure. In this case, data works as a source of interpretations and explanations upon which hypotheses will be formulated. This type of reasoning, starting from data and leading towards a theory, is called *inductive* reasoning. On the contrary, quantitative research follows *deductive* reasoning: it draws on theory in order to formulate hypotheses which will later be verified by data acquired in the field. When choosing a deductive approach, it is necessary to build a preliminary hypothesis, on which the research will be based and that will guide the researchers' methodological choices.

Going back to the example of learners within an immersion context, there are a large number of hypotheses that could be formulated by using the link between language stay and language proficiency. The first hypothesis could be that a language stay improves second language skills. A second hypothesis, similar to the first, but involving a different research methodology, could be that people who have spent time on a language stay have acquired better skills than those who have not. In order to verify the second hypothesis, we would have to test two groups of learners who may or may not have benefited from linguistic immersion, instead of one group of students before and after the stay. A third hypothesis could focus on one specific aspect of language proficiency, such as pronunciation in a foreign language (accent). We might imagine that the learners who have spent some time on a language stay may have a better pronunciation (an accent closer to that of the native speakers), than those who have not. In order to test this third hypothesis, two groups of students would be required, but this time they would be assessed on their pronunciation.

Even if they differ in their formulation and in the type of elements they have put to the test, the hypotheses mentioned above share a common feature, which is that they all postulate a relationship between what we call *variables*. In all the hypotheses, the first variable corresponds to linguistic immersion. In the first and second hypotheses, the second variable is the proficiency level in the second language. In the third hypothesis, the second variable corresponds to a weaker non-native accent. We will discuss the notion of variables in further detail in Chapter 2. For the time being, it is important to understand that a variable is something that varies, and can take different values. For example, a variable can be the age of participants in a study, which would result in a broad number of values. A second variable could be the fact of wearing glasses, or eye color, etc. These variables adopt fewer values: either yes or no for wearing glasses, and blue, brown, green or other for eye color.

Let us now take the example of a variable studied in language science: bilingualism. At first glance, this variable may seem to only adopt two values: either bilingual or monolingual. However, things get more complicated when we have *to define* what we mean by *bilingual*. For example, we may decide that anyone having knowledge of a second language is bilingual. In that case, there would be great heterogeneity within the bilingual group, containing people who can only speak or understand a second language superficially, and people capable of perfectly mastering both languages. A corollary of such a definition would be that very few people would belong to the monolingual group, since many people are familiar with one or more languages, apart from their mother tongue. On the other extreme, we could consider belonging to the bilingual group as only those with a perfect command of their second language. In this case, the bilingual group as would be more homogeneous, in the sense that all those belonging to it would have similar competences in their second language. But this definition raises additional questions: what do we mean by perfect command and how can command be measured? This example illustrates the need to clearly and precisely define the variables investigated in a research process. This definition procedure is called the *operationalization* of a research question. It represents a crucial phase in quantitative research, and we will discuss it in depth in Chapter 2.

To summarize, quantitative research aims to investigate the relationship between two or more variables. To do this, it starts from a hypothesis and defines the measures used for studying the chosen variables. Then, it relies on digital data collected from a large number of people and analyzes such data using statistical tests, in order to generalize the results.

1.1.2. *Observational research and experimental research*

Quantitative approaches in linguistics make an important difference between observational research and experimental research. The first example of a research tool, the questionnaire, is frequently used in linguistics to collect data in a quantitative manner. A questionnaire is a set of questions aimed at collecting different types of information about speakers, such as personal characteristics, their use of certain words or linguistic structures, or their point of view about certain linguistic phenomena. Let us now imagine that you wish to know whether there is a difference in the way that French speakers from France, Belgium and Switzerland refer to a *yogurt.* As Avanzi (2019) did, you could directly ask a large number of French, Belgian and Swiss people to tell you which of the two possible names, *yaourt* or *yoghourt*, they use on a daily basis. By counting the responses of more than 7,000 people, Avanzi showed that the form *yaourt* is mainly used in France, whereas it is never used in Switzerland, where *yoghourt* is the only form in use. In Belgium, the choice of *yaourt* and *yoghourt* varies from region to region.

In a slightly different way, instead of relying on the answers of people in a questionnaire, you could use linguistic data retrieved from natural productions and carry out a *corpus study*. In such studies, linguistic productions in the form of texts, audio or video recordings are used with the aim of counting the number of word occurrences, a grammatical form or any linguistic characteristic. In order to research the uses of *yaourt* or *yoghourt* in France, Belgium and Switzerland, first it would be necessary to select corpora comprising linguistic productions collected from these different regions. This data could come from French, Belgian and Swiss newspapers, for example. The number of occurrences of each form could be counted in each corpus and then compared, in order to reveal differences in the use of these forms from country to country.

Another way of studying quantitative data is to examine the link between two variables. Let us imagine that you wish to study the relation between learners' age and their ability to acquire a second language. Extensive research has already been devoted to this topic and suggests that the older people are when learning a second language, the more difficult it is for them to reach a high level of proficiency (see DeKeyser and Larson-Hall (2005) for a review). In order to confirm (or refute) this hypothesis, you could test a large number of people who start learning a language at different ages and measure their language proficiency after a certain period of time. In this example, the first variable, the age when learning begins, is a quantitative variable. Likewise, the second variable, language proficiency, can be measured quantitatively using a language test. Using an appropriate statistical test, it is possible to show the existence of a link between these two variables. This type of procedure is called *correlational* research and unveils the degree of dependence between two variables, which is called correlation. In the case of our example, if age plays a role in second language acquisition, the correlation obtained by our test would show that the older a person is when the process of learning a language begins, the lower their mastery of the language will be after a certain learning period.

The various studies described above correspond to research based on data observation. This type of research is generally used when, for practical or ethical reasons, it is necessary to observe variables from the outside. In this type of research, researchers do not interfere with the object of study, but observe the relationship between two variables at a given moment. As a consequence, the results of an observational study must be kept at a descriptive level, since it is not possible to infer a causal relation between two variables. In our example of a correlational study, the age when learning begins is related to language proficiency, but it is not possible to state that an increase in age is the *cause* for the decrease in language proficiency. It might be possible that other variables not considered in our research can also explain the relationship between the variables examined. We could imagine, for example, that the context in which second language acquisition takes place is not the same depending on the age when the learning process begins. It is likely that when young children learn a second language, this takes place within a family setting, where parents may speak different languages or a different language from that of the external environment. When older people start learning a language, it is probable that they grew up in a monolingual linguistic environment and later discovered a second language at school, or when moving to another country, for example. The type of linguistic exchanges may also differ depending on age, as well as the motivation to learn, cognitive skills or many other variables. These external variables that are left aside during research are called *confounding variables* and are related to the two variables examined, age and language proficiency. It could be, that language learning conditions rather than age itself can account for the differences in language levels. Since it is impossible to distinguish the variables examined, from confounding variables, research based on the observation of data should not draw a conclusion from a causal relation between two variables.

In order to determine a causal relation between two variables, it is necessary to exclude any confounding variable. By using experimental methodology, the variables of interest can be *manipulated* to determine what effect a variable has on another variable, regardless of other possibly interfering variables. In other words, rather than observing natural data, the experimental methodology defines the conditions under which a phenomenon could be observed and then sets up an experiment in which these conditions can be manipulated, in order to measure their influence on the phenomenon under investigation. In the rest of this chapter, we will describe in more detail the various characteristics of experimental research.

1.2. Characteristics of experimental research

In this section, we will first stress the fact that experimental research must be based on a research question that makes it possible to formulate precise hypotheses. We will then see that in order to empirically assess a hypothesis, an experimental study must manipulate variables of interest while controlling other variables, which may influence the outcome of the experiment. Finally, we will discuss some methodological aspects of data collection, so that they can be analyzed through the use of statistics. These points will be elaborated in detail in the chapters dedicated to these different aspects.

1.2.1. *Research questions and hypotheses*

We have already emphasized that experimental research is part of a scientific process. It builds on existing knowledge in a research field and aims to increase such knowledge by studying a *research question* generated on the basis of an existing theory. A scientific research question identifies the potential cause for a phenomenon and postulates a cause to effect relation between the cause and phenomenon. For example, the question “how do we understand a text?” is not a research question, as it is too vague. Such a question corresponds to a general research topic, from which many research questions can emanate. On the other hand, a question such as “what is the role of memory in readers’ comprehension of a text?” is a research question that can be investigated empirically. This question identifies a cause – memory – and a consequence – text comprehension –, and establishes a relation between the two.

Once the research question has been defined, it is necessary to transform it into a *research hypothesis*, which corresponds to an empirically testable statement. In other words, the hypothesis must be confirmed or rejected on the basis of objective data. In order to do this, the research hypothesis must be operationalized, that is, it is necessary to specify which variables will be examined and how these variables will be measured, in order to collect relevant data for the experiment.

If we go back to our example above, *memory* is still a vague concept. As a matter of fact, a distinction is generally made between long-term memory, short-term memory and working memory. Working memory is a system that simultaneously stores and processes verbal elements (verbal working memory) or visual elements (visuospatial working memory). It is typically the verbal working memory that we use for reading, for deciphering and for putting together the words in a sentence. The operational hypothesis should therefore define what type of memory will be the object of study, verbal working memory, for example.

In the same way, the operational hypothesis should explain the way in which *reading comprehension* will be measured. Reading comprehension involves many steps, from deciphering words to relating these words in a sentence, and then to a text. Therefore, it is impossible to measure *reading comprehension* in only one way or with one type of experiment. We need to narrow down this notion to a more precise variable, corresponding to a process involved in reading comprehension that can be measured. For example, this could be the elements included in the readers' representation of the text and stored in memory once the reading has finished. One way to assess comprehension would be to ask questions about the text at the end of reading and count the number of correct answers.

Let us look at a few more examples to understand what a research hypothesis is:

(1) Bilinguals have different cognitive abilities from monolinguals.

(2) Reading and understanding a text is difficult for children.

The above-mentioned hypotheses cannot be the basis of experimental research since they do not meet the criteria listed above. Their terms are too vague, they specify neither the cause nor the effect, and do not specify any measure to rely on so as to draw conclusions.

In order to be tested empirically, these hypotheses could be transformed into (3) and (4):

(3) Bilinguals perform better than monolinguals at a cognitive flexibility task.

(4) When reading a text, 10–12-year-old children draw fewer inferences than 14–16-year-old teenagers.

In these two examples we see that the vague terms used in (1) and (2) have been transformed into accurate terms in (3) and (4). *Cognitive skills* became *performance during a cognitive flexibility task,* and *understanding a* *text* became *drawing inferences*. By doing this, measures for quantifying the variables were defined. In addition, (4) specifies which groups will be included and compared in the study. Finally, both (3) and (4) indicate a clear relationship between variables.

In summary, a research hypothesis is based on existing knowledge in order to establish a relationship between two or more variables. It must also be operationalized, that is, clearly defining the measures that will be used for quantifying the variables being examined to verify the hypothesis.

The construction of a good research hypothesis is the result of different stages, among which the most important are conceptualizing the hypothesis, on the basis of knowledge acquired in the field, and then operationalizing the hypothesis. We will discuss the specific stages for conceptualizing a hypothesis in Chapter 6, which is devoted to the practical aspects of an experiment. We will discuss the stages involved in the operationalization of a hypothesis in Chapter 2.

1.2.2. *Manipulation of variables*

Let us now go back to the example of the influence of working memory on reading comprehension. In this example, the variable *verbal working memory* can be observed in two ways. The first possibility would be to measure the skills of the people taking part in the experiment by using a verbal working memory test. According to this evaluation and its results, participants could be sorted into groups. By doing so, every participant is included under a variable *modality* (e.g. high competence or low competence) depending on his/her own characteristics, as some people have better working memory capacities than others. In this case, the variable is simply observed during research.

A second possibility would be to *manipulate* the variable *verbal working memory*, by implementing conditions within the experiment where this variable has different modalities. In our example, the manipulation of the independent variable would aim at restricting the use of verbal working memory in some of the participants, in order to see the impact of such manipulation on reading comprehension, as compared to other participants whose working memory has not been restricted during the reading. A common task used for manipulating verbal working memory is to ask people to momentarily memorize different series of letters while reading the text, to report them and then to memorize others. Having to remember a series of letters while reading the text reduces the verbal working memory storage capacity used for reading and makes it possible to show a connection, if existent, between working memory and comprehension.

In general, in experimental research, the aim is to manipulate all the variables involved in the hypotheses. However, due to practical or ethical reasons, this is not always possible. For example, age, socio-economic level, bilingualism, etc., cannot be manipulated because they are inherent in people. When variables can be manipulated, the decision to manipulate them, as well as the way in which to manipulate them, must follow ethical principles, ensuring that research will not harm the participants during the test. The cost/benefit relationship must be clearly considered when pondering the possibility of manipulating a variable or not. For example, imagine that you formulate a hypothesis stating that in stressful situations, people tend to speak faster than in non-stressful situations. In order to study the influence of stress on articulation rate, you could decide to manipulate the participant's stress level. To set up a stressful condition, you could imagine putting some of the participants in a dark room in front of an audience booing at them. In experimental terms, such manipulation would be adequate, in the sense that a high level of stress would most likely result from your manipulation. On the other hand, it would be totally inappropriate from an ethical point of view. Actually, this type of manipulation would affect the participants to a much larger extent than needed, and they would probably not leave the experiment unscathed. Although this is an extreme example, it illustrates the fact that an experiment should not leave an impact trace on the participants once the experiment is over. We will develop this point in Chapter 6, which is devoted to the practical aspects of an experiment.

1.2.3. *Control of external variables*

We have seen that when operationalizing research hypotheses, variables need to be defined with accuracy. The main purpose of such a definition is to *isolate* the variables studied within the experiment, in order to reach a reliable conclusion as to the relationship between them. In parallel, and for the same purpose, it is necessary to *control* the other variables, known as *external variables*, which could influence the variables and the results obtained in the experiment. External variables can be multiple and we will return to them in Chapters 2 and 6, where we will discuss hypotheses and the practical aspects of an experiment. However, it is generally acknowledged that the characteristics of the participants are variables which may interfere with the variables investigated in an experiment.

Going back to the example of the influence of memory on reading comprehension, we may assume that educational level, general cognitive abilities, age, reading habits, etc., can influence both memory and reading comprehension. Likewise, the characteristics of the material used in the experiment may have an influence on the results. If, in the above-mentioned example, we use very simple text and questions, it is possible that everyone answers the questions perfectly well, regardless of their memory skills. On the contrary, if the text and the questions are very complicated, it is possible that very few people will be capable of answering. In these cases, we risk not finding a connection between memory and reading comprehension, not because the link doesn't exist, but because the material used for the experiment is not suitable for evidencing such a link.

1.2.4. *The notions of participants and items*

To attenuate these potential problems, and to reduce the importance of the characteristics of the participants or the material employed, experimental research is based on data collected from a large number of people, using a broad palette of materials. Referring back to our example, it would be necessary to test a large number of people by means of a comprehension test. This test should contain multiple texts and different questions for each of them. In general, the material used in an experiment is defined as a set of *items* (the texts or the questions in our example are items). The ideal number of participants, as well as the number of items necessary to undertake proper research, is a complex question, which we will address in Chapter 6.

Furthermore, experimental research is generally carried out by recruiting *naive* participants, who ignore the goals of the experiment and who have zero expertise in the subject under study. This precaution aims to try to control certain cognitive biases that could influence research results. The first bias is related to the fact that the participants who know the research hypothesis may try to base their answers on this hypothesis. Should this happen, the results obtained could suffer from what is called *confirmation* *bias*. Rather than answering naturally, participants could provide answers based on the hypothesis to confirm it, not because the assumption is correct, but rather because it seems adequate to them (even if this is not the case). The second bias is related to the fact that participants may want to help the researcher. If the participants know or suspect the goal of an experiment beforehand, the results obtained in this second scenario may not correspond to reality, but rather to the answers that the participants presume are expected.

Finally, in experimental research, participants are generally assigned to conditions in a *random* manner. This means that every person has the same chances of being included under one condition of the experiment or another. This random assignment offers additional protection against the effect of uncontrolled external variables. In addition to testing a large number of people, randomly distributing them to the different conditions reduces the probability that external variables could systematically influence the results. However, this random assignment is only feasible when all variables are manipulated. When one or more variables are simply observed, participants must be included in one condition or another on the basis of their own characteristics, such as gender or age, for instance. In this case, we speak of *quasi-experimental research*, since it is not possible to control all the variables. Leaving this question aside, experimental and quasi-experimental research is very similar, and the elements developed in the following chapters apply to both types of research.

1.2.5. *Use of statistics and generalization of results*

The last essential characteristic of experimental research concerns the way in which data is analyzed. Experimental research aims to collect quantitative data that can be statistically analyzed. As we will see in Chapter 7, quantitative data can be described using different indicators, such as the mean, for example. Based on these descriptive indicators, it is possible to obtain an overview of the data collected, to summarize and illustrate them, in order to communicate the results with simplicity.

At the second stage, data is used to draw conclusions about the research hypotheses. In experimental linguistics, the aim is to study and understand a linguistic phenomenon for a specific population. Since it is impossible to test an entire population, researchers collect data from a representative sample.

Through the use of inferential statistics, it is possible to determine whether the results of a particular sample are applicable to the whole population. This process is called *generalization*.

1.3. Types of experiment in experimental linguistics

Experimental research can be applied to all areas of linguistics, even if historically some areas have used such a methodology more consistently than others. Research questions vary widely between linguistic fields, meaning that many different methods and measures can be used in experimental linguistics. In this book, we do not aim to offer a detailed presentation of every research field and the methods associated with each, but rather to provide an overview of the principles of experimental methodology and the available techniques for linguists. Here, we will introduce some major classes of experiments that can be carried out in linguistics, and we will then develop these in every dedicated chapter.

In general, the experimental studies carried out in linguistics can be classified depending on the aspect of the language under study. Alternately, we will discuss studies on linguistic production and those relating to language comprehension. We will see that the study of comprehension poses many challenges, since this process is not directly observable. For this reason, research on language comprehension is based on the observation of indirect measures, which can be explicit or implicit. We will also see that it is possible to study comprehension by observing different stages of this process, either while it is in progress or once it has been completed.

1.3.1. *Studying linguistic productions*

The first type of linguistic experiment aims to investigate language production, all the manifestations of language that are produced by individuals in a certain language. Although these manifestations can be collected from diverse corpora and then studied through corpus analysis (see Zufferey (2020) for a detailed presentation of these methods), in some cases, the data contained in the corpus is not enough for studying a linguistic phenomenon. Some rare phenomena practically do not appear, if at all, in a corpus. What is more, the use of observation of naturally produced data is not suitable for showing the influence of a variable on the emergence of a specific linguistic phenomenon, as we have already seen. To counter this, different experiments can be implemented in order to study the production of linguistic phenomena. In these experiments, the goal is to purposefully elicit the emergence of certain linguistic structures, while controlling the context in which such structures appear. The experimental study of linguistic production will be described in further detail in Chapter 3.

1.3.2. *Explicit and implicit measures of comprehension*

The second type of experiments used in experimental linguistics include studies conducted on the mechanisms involved in language processing and comprehension. Such processes are numerous and range from the organization of the lexicon, to the comprehension of a text or a discourse. It is therefore the most broadly studied aspect in experimental linguistics. Unlike some aspects of the production component, the language comprehension component is unique, in that it cannot be directly assessed through mere observation. It is outright impossible to directly observe the processes involved in the comprehension of a text, for example. This is why it is necessary to find a way to measure these processes indirectly, based on indicators that can be associated with them.

The first way of collecting these indicators requires the use of *explicit tasks* in which participants have to reflect upon certain linguistic aspects. For example, this is the case for metalinguistic tasks such as grammaticality or acceptability judgments. This type of task could be used to test the participants' grammatical knowledge, by showing them syntactically correct or incorrect sentences in compliance with grammatical standards, and asking them to identify errors and justify their choice. While these tasks have the advantage of providing direct access to speakers' knowledge, they also have the defect of being based on their reflexive skills and their subjective appreciation of their own understanding. These tasks are also particularly complex for certain types of people, especially for children or people with language impairments, for whom it is often very difficult to explain the reasoning behind their decision. Other tasks make it possible to circumvent these problems, by setting up experiments in which the participants have to choose between several illustrations matching a linguistic stimulus. For example, Durrleman *et al.* (2015) tested the comprehension of relative sentences in people with autism spectrum disorder (ASD), asking them to point to the image corresponding to sentences such as "show me the little boy running after the cat”. Making use of such tasks offers the possibility of studying language comprehension in children and populations suffering from linguistic impairments.

Alternatively, methods for studying comprehension in an *implicit* manner (without asking the participants directly for a judgment or an explanation of their reasoning) have also been developed. This is the case in action tasks, in which some kinds of behavior adopted on the basis of a linguistic stimulus can be observed. For example, Pouscoulous *et al.* (2007) tested the understanding of scalar implicatures triggered by words such as *quelques* (roughly equivalent to *some*), by asking French-speaking children to arrange tokens in boxes so as to match statements like “quelques cases ont des jetons (some boxes have tokens)”. It is also possible to understand comprehension skills using recall or recognition tasks, in which questions are asked at the end of a reading exercise or after listening to a text or speech fragment. For example, Zufferey *et al.* (2015a) tested the comprehension of causal relations in children aged 5–8 years, by asking them to answer *why* questions after every page, when reading a story with them.

1.3.3. *Offline and online measures of comprehension*

The various tasks listed above, as well as the tasks proposed in the examples presented so far in this chapter, enable access to comprehension once the word, sentence or text has been processed and understood. These measures are described as *offline*, in that they affect the final interpretations resulting from the comprehension process. On the other hand, *online* measures allow us to study the processes that come into play in comprehension itself. Such processes have the characteristic of being extremely fast, transient and occurring out of people’s consciousness, therefore remaining inaccessible to traditional offline measures.

Borrowing scientific methods and paradigms from other disciplines, such as psychology, has allowed the study of online processes involved in language comprehension. The majority of online measurement techniques have something in common: they observe the time required for a process, by measuring the reading time or reaction time. These techniques are based on the idea that the time required to complete a process reflects certain characteristics of this process, particularly in terms of complexity. Longer reaction times and reading times are generally associated with a more in-depth processing of the linguistic stimulus. Tasks using these time measures typically involve asking participants to name words, read or produce sentences, or decide whether or not a series of letters matches a word in their language. Studies that have employed such tasks have shown that, **at the word level**, response times and reading times are influenced by properties such as frequency, length and predictability. Similarly, **at the sentence level, reading** is influenced by properties such as syntax complexity or the need to produce inferences (Just and Carpenter 1980; Rayner 1998; Smith and Levy 2013).

Studies based on time measures have benefited from significant technological developments since the 1970s, so that today, anyone can easily conduct research from their computer. In addition, new techniques have been developed to enable the recording of eye movement whilst reading or when observing an image. It is thus possible to gain an insight, not only into the time required to read certain words or sentences, but also the exact movements made by the eyes during reading. This data provides additional information, such as the time allotted for different words, the order in which words are fixated or even the eye movements associated with reading certain passages. These eye movement measures can be applied to the study of reading as well as to the study of spoken speech production or comprehension.

Finally, the methods used in the field of neuroscience have also been transferred to experimental linguistics. These methods provide access to the brain activity involved in language-related processes. Using small electrodes placed on the scalp, the electroencephalogram (EEG) records the activity of neurons on the surface of the brain. This technique gives an accurate *temporal* overview of the activity of neurons associated with a specific linguistic process. Functional magnetic resonance imaging (fMRI) aims to measure the activity of neurons based on their oxygen consumption. It thus provides a precise *spatial* overview of the brain areas involved in a specific linguistic process.

As we can infer by reading these lines, offline methods are the most accessible to researchers, since they require few technical means. In most cases, offline measures can be collected using paper and pencil tasks. A simple spreadsheet available on every computer can be used for organizing and analyzing the data from such studies. For some statistical tests, a program must be added to the list of necessary tools. Online methods for observing reaction time or reading performance require special software for programming experiments. Things get more complicated when you want to record eye movements. These recordings require the use of expensive tools, that also take time to control. Furthermore, the data from studies on eye movements is much more complex to process. Finally, EEG or fMRI studies are generally reserved for people benefiting from access to such techniques, which are extremely costly in terms of equipment and necessary skills for processing recorded signals. For this reason, such techniques will not be discussed in this book.

Finally, we should point out that the offline and online measures do not provide answers to the same type of research questions. It is therefore important to consider them as complementary measures, which shed different light on the same phenomenon. There are no good or bad measures in experimental linguistics; the choice must be made on the basis of the goals and hypotheses of the research project. More and more often, offline and online measurements are used in parallel in the same study. We will return to these measures, their specific characteristics and the means for combining them, in detail, in Chapters 4 and 5.

1.3.4. *Research designs and experimental designs*

Whether for the purpose of studying production or comprehension, research can be categorized according to the general framework in which data collection takes place or, in other words, the experimental design. On the one hand, there are *longitudinal* designs, in which the same subjects are observed on several occasions, following varying time intervals. This type of design is generally used in studies where a variable cannot be manipulated, but its effect can be observed through time. For example, to study the influence of age on the ability to distinguish sounds between the different languages spoken in the environment of babies growing up in bilingual homes, one possibility would be to test **the same** bilingual babies at 2 months, 4 months, then 8 months old. Another example of longitudinal design would be the study of the relationship between language development and the development of theory of mind. In this case, language skills and individual differences in theory of mind could be measured in children aged 3 and a half, 4, and 4 and a half, for example.

The major interest of longitudinal studies is that they make it possible to observe changes in real time. However, they also have two significant disadvantages. First, such studies imply that participants must be tested on several occasions in relatively short periods of time. It is thus inevitable to lose participants during the study, due to motivation and availability reasons. Secondly, these studies generate significant costs, since it is necessary to find and then test people repeatedly, and above all, keep in touch with them and convince them to return to the following test sessions.

In order to work around these problems, *cross-sectional* designs observe different people, who are subjected to different conditions. To use the example of bilingual babies, instead of testing the same babies at different ages, we could simultaneously test groups of babies of different ages. This method would imply making a sort of picture of a situation at a given moment, which would offer indications on the relationship between age and sound perception in bilingual babies. Cross-sectional designs are typically used in quasi-experiments, where the independent variable is not manipulated.

When the independent variable can be manipulated, it is possible to allocate the participants to different conditions, in which manipulation can either be present or absent. Two types of experimental designs can be constructed in this case. In the first, the *between-subject design*, the participants only take part in one condition or the other. For example, to study the influence of reading goals on reading comprehension, one option would be to carry out an experiment in which a group of people reads a text in order to briefly summarize it, while another group reads the same text in order to answer questions about it. The performance of the two groups can then be compared during a recall task after reading the text. The results of such a task would certainly show that the second group performs better than the first group (as in Schmalhofer and Glavanov (1986), for example).

In the second type of experimental design, the *within-subject design*, also called *repeated-measures design*, the participants take part in all the conditions of the experiment. For example, such a design can be used in an experiment on the influence of word frequency on their processing time. In this case, each participant would see frequent words and infrequent words in order to cover all the modalities of the variable *frequency*. Among other things, this type of design makes it possible to control the external variables associated with the participants, given the fact that everyone falls under all conditions. Between-subject and within-subject designs each have advantages and disadvantages, which will be developed in Chapters 2 and 6.

For the moment, the main thing is to remember that there are many ways to organize research and that experimental research may adopt different designs, depending on the conditions under which the participants are tested.

1.4. Advantages and disadvantages of experimental linguistics

So far, we have shown how experimental linguistics is set within the scientific process and involves the use of quantitative methodology for studying language. On the basis of these principles, the results from studies in experimental linguistics are considered representative and can be generalized, unlike those from studies based on a qualitative methodology. The possibility of generalizing results is one of the strong points of research in experimental linguistics. However, this approach has a less positive corollary. Due to its empirical and quantitative nature, experimental linguistics needs to *measure* the linguistic phenomena it intends to study. While this can be relatively simple in some cases, the operationalization of complex processes (let us consider language comprehension, for example) implies the decision to observe certain indicators which could, at some point, not exactly measure what is desired. This issue will be discussed in more detail in Chapter 2. For now, it is important to keep in mind that just because something can be measured, this does not necessarily make it valid or reliable. Conclusions drawn on the basis of inadequate measures may not correspond to reality and may therefore lead to erroneous generalizations.

We have also seen repeatedly throughout this chapter, that experimental linguistics aim to identify the variables, also called factors, which can influence linguistic processes. For this, it is necessary to establish causal relations between variables by manipulating them. In this respect, experimental linguistics differs from another quantitative method, corpus linguistics, which aims to observe linguistic phenomena on the basis of natural data. While corpus linguistics can only account for a relation between some variables, experimental linguistics also makes it possible to explain the reasons underlying such connections between variables. These two methods are considered complementary, since they take place at different stages of the research process. Implementing a corpus study makes it possible to explore data and to uncover relationships between variables, relationships which can later be investigated experimentally, based on the hypotheses formulated as a result of data observation.

One of the prerequisites for establishing a cause and effect relationship between two variables, is the control of the conditions in which these variables are manipulated, as well as the identification of all the external variables that could influence the results. This necessity has the advantage of enabling solid conclusions as to the relationship between the variables, but risks keeping the experiment too far from reality. In certain comprehension experiments, such as reading tasks, for example, the sentences are presented word by word in order to measure the time allotted to each word. This way of reading differs enormously from natural reading conditions, where it is notably possible to go back in the text. For these reasons, experimental studies may lack ecological validity and not be completely generalizable.

This need for control also implies that each experiment can only investigate a specific hypothesis, in which every variable is operationalized in a certain way. For this reason, a specific experiment can only respond to a narrow research question. It is therefore essential to conduct a lot of different experiments to arrive at the comprehension of a phenomenon. Knowledge can then be built on the accumulation of experimental research related to such a phenomenon.

Beyond these limitations, the experimental methodology is a very important tool for linguistics, as it makes possible to study almost any research hypothesis. When linguistic phenomena are very rare or hardly accessible to the consciousness, it becomes essential for the construction of knowledge about these phenomena.

1.5. Where to access research on experimental linguistics

Before going further in this book, we offer a list of scientific journals publishing studies on experimental linguistics. As this is an extremely large and varied field, we cannot set up an exhaustive list of such resources. We will limit our choice to reputable journals in different fields of application. A large part of these journals originate from, or are related to, the field of psychology. As we have already seen, there is a close connection between psycholinguistics and experimental linguistics, due to the fact that they share common methods and measures. It is therefore unsurprising that studies in experimental linguistics are found in journals classified under the *psychology* section.

The following journals are excellent sources for finding research in experimental linguistics: Discourse Processes; Journal of Pragmatics; Journal of Phonetics; Journal of Experimental Linguistics; Applied Psycholinguistics; Second Language Research; Studies in Second Language Acquisition; Bilingualism: Language and Cognition; Cognition; The Quarterly Journal of Experimental Psychology; Journal of Memory and Language; Journal of Experimental Psychology: Learning, Memory, and Cognition; Language, Cognition, and Neuroscience; Behavioral and Brain Science; Psychological Science.

1.6. Conclusion

In this chapter, we first saw that the scientific process is based on the observation of concrete phenomena, whose systematicity enhances the development of explanatory theories. On the basis of these theories, it is possible to develop specific predictions which will then be tested, in order to refine or revise the existing theories. We then presented the difference between qualitative and quantitative approaches, in terms of the types of reasoning and possibilities of generalization. We also saw that quantitative research can adopt different types depending on the manner of observing the variables, as well as the control procedures carried out on them.

We then presented the characteristics of experimental research. Such research is based on a research question, making it possible to formulate clear hypotheses as to the relationship between two or more operationalized variables. In order to test these hypotheses, it is necessary to manipulate the variables involved in an experimental study, while controlling the other variables which may influence the results. In parallel, it is essential to test naive individuals, using numerous items, and to distribute the participants randomly under the different conditions. Finally, the data collected in an experimental study is mostly quantitative in nature, so that it can be synthesized and analyzed by means of statistical tests.

Studies carried out in experimental linguistics can examine linguistic production or comprehension. For the study of the latter, we have seen that there is a first differentiation between explicit and implicit measures, depending on the tasks. A second differentiation lies in the processes examined: while offline tasks focus on the results of comprehension, online tasks look into the comprehension processes.

Finally, we discussed the advantages and disadvantages of the experimental approach in linguistics, before suggesting useful resources for becoming familiar with this type of research.

1.7. Revision questions and answer key

1.7.1. *Questions*

1) How can inductive and deductive approaches be complementary for the construction of a research question?

2) Imagine a way to study the influence of fatigue on retrieving tip-of-the-tongue (TOT) words in a qualitative and then in a quantitative manner.

3) Which of the following assertions are empirically testable research hypotheses? How could you transform non-testable propositions into testable hypotheses?

a) There is a key factor in language learning.

b) The elderly suffer more from tip-of-the-tongue (TOT) failures in everyday life than younger people do.

c) French speakers find it easier to learn Italian than English speakers.

d) Short words are processed faster than long words while reading.

4) A researcher wishes to examine the influence of alcohol consumption on fluency while speaking a foreign language. Imagine how to observe the variable *alcohol consumption* and then how to manipulate it. What should you consider if you decide to manipulate this variable?

5) Why is it important to manipulate variables in experimental research?

6) What different types of external variables are there in research? What strategies are used for controlling them?

1.7.2. *Answer key*

1) The inductive approach consists of observing a phenomenon in its context in order to build comprehension. This approach is essentially flexible and gradually adapts to the emergence of new elements as they appear while studying a certain phenomenon. By contrast, the deductive approach

formulates hypotheses on the basis of a theory and verifies them by using data acquired in a controlled manner. When constructing a research question, it may be interesting to start with an inductive approach in order to understand the phenomenon that one wishes to study. The elements resulting from this approach can then be translated into the form of empirically testable hypotheses. For example, when interviewing bilingual people, it could appear that these people often report feeling like they do not have the same reactions or the same personality when they speak their mother tongue or their second language. On this basis, a hypothesis could be formulated as to the relationship between the language spoken and the personality of bilingual people. This hypothesis could then be tested by collecting objective data on the personality of the speakers in their mother tongue and in a different language.

2) One possibility for implementing a qualitative study would be to conduct in-depth interviews on the theme of the tip-of-the-tongue (TOT) phenomenon with a limited number of people. During these interviews, we could first let people express their feelings spontaneously, and then later ask them one or more questions specifically intended to collect their opinion on the influence of fatigue on this phenomenon. From these interviews, we could discover that people feel like they tend to have more TOTs when they are tired. Attempts to explain this cause could also be suggested.

One possibility for carrying out a quantitative study would be to give a task to a large number of people aimed at provoking TOTs, and then to compare the number of TOTs encountered by tired and by less tired people. For example, an adequate task would be to give word definitions to the participants and then ask them to name the word corresponding to each definition. The tiredness variable could simply be observed in the study, by asking the participants to assess their level of tiredness in order to classify them into two groups. Tiredness could also be manipulated by researchers, by creating conditions of tiredness to place participants in. One could decide to manipulate tiredness by asking one group to carry out a tiring task before the naming task, while the other group could rest for a moment before the task.

3) Remember that a testable research hypothesis must identify a cause and a consequence, as well as a directional relationship between them. Furthermore, it must be operationalized, that is, the measures of the various variables should be determined. Proposals (b) and (d) meet these criteria, which is not the case for proposals (a) and (c).

a) is a general assumption, which does not identify the key factor in question. This proposal could lead to a large number of different research questions. In order to turn it into a testable hypothesis, a factor for focus should be determined, as well as what is meant by learning a language. One possibility among many would be to hypothesize that the age when one begins learning a language determines the proficiency attained in such language a year later. The means of determining language proficiency should be further clarified; for example, it could correspond to the results of a standardized language test.

c) should clarify what is meant by ease. Ease of learning a language could be objectively measured by using a language test after a certain period of learning. This easiness could also correspond to another aspect of learning, such as the learners' perception about their learning processes. In this case, it could be measured using a scale on which learners would rate the perception of their ability to learn Italian, for example, from 0 (very difficult) to 10 (very easy).

4) To simply observe alcohol consumption, the researcher could go to a bar in the evening, ask those present for the number of glasses of alcohol consumed and then measure their fluency in speaking a foreign language. In this case, the danger would be that the different groups (less than two glasses vs. more than five glasses, for example) would not be equal with respect to other variables that could influence the dependent variable, such as mastery of the foreign language, or the habit of expressing themselves in that language. External variables could be controlled by manipulating the variable, for example, by choosing two equal groups of participants who would be asked to drink water or alcohol, before testing their fluency in speaking a foreign language. In the second case, ethical questions would arise as to the cost/benefit ratio of such research. It would also be advisable to help participants having consumed alcohol, leave the experiment in a similar state to when they arrived.

5) Manipulating variables makes it possible to establish a cause and effect relationship between independent variables and dependent variables. By controlling the external variables, so as to isolate the independent variable and by establishing precise conditions, it is possible to draw conclusions about the influence of one variable on another.

6) External variables are uncontrolled variables which can influence the variables examined in a study. These variables can be related to the participants and/or to the items of the experiment. In order to control them, it

is necessary to recruit many people and to use many items in an experiment. Moreover, people should be randomly assigned to the different conditions and should be naive, that is, they should ignore the goals of the experiment. Confounding variables, a class of external variables, are related to both independent and dependent variables. The existence of confounding variables casts doubts on research findings, which is why it is necessary to identify and control them, for example, by keeping them constant between conditions. In Chapters 2 and 6, we will see that there are different ways of controlling external variables.

1.8. Further reading

Litosseliti (2018) provides an overview of the difference between qualitative and quantitative research in linguistics, as well as the means of combining them. For a more detailed introduction to quantitative research and experimental research, we recommend Gass (2015) and Phakiti (2015), devoted to applied linguistics but whose principles are properly suited for all areas of linguistics. To deepen the comprehension of the different types of quantitative research and scientific logic, we recommend Chapter 1 by Field and Hole (2003). Even if this work is primarily intended for psychologists, the examples are clear enough for all people to understand. Finally, for a point of view committed to experimental linguistics in the fields of syntax and semantics, we recommend the articles by Gibson and Fedorenko (2010, 2013), as well as the response by Culicover and Jackendoff (2010) for the opposite opinion. For the principles of corpus linguistics, see Zufferey (2020).

2

Building a Valid and Reliable Experiment

In Chapter 1, we reviewed the characteristics of an experimental study. We saw that carrying out an experiment comes down to manipulating at least one variable, in a controlled manner, in order to bring to light its effect(s) on one or more other variables. We also stressed the fact that each experiment should be used for testing a precise research hypothesis in which the observed variables are defined via an operationalization process. In this chapter, we discuss the different stages involved in the operationalization of a research question in detail. We see that the operationalization process requires making many choices as to the variables studied and the manner of measuring them and the conditions examined and the experimental design chosen. At the same time, these choices have consequences for the validity and reliability of the experiment. Thus, we begin this chapter with a presentation of the key concepts of validity and reliability. Secondly, we develop the notion of a variable introduced in the first chapter, in order to accurately define the types of variables involved in an experiment. Once this framework has been set, the rest of the chapter will deal with the stages for operationalizing a research question.

2.1. Validity and reliability of an experiment

In Chapter 1, we saw that the purpose of an experiment is to collect data in order to test a research hypothesis that states a cause-and-effect relationship between variables. In order to be valid, an experiment should lead to a trustworthy conclusion about this relationship, while ensuring that the results are not influenced by other variables not considered in the study. In other words, the cause identified in the research hypothesis must be the

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

origin of the effects observed in the results. This is called the *internal validity* of an experiment. For example, in the case of an experiment aimed at showing a relationship between word length and reading time, it is necessary to ensure that the other variables that could have an impact on reading time do not influence the results. In this experiment, only word length should vary, whereas word frequency, the grammatical category or the number of phonological neighbors, for example, should be controlled. We will later return to this notion of control.

In the first chapter, we also saw that the results of an experimental study must be generalizable, and make it possible to draw conclusions concerning the relationship between two variables, regardless of the sample of participants and items included in the study, and of the conditions under which the study was carried out. This is called the *external validity* of an experiment. For example, going back to the afore-mentioned study on the relationship between word length and reading time, if the subjects studied were all 35–45-year-old women, the external validity would not be met, since the results of the study could not be generalized to men or women belonging to other age groups.

In addition to being valid, an experiment must also be *reliable*, that is, it must produce consistent results. In other words, if the same experiment was carried out several times, the results should demonstrate the same effects. For this reason, the results obtained in an experiment should be replicated in successive experiments before being communicated. However, in practice, this has rarely been the case, due to the fact that replicating an experiment is costly in terms of time and resources. Thus, the use of similar methodologies or similar tasks by the same or by different research teams has long been considered as a roundabout way of ensuring the reliability of a result. This practice is now being called into question, and more and more voices are rising in favor of the application of different means for ensuring reliability. One of these means is, for example, pre-registering the hypotheses, the research method and the analyses planned for each study. Another means is the establishment of *open science* platforms for sharing the data collected, the analyses carried out or even the different versions of the scientific articles reporting on the study. The discussion of the problem of reproducibility goes beyond the scope of this book, but we can only encourage readers to learn about these practices before embarking on the path of research. A good starting point is the *Center for Open Science*[1] site which presents the steps to be followed for conducting transparent and open research.

Internal and external validity, as well as the reliability of an experiment, are influenced by many factors, which we will address throughout this chapter, and which we will illustrate by numerous studies in the following chapters. The validity of an experiment also crucially depends on the way in which the variables are chosen. In section 2.2, we will detail the types of variables that can be included in an experiment.

2.2. Independent and dependent variables

Let us recall that experimental research aims to empirically verify a cause-and-effect relationship between at least two variables. In general, a distinction is made between *independent* variables (the causes) and *dependent* variables (the effects). Independent variables are the parameters we identify as being responsible for influencing the value of one or more dependent variables. In other words, an independent variable is the variable whose effect we want to evaluate, the one that is manipulated in the experiment. The dependent variable, on the other hand, is the variable that is modified in accordance with the independent variable, the one whose change we want to measure. Let us take a first intuitive example. Let us imagine that we wish to find out whether the lack of sunshine causes seasonal depression. In that case, the independent variable would be the sunshine rate, and the dependent variable, depression. Let us now take linguistic examples. If the research question is “Do Australian people speak faster than American people?” the independent variable is the person’s nationality (Australian or American), and the dependent variable is the articulation rate. For the question “At what age do children begin to understand scalar implicatures in the same way as adults?” the independent variable is the age of the children and the dependent variable is the understanding of scalar implicatures. Each experiment includes at least one independent variable and one dependent variable, but it can also include several independent variables and/or several dependent variables.

[1] http://cos.io.

2.3. Different measurement scales for variables

Experimental research is based on the quantification of observable responses or types of behavior. Whether independent or dependent, a variable must be measured in order to be included in experimental research. According to the type of measurement scale used, a variable can be either qualitative or quantitative. These two general categories can, in turn, be subdivided, as we will see later.

2.3.1. *Qualitative variables*

Qualitative variables correspond to variables that are not numerical, but describe categories, such as having a specific mother tongue or a certain nationality. This type of variable includes variables which can be measured on two types of scales: nominal and ordinal.

The values of *nominal* scales correspond to categories including individuals or similar things sharing some characteristic, for example, the fact of defining oneself as male or female. These values can be defined by nouns (masculine gender or feminine gender), or numbers (e.g. 1 for the masculine gender and 2 for the feminine gender), which bear no size relationship to each other. In the case of our example, value nos 1 and 2 do not offer any indication of a size difference between the feminine and masculine genders (the feminine gender is not worth twice the masculine gender), but simply corresponds to a means of defining or of categorizing a group. Numbers assigned to nominal scale values are often used for data coding purposes and should not be subjected to arithmetic tests. It would indeed be inappropriate to calculate an average of the kind of people participating in a study. On the other hand, it is possible to calculate a number for each of the variable’s condition or modality, in other words, from the total, how many people taking part in the experiment defined themselves as male or defined themselves as female.

Other examples of nominal scales can be mother tongue, marital status or a *yes/no* answer to a question. In all these cases, there is no hierarchy between the different categories, which are simply a list of possibilities. Frequently, the independent variable of a research project is measured on a nominal scale for creating two or more conditions under which the dependent variable can be observed. Most of the examples presented so far and in the previous chapter illustrate this scenario: studies comparing monolingual versus bilingual people, less frequent words versus very frequent words, people who have had a language stay versus those who haven't, people with a high working memory capacity versus those with a lower capacity, or people who have to perform a verbal working memory task while reading a text versus those who do not have to perform this task.

The second type of scale associated with qualitative variables offers more information on the relationship between the different values of the scale. This is the *ordinal* scale, whose values can be ordered, although the size of the difference between the values cannot be evaluated. These values can also correspond to tags such as *a little*, *a lot* and *passionately*, or to categories.

For example, imagine that people taking part in an experiment are included in the following age categories: 15–25 years old, 25–35 years old, 35–45 years old. By assigning every participant to their age category, it is possible to classify them. For example, if participants 1, 4, 6 and 7 belong to the 15–25 years old category and participants 2, 5, 8 and 9 to the 25–35 years old category, participants 1, 4, 6 and 7 should be ranked before the others on the age scale. However, in this configuration, it is not possible to determine the order in which participants 1, 4, 6 and 7 appear, because the only known indicator is that they belong to the same category. In other words, even if these participants are not the same age, this type of information cannot be retrieved from the data.

Ordinal scales do not offer an indication of the size of the difference between the values of the scale. In the case of our example, even if every category spans 10 years, it is not possible to conclude that a participant in the first category and a participant in the second category are 10 years apart. It is indeed possible that the first participant is 15 years old and the second 35 years old. The response scales typically used in questionnaires are another illustration of the impossibility of assessing the size of the difference between the values in ordinal scales. For example, in order to measure language proficiency, one could ask people to assess their level on a scale between 1 and 7, where 1 represents poor fluency and 7 represents perfect fluency. By observing the values chosen by people on such a scale, it would be possible to deduce that people with a score of 4 have better skills than people with a score of 2. However, it would not be possible to say that people with a score of 4 have twice the language proficiency than people with a score of 2. For this reason, it is not appropriate to perform arithmetic tests on the data obtained by means of ordinal scales.

In experimental linguistics, some independent variables are often measured on an ordinal scale through the use of categories. This is the case for the age of speakers and the number of years of residence in a country, for example.

2.3.2. *Quantitative variables*

Unlike qualitative variables, quantitative variables can be subjected to arithmetic tests because they use scales based on quantifiable values. There are two types of scales: interval scales and ratio scales.

Interval scales are similar to the ordinal scales we have already described, but differ from those in that the interval between the different categories always has the same value. It is thus possible to perform arithmetic operations on the differences between the values of the scale. Conversely, interval scales do not have an absolute zero. Therefore, it is not possible to perform operations on scale values. A simple illustration of an interval scale and its properties is the temperature scale. On this scale, the difference between 5°C and 10°C is the same as that between 20°C and 25°C, that of 5°C. On the other hand, a temperature of 30°C does not correspond to a heat three times higher than a temperature of 10°C.

The difference between ordinal scales and interval scales is simple on paper, but complicated in some cases. Let us take the example of an experiment in which the participants have to judge the acceptability of sentences, on a scale of 1–9, with 1 being equivalent to *totally unacceptable* and 9 to *totally acceptable*. In order to be able to consider this scale as an interval scale, we should assume that the difference between values 1 and 2 is the same as between values 4 and 5, or 8 and 9, for example. It would also imply that a sentence with a score of 9 is considered as more acceptable than a sentence with a score of 6, in the same proportion that the latter would be more acceptable than a sentence with a score of 3. As we can see here, it is impossible to formally verify these conditions and, in absolute terms, the scale of acceptability of this example should be considered as an ordinal scale. However, it is accepted that if a scale is presented in such a way as to highlight equal differences between the different scores, it is likely that people will assess the differences between the scores as being equal. This is why, in practice, we often consider these response scales as interval scales.

The second type of scale that quantitative variables are based on corresponds to the *ratio scale*. Just like the interval scale, this scale has equivalent intervals between its values. In addition, this scale has an absolute zero. This means that it is possible to perform operations on the values themselves. For example, sentence length calculated per number of letters is a ratio scale. If a sentence contains 178 letters, it is twice as long as a sentence containing 89 letters. Likewise, when we measure it continuously, age is considered as a ratio scale. A 60-year-old woman is three times older than a 20-year-old woman. Online measurements used in experimental linguistics, such as response time or reading time, are typically considered as ratio scales. It is also possible to measure the opinion of people using a ratio scale, for example, by presenting a non-gradation line on which the interviewees have to indicate their degree of agreement. We can then measure the distance between the start of the line and the answer in order to obtain the value representing the level of agreement with each statement.

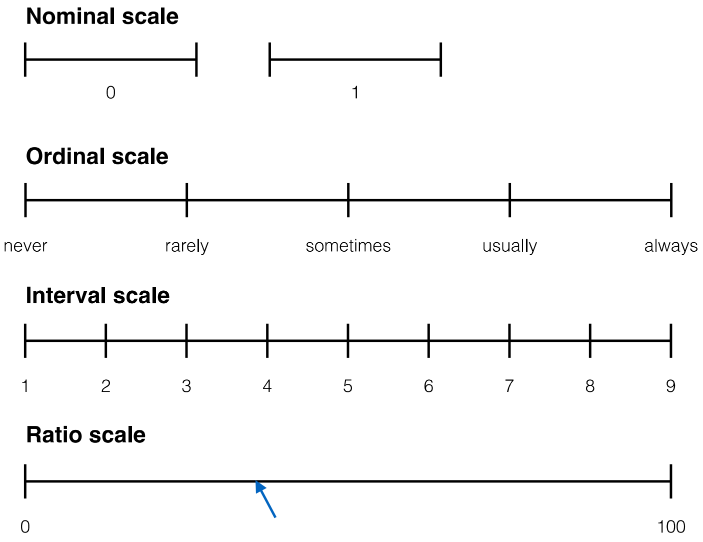


Figure 2.1. *Illustrations of nominal, ordinal, interval and ratio scales*

To sum up, the different measurement scales have divergent properties and cannot be subjected to the same arithmetic operations. The simplest scales, called nominal scales, only make it possible to differentiate between categories. With ordinal scales, categories can also be ranked. Interval scales also make it possible to take into account the distance between the different categories. Finally, ratio scales make it possible to perform all arithmetic operations on the scale values. As a corollary, ratio and interval scales can be transformed into ordinal or nominal scales. To do this, values can simply be grouped into categories. For example, if the exact age of the participants is recorded at the time of the experiment, it is then possible to set up different categories. Similarly, an ordinal scale can be transformed into a nominal scale, by simply decreasing the number of categories. In this way, we can see that it is always possible to go from a more accurate scale to a less accurate scale, but the opposite cannot be done.

Due to their different properties, the types of scales we have just described do not allow the same statistical tests. It is useful to know that data from ratio scales and interval scales can be subjected to what are known as parametric tests. These are compatible with a wide variety of statistical analyses and are most commonly used in research. We will present them in Chapter 7. The data obtained from nominal and ordinal scales can be subjected to non-parametric tests, which offer fewer possibilities for analysis. From the very beginning of research, it is extremely important to proceed with caution when choosing the types of scales to be used, since these will not only shape the analyses that can be performed on the data, but also the conclusions that will be drawn on the basis of the analyses.

2.4. Operationalizing variables

Now that we have seen the characteristics of the variables involved in an experiment, we will describe the operationalization process, whereby the variables of interest are defined in terms of measurements. While some variables can be easily measured using objective indicators, others are more difficult to operationalize. On the one hand, age, mother tongue, word length, sound frequency and reading time are variables that can be measured directly and objectively. On the other hand, language proficiency, the derivation of inferences, the understanding of discourse connectives and even the access to the meaning of a word cannot be measured directly. As a matter of fact, this type of variable refers to abstract concepts which (1) cannot be observed directly and (2) are based on definitions or theoretical models. These variables require a process of reflection on how to operationalize them, by which we try to define a signal of the abstract concept. Measuring this signal amounts to measuring the abstract concept, in a roundabout manner.

To illustrate this process with a concrete example, let us imagine a plane flying in the sky and leaving a white trace behind. By seeing its trace, we know that an airplane has passed, even if the airplane is no longer visible. Thus, the trace works as a signal, making it possible to deduce the presence of the aircraft in a more or less precise manner, depending on its quality. If we apply this idea to some of the abstract concepts mentioned above, we can consider the results of language tests as a signal of language proficiency and certain eye movement patterns while reading a text (e.g. going back on certain words, longer reading time) as a signal of inference construction. Drawing a parallel, the goal of researchers is to measure the signal that best reflects an abstract concept, in a similar way as a clear and good quality trace should be as close as possible to the path of the plane.

During operationalization, we have to define not only the signal we want to measure, but also the scale on which this signal will be measured. We have already pointed out that experiments very often compare different groups of participants in relation to the values of the dependent variable. In general, the independent variables involved in research are measured on nominal (monolingual vs. bilingual, for example) or ordinal (different age categories) scales, in order to be able to create conditions, whereas dependent variables are preferably measured on interval or ratio scales (e.g. the number of correct answers given to a linguistic task). We will first review the general choices to make when choosing a measure for the variables before turning to the specifics of the independent variable.

2.5. Choosing a measure for every variable

There are different ways to operationalize the linguistic concepts investigated in experimental research. The choice of the measure mainly depends on the process that one wishes to examine, as well as the possibility of gaining access to it in a more or less direct manner. When the process is accessible to consciousness, one possibility is to use a survey response scale, as is the case of studies aiming to measure the acceptability of sentences according to their syntactic structure. When the process is not accessible to consciousness, or when one wishes to measure it implicitly (see section 1.3.2), it is possible to use behavioral measures. For instance, this is the case of studies using action tasks (such as performing an action on the basis of a sentence), and those measuring reading time or reaction time. Finally, when the process is not accessible through behavioral measures, it is possible to measure physiological reactions, as is the case in studies using an electroencephalogram, or magnetic resonance imaging (MRI), for example. Certainly, these different methods can be combined within the same study, in order to shed light on the same process from different perspectives.

These multiple operationalization methods do not only concern different concepts. In fact, there are many ways of operationalizing the same abstract concept through the use of different types of measures. Let us suppose that you are interested in the theory of linguistic relativity, according to which the speaker's language may have an influence on their world view or cognition. Studies have shown that the way a language encodes different phenomena such as time, colors or gender has an influence on the representations that speakers have of such phenomena (e.g. Vigliocco *et al.* 2005; Athanasopoulos *et al.* 2011; Boroditsky *et al.* 2011). The representations often examined in these studies are those built on the basis of grammatical gender. Some languages, such as French or Italian, have two grammatical genders, feminine and masculine, and all nouns in these languages are related to a grammatical gender. In French, for example, this association can be based on the person's gender, such as *un infirmier* (a nurse, masculine) or *une astronaute* (an astronaut, feminine). It can also be completely arbitrary, as when talking about *une chaise* (a chair, feminine), *une tomate* (a tomato, feminine), *un train* (a train, masculine), *un clavier* (a keyboard, masculine) or even *une envie* (a desire, feminine) or *un souhait* (a wish, masculine). In other languages, such as English or Japanese, nouns do not have a specific gender (English has certain exceptions, such as a ship or a bell, which can be regarded as feminine). On the basis of this difference, one might wonder whether the speakers of languages with grammatical genders associate feminine or masculine characteristics with certain words in the language, depending on their grammatical gender. In order to carry out a study on this phenomenon, it would be necessary to define what we understand by the *presence of grammatical gender in the language* and *feminine or masculine characteristics assigned to certain words* and how these variables would be measured.

A first possibility for operationalizing such a question can be drawn from the study by Konishi (1993) in which German-speaking and Spanish-speaking participants had to evaluate the words *man* and *women* in their language, as well as nouns for objects (*newspaper*, *cigarette*), places (*mountain*, *desert*) or abstract concepts (*love*, *record*) on a potency scale. The nouns to be evaluated (except *man* and *women*) were selected according to their German grammatical gender, which was always different from their Spanish grammatical gender. Half of the words presented in each language were masculine and the other half feminine. In this study, the independent variable was the words' grammatical genders in the participants' language. Words were either masculine in Spanish and feminine in German, or feminine in Spanish and masculine in German. The score on the potency scale made it possible to operationalize the concept of masculinity, in order to measure the dependent variable. In this example, we can see the transformation of an abstract concept, *considering an object as more or less masculine*, into a response on a potency scale. The results of this study showed that the participants subjectively placed the word *man* on a higher rank on the potency scale than the word *woman*. In addition, masculine words were also rated as more powerful than feminine words in both languages.

A second example on how to operationalize this research question comes from a study described by Boroditsky *et al.* (2003), in which Spanish-speaking and German-speaking participants carried out a memorization task in a language that does not have a grammatical gender, namely English. As in Konishi's (1993) study, Boroditsky *et al.* compiled a list of objects with opposite grammatical genders in German and Spanish. Half of the nouns referring to the objects were masculine in Spanish and feminine in German, whereas the other half were feminine in Spanish and masculine in German. Setting up object–name pairs, each object was associated with a first name, congruent with the object's gender (e.g. *apple*-*Patricia*) for half of the participants and incongruent for the other half (e.g. *tomato*-*Peter*). The participants had to learn the first names associated with the objects and were then tested on their memory of the pairs. Here, the independent variable was operationalized as the association between a first name and an object's grammatical gender in the participant's mother tongue and had two modalities: congruent (masculine object and masculine first name or feminine object and feminine first name), or incongruent (masculine object and feminine first name, or feminine object and masculine first name). The dependent variable corresponded to the number of correct answers provided during the recall task. The results of this study showed that the participants remembered the first names associated with the objects under the congruent condition better than those under the incongruent condition.

A third example of operationalization can also be found in Boroditsky *et al.* (2003). Once again, participants whose mother tongue was German or Spanish completed a task in English, using a list of words of opposite grammatical genders in German and Spanish, similar to that of the previous study. This time, the participants were asked to list, for each word, the first three adjectives that came to mind. A group of English speakers then evaluated these adjectives to determine whether they predominantly related to female or to male characteristics. In this study, the independent variable was operationalized as the grammatical gender of a noun for an object in the participants' mother tongue (feminine or masculine). The dependent variable corresponded to the feminine or masculine perception of the adjectives attributed to words. The results showed that the adjectives associated with feminine words in the participants' mother tongue were assessed as predominantly feminine when compared to the adjectives associated with masculine words. The words that were feminine in Spanish but masculine in German were associated by the Spanish-speaking participants with adjectives perceived as predominantly feminine, whereas they were associated by German-speaking participants with adjectives perceived as predominantly masculine. The opposite was also true for words that were masculine in Spanish and feminine in German.

These three examples illustrate the fact that it is possible to operationalize the same concept in different ways. How can we decide on how to operationalize variables for research? A first clue can be found in the scientific literature already published on the subject of interest. It is therefore strongly advised to build on existing studies and to pay special attention to the way in which the variables have been operationalized. An in-depth literature review should make it possible to identify the different measures that have been used so far, as well as the results obtained on the basis of these measures.

The choice of a measure for operationalizing a variable should also be made keeping in mind the statistical analyses that will later be performed on the data. In fact, the quantitative data acquired in an experimental study must be statistically tested, in order to check whether an effect is real or not. In addition, as we have already discussed, the different measurement scales are not compatible with the application of all the statistical tests. This is something that is very important to think about before collecting the data, in order to avoid reaching the last stage of the research process and realizing that the data cannot be analyzed as they should be.

2.6. Notions of reliability and validity of measurements

Finally, the essential element when choosing how to operationalize variables is to ensure the quality of the measurement. In the same way as for an experiment, this quality can be assessed by means of two concepts: the validity and the reliability of the measurement. The *validity of a measurement* refers to how well it measures what it intends to measure. Imagine an experiment in which we want to study the effect of word length on reading time. One way of measuring word length could be to count the number of letters in each word. It could also be possible to decide to count the number of syllables, rather than the number of letters. To measure the reading time, you could present isolated words on a computer screen and ask people to press a key when the word has been read. By measuring the time between the word's appearance on the screen and the key press, it would be possible to deduce how long it took for each word to be read. In this example, the number of letters and the number of syllables are both valid measures for calculating word length. As regards the measurement of reading time, the proposal made here also seems valid, in that it makes it possible to accurately record the amount of time taken by people to read each word.

Let us think about an experiment designed to examine the impact of people's personality on their ability to learn a foreign language. We can see that the variables of this research question are much more abstract than those previously examined and that they cannot be directly observed. Let us first examine the dependent variable in this question, namely the ease of learning a foreign language, and explore some different ways of operationalizing it. A first option would be to directly ask for people's opinion by offering them to assess this ease on a scale from 0 to 5, for example. A second option would be to evaluate them after a few months of learning by means of a dictation in their second language and counting the number of errors. A third option would be to assess people's skills in the areas of language production and comprehension after a few months of learning, using standardized tests, that is, tests developed and validated in previous studies.

The first option is the least valid measurement, because it is based on a subjective assessment and on a single question, which can be interpreted in many different ways among participants. It is therefore very likely that the scores on this scale do not directly reflect the ease of learning, or, in some cases, not the concept that the researchers desire to measure. The second option seems to be a more valid measurement than the first, in the sense that it practically leaves no room for interpretation, since a number of correct answers is a concrete and objective measurement. However, it measures the ease of learning a foreign language on the basis of a single language-related task, a dictation, only reflecting spelling competence. Language-related skills obviously go well beyond the simple fact of not making mistakes in a dictation. Consequently, the validity of this measurement is not suitable, since it is too distant from the construct it is supposed to assess, and only targets one facet of language. On the other hand, the third option, which evaluates different skills in the second language, makes it possible to measure the dependent variable more comprehensively. Besides, as this measurement has already been used and validated, it seems the most adequate one.

Now, let us go back to the examples on how to operationalize the concept *assigning feminine or masculine characteristics to words*, described above. We can notice that, in all cases, the measurement was a relatively distant signal from the original concept. In the first case, measuring the potency associated with a word as a signal of its masculinity is based on the idea that potency is a good indicator of masculinity, as a result of the gender stereotypes present in society. This measurement also assumes that the perception of potency is directly related to the grammatical gender of the word, rather than to other characteristics, such as its phonology, for example. In the second study, measuring the recall of word-first name associations is based on the idea that people generally remember congruent things, and that congruence is based on the association of the word's grammatical gender with the gender of the first name, and not on other aspects. Finally, in the third study, the perception of the femininity or masculinity of the adjectives associated with the words also draws on other concepts, which are directly associated with the basic concept to greater or lesser degrees. This measurement is based on the stereotypes pervading society which potentially encourage people to associate feminine words with feminine characteristics and masculine words with masculine characteristics. It also depends on the evaluation of the masculinity or femininity of the adjectives chosen, made by external persons.

If we go back to the metaphor of the plane we introduced above, these measurements would correspond to the plane's distant signals, which have a lower intensity in the sky and a less clear course. The quality of this signal, or in other words, its validity, could be questioned more easily than that of measurements such as reading time or the actions performed following a set of instructions. This illustrates the fact that the more abstract a concept, the more difficult it becomes to operationalize, and the more its measurements can become a topic for discussion. In cases like these, it would be appropriate to choose different ways of operationalizing the concept and to check that the results are consistent between the measurements, as has been done by the authors of these studies.

In addition to being valid, the measurement must be reliable. The *reliability of a measurement* denotes the fact that it always produces the same or almost the same result under the same conditions. For example, your scale gives a reliable measurement if, when you weigh yourself several times in the same day, the result is the same. When working with measurements, such as those used in an experimental linguistics study, things can be a bit more complicated. Indeed, these measurements are dependent on many factors, such as the fact that they are carried out on different people and that the conditions are impossible to keep completely constant. Thus, if we take the example of a study aiming to measure foreign language skills using standardized tests to assess production and comprehension skills, it is unlikely that the participants' responses will be exactly the same if the tests are carried out several times. However, these responses should be similar enough so as to ensure the reliability of the measurement. There are different ways to assess the reliability of a measurement, but their presentation is beyond the scope of this chapter. Those interested can turn to the resources listed at the end of the chapter.

The validity and reliability of a measurement are two distinct concepts, and the fact that a measurement may be reliable does not necessarily make it valid. Let us take the example of the scale again. If you weigh yourself several times a few minutes apart and every time the scale tells you the same weight, which you know is yours, then this can be considered as reliable and valid. If it indicates a weight twenty pounds higher than yours every time, the measurement is reliable, but not valid. A scale indicating your weight with +/- one pound every time would be valid but not completely reliable. Finally, if the result was different every time and far from your weight, the measurement would be neither valid nor reliable. When choosing a measurement, whenever possible, you should try to find a measurement that is both the most valid and the most reliable one.

Finally, we should point out that the validity and reliability of measurements will greatly influence the overall validity and reliability of the experiment. As a matter of fact, the internal validity of an experiment partly depends on its ability to measure the variables, in order to be able to draw solid conclusions on the relationships between them. Likewise, its reliability depends on that of the measurements. In order to be replicated, an experiment requires reliable measurements that assess the phenomenon we want to observe in a consistent manner.

2.7. Choosing the modalities of independent variables

When the different variables have been operationalized, we still have to define the modalities of the independent variable, that is, we have to determine the conditions in which the participants will be included. These conditions must make it possible to clearly evaluate the influence of the independent variable on the dependent variable. For this, every experiment should at least compare two conditions: one in which the independent variable is present and one in which it is absent. For example, we could decide to compare the linguistic competences of children with language impairments with children without language impairments, or high and low level learners.

There are different general ways to build conditions. In a between-subject design, each participant only takes part in one experimental condition (or modality). In a within-subject design, each participant takes part in all the experimental conditions. Between-subject designs must be used for evaluating the influence of an independent variable that cannot be manipulated, for example, being monolingual or bilingual. When the independent variable can be manipulated, it is possible for the same person to take part in all the conditions, or in only one condition. For instance, going back to the examples of studies on linguistic relativity, it is possible to ask a participant to take part only in the word-first name congruent condition, or in both conditions (congruent and incongruent). In the first case, we could observe whether the participants in the congruent condition recall more pairs than the participants in the incongruent condition. In the second case, we could compare the numbers of pairs recalled between the two conditions for all participants. We will return to these different designs, as well as their advantages and disadvantages, in Chapter 6, which will describe the practical aspects of an experiment.

As to the choice of the independent variable modalities, it is often less simple than it seems at first glance to construct a condition in which the independent variable is present and one in which it is absent. To illustrate this difficulty, we will begin by an example outside the field of linguistics, namely the question of drug effectiveness. To test its effectiveness, the drug should be given to one group of patients, not given to another group of patients, and the condition of the two groups should be compared after some time. If the experimental group, who took the drug, reports a decrease in symptoms greater than the reference group, who took nothing, then we can conclude that the drug is effective. However, this conclusion would be wrong in the presence of the so-called placebo effect. This effect reflects the notion that the mere act of taking a pill leads certain people to believe that their condition will improve, something which can actually have an impact on their general condition. For this reason, the results of the experimental group should be compared with those of a control group, made up of people who do not take the medicine, but a pill looking exactly like it except that it lacks the active substance. By comparing the results of the different groups, we can thus show the real effect of the substance (the experimental group vs. the control group) and the placebo effect (the control group vs. the reference group).

In experimental linguistics, an effect similar to the placebo effect could be problematic in an experiment aimed at determining the effectiveness of a language teaching method, for example. Let us imagine the hypothesis that the positive feedback from a teacher improves learners’ skills. Comparing only one condition including a positive comment with a condition where there is no comment would not lead to reliable conclusions. Indeed, if the results of the group receiving positive comments are better than those of the other group, this could be due to the simple presence of a comment. For this experiment to be valid, a group receiving another type of comment (e.g. a neutral or a negative one) should then be added. The results of this group should then be compared with those of the experimental group. This would make it possible to differentiate the effect of the presence of a comment (neutral comment vs. no comment) from that of a positive comment (positive comment vs. neutral comment).

Certain research questions also require the presence of more than two modalities for testing the independent variable. For example, in order to investigate the effect of age on certain linguistic competences, it might seem appropriate to examine more than two age categories, in order to offer a complete vision of the phenomenon. This is illustrated by an experiment in which Zufferey and Gygax (2020) studied the knowledge of connectives such as *aussi* (which roughly corresponds to *therefore*) and *en outré* (which roughly corresponds to *in addition*) in French-speaking adults. As connectives may differ on many aspects (their preferential use in the spoken or written discourse, their frequency, the type of coherence relation they encode), a choice was made of four connectives mainly used in spoken speech and four connectives mainly used in the written modality, also differing on other aspects. The connectives were inserted into sentences correctly or incorrectly. Participants had to judge whether each sentence was correct or incorrect. Using several connectives made it possible to highlight certain variables which influence the mastery of connectives, something which would not have been possible in an experiment using only one type of connective.

In the above-mentioned examples, we can see that there is no simple answer to the question about the number or type of modalities to be chosen for an independent variable. The choice strongly depends on the research question and the conclusions that the researcher wants to draw. The following chapters, devoted to studies in language production (Chapter 3) and language comprehension (Chapters 4 and 5), will describe research related to different fields in experimental linguistics. These chapters will offer additional illustrations of choices related to the operationalization of different research questions, as well as the characterization of the experimental conditions for the chosen variables.

2.8. Identifying and controlling external and confounding variables

From the examples discussed so far, we can conclude that, when choosing the modalities of the independent variable, it is not only necessary to build a condition in which the variable is present and another in which it is absent, but that these conditions should be comparable in all other respects. If the two conditions differ, it is not possible to draw conclusions on the effect of the independent variable. When reflecting on the operationalization of variables, it is therefore essential to think about external variables to be taken into account during the construction of the experimental design. External variables are the variables which can influence the results but are not directly investigated in an experiment. Let us take an example studying reading time so as to investigate the effect of a variable. In this case, the reading time should vary according to the modality of the independent variable. But the reading time will certainly also be influenced by other variables, such as the participants' reading habits, their reading speed, their personal reaction to the variable under study, or even the characteristics of the items (word length and word frequency, or sentence complexity, for instance), and those related to the act of conducting the experiment (time of day, place, etc.).

All of these external variables add "noise" to the dependent variable. This means that the measurement does not only depend on the influence of the independent variable, but also on that of all the external variables. One of the ways to minimize the impact of these external variables on the dependent variable is to test a large number of people using a large number of items. By doing this, the measurement portion associated with noise will decrease, since the external variables generally have a random influence on the measurement. This influence could be high for one trial or one participant, and weak for another trial or another participant. By testing many participants and many items, the noise portion in the measurement should therefore tend towards zero. On the other hand, the measurement portion associated with the independent variable should be maximized, since the effect of this variable should be the same for every trial and every participant.

Another way to minimize the effect of external variables is to implement, whenever possible, a within-subject design. As we said above, in this type of design, every participant takes part in all the conditions. Since the external variables associated with each participant (reading speed, reaction to the independent variable, for example) remain constant from condition to condition, in principle, the comparison of the measurement between conditions should provide a precise indication of the effect of the independent variable. Similarly, whenever possible, a within-item design should also be implemented. In such a design, every item is presented under the different conditions, in order to minimize the impact of external variables related to the items themselves. We will return in detail to the means for building such designs in Chapter 6.

When it is not possible to test every item or every participant under the different conditions, it may be useful to control the external variables in other ways. A first possibility would be to randomly choose and include the participants in the experiment's different conditions. By doing this, we acknowledge the principle that happenstance does things properly and that there is a good chance that the different modalities of the external variables will be evenly distributed in the conditions of the experiment. However, this solution is not suitable for experiments using a limited number of participants.

A second possibility would be to keep the external variable at a constant level, by choosing a modality of the external variable and only testing people or items corresponding to this modality. For example, we could decide to test only people of the same age, of a similar educational level or to take into account only low frequency words or sentences with the same complexity. However, in this case, the external validity of the study might be threatened, since the results cannot be generalized to other groups of people or to other types of items.

Another possibility would be to gather groups of participants in which all the modalities of the external variables are represented. For example, we could include as many women as men, or as many low and high frequency words for every modality of the independent variable examined. This possibility might solve the problems of external validity raised above but could complicate data collection, depending on the number of external variables to be controlled. It is also practically impossible to control external variables such as reading speed, the level of involvement of the participants in the study or their reaction to the independent variable, because the participants belonging to the different modalities of these variables can only be known while or after conducting the experiment.

To conclude, we can note that these different possibilities can be combined in the same experiment, in order to control the effect of several external variables. In general, a perfect command of external variables is unattainable, since these variables are multiple and varied. Therefore, researchers generally only focus on the most relevant external variables for a specific experiment. However, there is a type of external variables that cannot be neglected, namely the confounding variables.

Confounding variables are variables whose levels vary systematically along with the levels of the independent variable. Due to their systematic variation along with the independent variable, confounding variables offer an alternative explanation to the results found in the project and can thereby threaten the internal validity of the experiment.

Let us take a few examples to illustrate the problem of confounding variables. Let us imagine that an experiment has shown that people take it longer to read infrequent words than frequent ones. Let us admit that in this experiment, the participants had to read either frequent words or infrequent words. The words appeared one by one on a computer screen and the participants had to press a key after having read each word, which recorded the reading time. Frequent words were: *body*, *hotel*, *husband*, *water*, *paper*, *table*. Less frequent words were: *abdomen*, *clarinet*, *eloquence*, *manuscript*, *obelisk*, *rosemary*. By examining the words used in the experiment, we quickly perceive that these words differ not only in terms of their frequency of appearance in the language, but also as regards their length, and that this occurs systematically. In other words, infrequent items are longer (three syllables) than frequent items (two syllables). The result could therefore just as easily stem from the fact that longer words take longer to read. The existence of this confounding variable makes it impossible to draw a conclusion as to the relationship between frequency and reading time. In order to overcome this problem, items of similar length should have been chosen when considering high and low frequency conditions.

Imagine another experiment in which we are interested in the role played by practicing a language in tandem, so as to speed up the learning process. Let us admit that, in this study, we decide to recruit learners who take part in exchanges with other learners one evening per week, and a group of learners who do not take part in any activity of this type. These two groups of learners are then compared using language proficiency indicators. Here, a confounding variable could be the motivation to learn a language. It is indeed very likely that people who invest their own time in an additional activity for studying a foreign language are more motivated to learn it. This characteristic will probably have effects on their foreign language proficiency. One way of avoiding the confounding variable would have been to manipulate the independent variable by only choosing participants of the same level who do not practice the language outside the classroom, and then to ask half of them to make tandems. In this way, we could not say that some people are intrinsically more motivated to study than others.

In the examples above, we can see that a confounding variable is more likely to appear in the experiments having between-subject or between-item designs. In these designs, different participants or items are included in the different conditions. As a consequence, there is a higher risk that an additional variable plays a role in the results than in within-subject or within-item designs, where the participants and the items are included in all the conditions. Likewise, a confounding variable is more likely to appear in a quasi-experiment in which the independent variable cannot be manipulated by researchers and is inherent to the participants or the items. When we build conditions on the basis of a variable that cannot be manipulated, the groups have a significant probability of systematically differing on other aspects than the one examined. Therefore it is essential to think about the different designs possible for an experiment. When possible, variables should be manipulated instead of simply observed, and repeated measurements should be used.

2.9. Conclusion

In this chapter, we first saw that a good experiment must be valid and reliable. The validity of an experiment is based on two main aspects: internal validity and external validity. From the point of view of internal validity, an experiment should lead to a clear conclusion concerning the influence of an independent variable on a dependent variable. In other words, the changes observed on the dependent variable should only stem from the manipulation of the independent variable. From the point of view of external validity, the conclusions observed at the sample level should be generalizable beyond the specific conditions of the experiment. We have also seen that an experiment must be reliable, that is, it should lead to similar results if it is conducted several times.

Next, we defined the different variables involved in an experiment and described four types of scales used for measuring them. We saw that the different types of scales do not have the same properties and therefore do not support the same analyses. Finally, we presented the steps involved in the operationalization of a research hypothesis. Firstly, we have to choose a valid and reliable measure for quantifying the variables. Secondly, the modalities of the independent variable must be defined in order to make it easier to reach a clear conclusion as to the effect of this variable. Finally, we discussed the control of those external variables which possibly influence the results of an experiment, as well as the importance of identifying any confounding variables that could jeopardize the conclusions drawn from a study.

2.10. Revision questions and answer key

2.10.1. *Questions*

1) Identify the independent and dependent variables for the following hypotheses:

a) bilingual children have better math skills and a better ability to learn a new language than monolingual children;

b) mastery in the use of connectives depends on their frequency in language and the reading habits of speakers.

2) What type of scale (nominal, ordinal, interval, ratio) correspond to the different variables below?

a) The time required for fixating words, measured using a device that records eye movements.

b) Each participant's mother tongue.

c) Each participant's year of birth, from 1990 to 2000.

d) Agreement with a statement, measured on a scale with the following options: strongly disagree, somewhat agree, strongly agree.

3) List different ways of operationalizing the following question, specifying the measurements used for the different variables and the conditions chosen for the independent variable: does a person's empathy level (ability to understand the emotions of others) have an influence on understanding emotions when reading?

4) What type of validity is threatened in the following studies? For what reasons?

a) An experiment carried out on bilingual university students has shown that the comprehension of anaphora depends on verbal working memory capacities.

b) A study has shown that bilingual people change their personality depending on the language used. In their mother tongue, people were described by their friends as being more extroverted and communicative than in their second language.

5) Which external variables should be controlled to investigate the following hypothesis: sentences conveying an emotional content are read more slowly than neutral sentences?

6) Identify the confounding variables that may be involved in the following study: an experiment has investigated the influence of private lessons on the reading skills of deaf children and children without any hearing loss. To do this, the two groups of children had to read a half-page text and then answer questions about it. They then benefited from private lessons for two months, after which they had to read a one-page text and answer questions. The results showed that private lessons did not reveal significant benefits for children. Children without any hearing loss provided the same number of correct answers, whereas deaf children gave fewer correct answers on the second test than on the first one.

2.10.2. *Answer key*

1) a) The independent variable is the number of languages spoken by children (one vs. two). This hypothesis has two dependent variables, math skills and the ability to learn a new language.

b) In this case, there are two independent variables. The first corresponds to connective frequency in the language and the second to reading habits. The dependent variable is the mastery of connectives.

2) a) Word fixation time is a quantitative variable measured on a ratio scale. The data acquired on this scale can range from zero to several thousand milliseconds. It is also possible to rank the fixation times and to apply arithmetic operations to them.

b) Each participant's mother tongue corresponds to a qualitative variable, measured on a nominal scale (e.g. French, Italian, German).

c) Each participant's birth year is a quantitative variable measured on an interval scale. It is possible to rank the participants and to find out their age difference. However, it would be inappropriate to apply other operations, such as multiplication or division, to birth years.

d) Agreement on a scale with four options is a qualitative variable, measured on an ordinal scale. The answers can be ordered, but the gap (difference in size) between the options cannot be guessed.

3) In order to operationalize the question, it is necessary to identify the variables and then choose an objective measurement for these variables.

Here, the independent variable corresponds to the empathy level of the participants, whereas the dependent variable corresponds to the understanding of emotions while reading. In order to measure the empathy level, one can turn to a standardized test for measuring this ability, for example, the *Interpersonal Reactivity Index* (Davis 1980) or the *Empathetic Quotient* (Baron-Cohen and Wheelwright 2004). On the basis of the score obtained in one of these questionnaires, it would be possible to classify the participants into two groups. There are different possibilities for quantifying the understanding of emotions while reading. A first solution could be to make a presentation of short excerpts describing emotions and then to ask the participants to name the emotion of the characters, and then count the number of correct answers. Another possibility would be to turn to online measurements (see Chapter 5) to evaluate the derivation of emotional inferences during reading. To do this, we could present the participants with short excerpts, again in which the character feels an emotion, and then to measure the reading time of target sentences displaying the emotion. Of course, there are other ways of studying this question, using the different methods presented in Chapters 4 and 5.

4) a) The validity of an experiment depends on the possibility of drawing reliable conclusions concerning the relationship between variables under study (internal validity), as well as on the possibility of generalizing the results beyond the method, the participants and the items examined in the experiment (external validity). In the first case, the external validity of the study would be limited, as only female students were tested. It would not be possible to say that, in general, the understanding of anaphora depends on verbal working memory capacities; this conclusion should be limited to the population the sample of participants comes from.

b) In this case, the study aimed at evaluating changes in the personality of people speaking a foreign language. Personality aspects were assessed by friends of the participants, who probably had to agree with statements such as “This person is more extroverted when they speak their mother tongue than when they speak another language.” The internal validity of this study is compromised, since the validity of the measurement employed can be called into question. It would have been appropriate to choose a more objective personality measurement indicator or to use other personality assessment tools for confirming the results.

5) The dependent variable of this hypothesis corresponds to the time spent reading sentences (reading time), whereas the independent variable corresponds to whether such sentences convey emotional content or not. In

order to isolate the effect of the independent variable on the dependent variable, it is essential to create conditions in which the sentences differ only in terms of emotional content and not on other variables that may influence the reading time. At the item level, the variables to control are typically sentence length, their syntactic complexity and the frequency of the words used. At the participant level, their general reading speed or their comprehension skills can also influence the dependent variable. It would be appropriate to set up a design with repeated measurements, in which each person would take part in all the conditions in order to keep the external variables related to the participants at a constant level. It could also be interesting to measure the general competences associated with the participants' understanding of emotions (e.g. empathy levels) in order to see whether and how they influence the reading times in the different conditions.

6) Different points make the conclusions of this study questionable. First, children's comprehension was operationalized as the number of correct answers given to questions that had to be answered in writing. By asking the children to respond in writing, an additional variable comes into play in the experiment, namely the children's writing skills. It was therefore not only the comprehension skills that were measured, but also the children's skills for writing down their understanding. Second, the tests performed two months apart were different. While the first text was half a page long, the second text was one page long, twice as long than the first. This introduced an additional variable into the experiment, namely the memory capacities of children. These are likely to play a more significant role during a test on a one-page excerpt than on a half-page excerpt.

2.11. Further reading

For more detailed explanations on the different types of measurement and the choices to be made during operationalization, we recommend Chapter 2 of Field and Hole (2003). For more details on the different types of validity and reliability, for experiments or measurements, we refer readers to Chapters 5 and 6 of Price *et al.* (2013). This book illustrates, with simple examples, the different validities involved in research, as well as their influence on each other. For more information on questionnaires, it is possible to turn to Rasinger (2010) and Wagner (2015), among others.

3

Studying Linguistic Productions

This chapter presents various methods for studying the linguistic productions of speakers, namely through elicitation and repetition tasks. We start by discussing the differences that exist between the ability to produce and to understand language; and we argue that it is necessary to examine these two components of the language faculty separately, in order to have an overall picture of the functioning of a certain linguistic phenomenon. We then present the fundamental methodological differences which separate the observation of linguistic productions in a corpus and the experiments aimed at eliciting such productions. In the rest of the chapter, we introduce the different methods used for generating productions in an experimental context. We start with so-called free elicitation tasks, which imply a minimum level of constraint on productions. Then, we move on to constrained elicitation tasks and finally to repetition tasks, which imply an even greater control over production. In every case, we discuss the possibilities that these tasks offer for the study of language, as well as their limitations. We arrive at the conclusion that these tasks are complementary and that the most reliable method for studying linguistic productions is to combine them.

3.1. Differences between language comprehension and language production

Mastering language involves being able to use it appropriately for communicating with others, as well as decoding and interpreting discourse (spoken and written) produced by others. These two skills, respectively, involve the ability to produce and to understand language. As we shall see in

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

this section, these two elements of the language faculty are nonetheless partially dissociated and should be studied separately in order to obtain an overall picture of the speakers' linguistic competence.

The dissociation between language production and comprehension abilities is particularly evident during the language acquisition period in the first years of life. Indeed, between birth and the age of 1 year, infants do not really produce language. During their very first months, babies only cry. Then, when they reach 2–4 months, children start producing vowels like "aaaa" with different intonations. It is still necessary to wait until between 6 and 9 months of age for the so-called babbling stage to begin. This period is characterized by the repetition of syllables like "da-da-da" or "goo-goo", which reproduce certain features of their mother tongue. Finally, it is only around their first birthday that babies produce a few isolated words like "bye-bye" and "no".

Observing babies' productions during their first year of life could give the impression that no aspect of language is mastered. However, this is far from being true, as shown by experimental techniques which make it possible to indirectly measure language comprehension in babies. One of these techniques, non-nutritive sucking, consists of measuring differences in suction intensity and rhythm by means of a teat containing sensors. It has shown that babies are already sensitive to many aspects of language before they can speak because they react systematically to changes in stimuli, as revealed by the differences in the intensity and rhythm of their sucking. To quote only a few examples, from birth babies are able to distinguish their mother tongue from other very different languages (e.g. French and Chinese) and can perceive phonetic contrasts, even in languages that are not their own native language. Between the age of 4 and 6 months, babies recognize the differences between even very close languages (e.g. German and Dutch) and can already understand a few isolated words. When they reach 1 year of age, they are able to recognize words heard several weeks earlier, in stories, and detect violations in the word order of their mother tongue. By the time they finally start producing a few words, babies have already developed sophisticated comprehension skills in their mother tongue (for a more in-depth discussion of these early skills, see Rowland (2013)).

The dissociation between language comprehension and language production persists throughout the language acquisition period. In most cases, children understand more than they are able to produce, but the reverse asymmetry also occurs. In particular, the first productions of a word do not imply that children really understand its meaning. For example, children pointing at the dog in their home using the word “dog” give the impression that they understand the meaning of the word. However, children go through a phase known as underextension, during which they assign a linguistic label not to a category (all the dogs in the world) but to a specific referent (the dog in their house). During this period, they do not yet master the meaning of this word, even if their productions are technically correct. These examples illustrate the need to study not only children’s linguistic productions, but also their comprehension of language.

However, the dissociation between comprehension and production is not the prerogative of young children during the language acquisition period. It is also found in foreign language learners, both children and adults. For this reason, so-called receptive and productive language skills are evaluated separately in foreign language assessment tests. In the same way as young children, foreign language learners often have better receptive skills (also called passive skills) than production skills. Production is also sometimes ahead of comprehension in the interlanguage of learners (Ortega 2008).

Another example showing the need to dissociate comprehension and production is that of those suffering from language impairments. As a matter of fact, some aphasia types such as Broca’s aphasia (also called expressive aphasia) primarily affect the productive aspect of language, whereas others such as Wernicke’s aphasia (also called receptive aphasia) primarily trigger problems in language comprehension. The study of those suffering from aphasia also illustrates the fact that language production, like comprehension, can be used for analyzing the different components of language. For example, the inability to carry out a very simple lexical production task, such as naming an object represented in an image can have various causes: problems accessing conceptual information about this object (the object is no longer recognized), an inability to access the phonemes that make up its name (a problem that healthy people also encounter when they have a word on the tip of their tongue) or even a motor inability to pronounce these phonemes.

Finally, note that the difference between language production and comprehension is also present in adults who do not suffer from any language impairments and who speak in their mother tongue. There is notably a big difference between production lexicon, for example, the words used every day for speaking to somebody we know, for writing letters or for teaching a course, and comprehension lexicon, which corresponds to the number of words we can actually understand. Again, production is clearly below comprehension. For example, in his novel *Madame Bovary*, Flaubert used no more than 14,000 different words (or word types) even when counting the conjugated forms of verbs, plurals, etc., separately, and just over 7,500 words if we only count semantically different words (lemmatized forms). These relatively low numbers might suggest that if the lexicon of a great author is no greater than 10,000–15,000 words, then the lexicon of an average person should be much smaller. But once again, it is necessary to differentiate between the production lexicon, that is, the words that people have the opportunity to use, and those that they are able to understand. In fact, the comprehension lexicon contains at least 40,000 words for someone with a high school diploma and may amount to 60,000–80,000 words for speakers with a college education (Aitchison 2003).

All the examples discussed in this section confirm that language production and language comprehension are clearly dissociated and that these two aspects of linguistic ability should be studied separately, in order to have an overall picture of the linguistic competence of speakers. However, for several decades, only language comprehension was considered an adequate reflection of linguistic competencies. This exclusion of the productive aspect as a component of the language faculty finds its origins in the works of the American linguist Noam Chomsky, and in his definition of I-language. According to Chomsky, linguistics has the task of studying the linguistic representations of speakers, which he denominates the *I-language* or *internal language* (see, in particular, Smith (2004) for an introduction to Chomsky's thought). These representations reflect what people intuitively know about their mother tongue, in other words, what they understand. Contrary to this, according to Chomsky, linguistic productions do not represent competences but are barely performances or implementations of the language faculty. However, the latter are not always representative of competence. Indeed, a person may make mistakes when speaking, for example, by using a word in the place of another, not because he or she does not know the word's meaning but due to tiredness, stress, etc. Therefore, according to Chomsky, studying linguistic productions offers a biased reflection of the internal language, which should be the only study object for linguists. Recently, the study of linguistic productions has returned to the heart of linguistic research, thanks, in particular, to the development of corpus linguistics (see section 3.2). From an experimental point of view, Chomsky's objections can be avoided by using quantitative methods, which make it possible to sort isolated occurrences, which are not representative of linguistic competence, from recurring facts. For example, if a person produces the form "*he goed*" 10 times in a 30-minute interview and never the correct form "*he went*", it seems unlikely that these productions are random errors but rather that they reflect the fact that the person does not know the irregular form of this verb.

In summary, the study of language can either relate to the aspect of production or to that of comprehension, but we should keep in mind that the results in one of these areas cannot be generalized to the other. In the following chapters, we present different types of experiments aimed at measuring language comprehension, as this can be done through many experimental paradigms. In this chapter, we focus on the production component and review the pros and cons of the different methods for studying it.

3.2. Corpora and experiments as tools for studying production

Different empirical methods can be used for studying linguistic production. A first important distinction between these methods, which we will study in this section, is the one that separates the observation of corpus linguistics productions from the elicitation of productions within an experimental context.

Corpora are large collections of texts or recordings gathered in an electronic format so as to be representative of a certain type of language. For example, some corpora aim to represent a discourse genre (journalistic, literary corpus or online discussions), types of speakers (adults vs. children, learners vs. native speakers) or linguistic regions (the UK, the US, Australia, etc.). Whatever the type of corpus considered, corpus linguistics aims to study natural linguistic productions from a quantitative perspective. For example, a corpus study could be used for studying the differences in pronunciation of English vowels between speakers from London and New York. Or another study could compare the development in the production lexicon of neurotypical children and children with autism spectrum disorder (ASD), at the same chronological age. The common point between these studies, albeit with totally different themes, is that the language samples produced in the corpora as a study object were collected in their natural context, without any intervention on the part of researchers.

When collecting data from a corpus, the primary aim is to collect spontaneous interactions in the same way that they might have occurred in the absence of a recording. It is not always possible to reproduce entirely natural conditions, because the simple fact of recording the participants can cause them to unconsciously change their behavior, but the goal is to get as close as possible to a natural environment. For example, children are usually recorded at home when interacting with family members. This is a big difference with the experimental contexts, in which the participants do not evolve in their natural environment but in the laboratory, or sometimes in their classroom, in the case of children. Thus, one of the main advantages of studying corpora productions in comparison with the experimental context is their natural character, which better reflects the real skills of people than productions collected in a non-familiar context, in the presence of strangers.

Another advantage of using corpora is that, due to their large size, they make it possible to observe a large number of occurrences of a phenomenon, produced by a large number of different people. Conversely, in an experimental context, it is not possible to have more than a limited number of occurrences produced by each participant, in order to avoid tiredness and learning effects. In addition, the number of participants in a study is often limited for practical reasons. However, there are many occurrences which can only be observed in a corpus for frequent linguistic phenomena, for example, the use of basic vocabulary or frequent verbal tenses such as the simple past or the future. For rarer linguistic phenomena, such as the use of a specialized lexicon or the use of infrequent verbal tenses, it is very likely that even a large corpus will not make it possible to find many occurrences. Conversely, in an experimental context, it is possible to encourage participants to produce infrequent elements by constraining the production context. For example, it is possible to ask the participants to continue a sentence which can only be completed by the subjunctive form, or to name objects represented in images that correspond to rare words. This experimental method makes it possible to collect more occurrences of rare phenomena than the use of a corpus.

In addition to testing rare linguistic phenomena, the experimental method also has another advantage compared to the observation of natural phenomena in a corpus. If an element does not appear in a corpus, for example, if children produce no passive sentences, it is not possible to conclude that they do not know the passive form. These children may simply not have had the opportunity to produce passive forms during the recordings, although they are capable of doing so. In other words, the lack of evidence in the corpus of children producing passive sentences does not suffice to conclude that they avoid this form because they cannot master it. On the other hand, in an elicitation context where children are invited to complete the transformed sentence in (1) to keep its meaning, it would no longer be possible to avoid the passive form:

(1) The cat chases the dog.

The dog ___________ by the cat.

Thus, the experimental method makes it possible to determine whether people are capable of producing a specific linguistic form or not, whereas corpora only make it possible to observe whether a certain form is used or not, and how often it is produced. This difference may have a significant impact on the conclusions of a study.

Let us examine an example that illustrates this problem. Royle and Reising (2019) studied the ability of children with specific language impairment (SLI) and children without language impairment – matched on age or on the mean length of utterance (MLU) – to produce correct agreements between the elements within noun phrases, both in the context of natural observation recorded in a corpus, and during an elicitation task. In the elicitation task, children had to make a puzzle and name the pieces. This task was designed to elicit the production of complex noun phrases, combined with adjectives (“the little house”, “the big blue house”, etc.). The same children were recorded during natural interactions in the context of play. The results showed different errors under the two conditions. During spontaneous interactions, children with specific language impairment (SLI) essentially omitted elements of the noun phrases, such as determiners. The elicitation task, on the other hand, revealed specific difficulties in adjective agreement. A generally high level of agreement errors was also found. This difference reflects the fact that children tend not to produce adjectives or, more generally, complex noun phrases in spontaneous speech. Thus, the elicitation task made it possible to reveal linguistic difficulties in children with SLI which were not apparent during natural interactions.

Another inherent limitation in corpus linguistics is that the identification of some linguistic phenomena in a corpus requires manual processing of data, which is very time-consuming. As a matter of fact, only words can be searched automatically in a corpus, and these searches must be refined to eliminate the irrelevant occurrences of homonyms. For example, let us imagine a study aiming to identify all the uses of relative sentences in a corpus. One idea might be to look for all the occurrences of relative pronouns, such as “who” or “which”. However, this research would not be enough to identify the relevant occurrences, since these words are also used as interrogative pronouns. It would therefore be necessary to sort all the search results manually and stick to the relevant occurrences. Imagine a search aiming to determine the different ways in which requests are formulated. This time, a word search would be of little use, as there is no conventional link between the form and function of speech acts. To summarize, research on corpora may become very complex and time-consuming in cases where the phenomenon investigated is not associated with an unambiguous linguistic form that can be automatically queried. Experimental research makes it possible to circumvent these problems by formulating a task encouraging participants to specifically produce the element under study.

Furthermore, the meaning that a person tried to convey during the discussions recorded in a corpus may be ambiguous. For example, when a young child uses a name designating an object which is not present in the immediate context, as, for example, the word “island”, it is difficult to know whether the word is being used to convey the appropriate concept or not. This problem is all the more important since children regularly produce underextensions and overextensions of meaning, as we have already pointed out (on this, see, for example, Bloom (2000)). The intended meaning in elicitation tasks which involve image description does not pose the same ambiguity problems.

Finally, because of the ambiguity inherent in corpus data, there is the question of how many spontaneous occurrences make it possible to conclude, with absolute certainty, that an element has been acquired. After the first occurrence? After three occurrences during the same recording? Elicitation experiments allow more precise control over the production context and thus make it possible to determine, with certainty, whether a person is capable of producing a certain linguistic form repeatedly or not. On the other hand, these experiments imply additional difficulties for the participants, such as the need to understand the task and to carry it out in an unnatural production context. Due to these additional difficulties, elicitation tasks generally indicate a lower level of competence compared to the observation of productions in a corpus, and thus provide a conservative image of the linguistic level of the participants.

In sum, the study of linguistic productions can be done either through the observation of a corpus or by experimentally eliciting productions thanks to the use of specially designed tasks. Both methods have advantages and disadvantages which we have discussed in this section. We should also observe that in many cases, these two approaches can provide complementary points of view, which are very useful. For example, before deciding to create an elicitation task, the frequency of a certain linguistic phenomenon in different contexts or discourse genres can be assessed by means of a corpus.

In the rest of this chapter, we will focus on the presentation of different elicitation tasks aimed at experimentally eliciting linguistic productions. We will see that these tasks are placed on a continuum, spanning from a very low level of control, in the case of free elicitation tasks, to a higher level of control, in the case of constrained elicitation tasks, reaching a maximum level of control in repetition tasks.

3.3. Free elicitation tasks

In order to overcome some limitations associated with the natural observation of data in a corpus, and to complete their production database, researchers sometimes resort to the free elicitation technique, which consists of orienting the productions by placing the participants in a previously set context. These elicitation tasks often take the form of interactive games for obtaining certain dialogue-related elements, or the description of films, images or the retelling of memories, in order to collect monologues.

The great advantage of this technique is that it makes it possible to preserve the naturalness of corpus data to a large extent, since the participants are free to produce language samples without the intervention of the researchers in charge of data collection. Unlike the observation of corpora discussed previously, this method involves a form of experimentation in that it makes it possible to manipulate the production contexts in order to study their influence on the type of linguistic productions. This technique offers many advantages.

First of all, it makes it possible to generate linguistic elements which are rarely found in corpora and whose low frequency hinders a quantitative analysis of data. For example, by asking people to describe events taking place in a video, it is possible to test their ability to retell a series of events and to study the use of verbal tenses, for instance. Some studies have used Charlie Chaplin's silent films as stimuli for eliciting production. Many studies (e.g. Berman and Slobin (1994)) used a story without any text captions in the form of a 24-vignette series called *Frog, Where Are You?* (Mayer 1969) for elicitation tasks with children and learners. This story has become a classic of elicitation studies and data exists in many different languages, available via the CHILDES online database (MacWhinney 2000). In addition to films and stories without text, another medium to encourage the production of nouns for designating specific objects is to have participants play with objects or cards containing images of such objects. For example, participants may be instructed to describe these cards with sufficient precision so as to allow someone else to identify the correct card. By using a card deck representing similar objects, for example, a red car and a blue car, it is possible to test the production of complex nominal phrases.

Another advantage of free elicitation over corpus observations is that it is easier to control the meaning the speaker intended to produce. The observation of corpora, where the context is not controlled, leaves ample room for interpretation. On the other hand, when an object is represented on a card, the target word is clearly identified and naming mistakes can also be easily identified.

At a syntactic level, free elicitation tasks do not always enable participants to produce the structures targeted in the study. In fact, there are often several ways of freely expressing the same proposition, and the avoidance strategies we mentioned in relation to corpus data may also appear in free elicitation tasks. In order to specifically test certain complex syntactic structures, which tend to be avoided in everyday spoken language, such as subordinate clauses or passive constructions, constrained elicitation tasks seem to be better suited, as we will discuss in further detail in section 3.4.

Furthermore, free elicitation tasks do not provide representative data on the actual production frequency of certain words or syntactic structures. This is why it is preferable to use them for supplementing, rather than for replacing the analyses of spontaneous productions in a corpus, as Evers-Vermeul and Sanders (2011) did when studying the productions of subjective causal relations in young Dutch children. In the literature, different types of subjective causal relations are often separated into two sub categories. An important distinction separates the relations involving speech acts, as in (2), and so-called epistemic relations, which imply arguments and conclusions derived from them, as in (3):

(2) Lend me your umbrella, because I lost mine.

(3) Perhaps it will rain, since everyone has taken their umbrella.

Evers-Vermeul and Sanders (2011) wanted to find the order in which young Dutch children begin to produce these two types of causal relations. The children took part in a free elicitation task in which they had to pick either a character from among many on printed cards, and then convince a doll that they had made the right choice, or give instructions to the doll so that it placed stickers in certain places on a picture. While the first task provided a context favoring the production of epistemic relations, the second focused on a context encouraging the production of relations involving a speech act (giving orders to the doll). The experiment was carried out with children of two age groups, the first of about 4 and a half years old and the second, with a group of 6-year-olds. The results showed that children in both age groups were able to produce both types of causal relations. In addition, contextual manipulation biased the type of production, since the children systematically produced more epistemic-type relations in the argumentation task and more speech act relations in the directive task.

This experiment indicates that context plays an important role in the production of causal relations and this should therefore be taken into account. On the other hand, it does not provide enough information about the age from which children are able to produce these two types of causal relations, as even the youngest children were able to produce both types. In order to answer the question of when these productions begin, it was necessary to complete the experiment with an analysis of children's spontaneous productions in a corpus. This analysis revealed that there is a difference in the age of the first productions, since children first produce relations involving speech acts, on average, several months before producing epistemic relations. On the other hand, in the corpus analysis, the important role of context on the type of relation produced could not be established. The elicitation task therefore provided an important additional element for understanding when and how children produce different types of causal relations. This study thus illustrates the advantages of combining corpus data and a free elicitation task.

Another great advantage of free elicitation tasks is the low level of linguistic constraint they impose, which makes them easy to implement in different languages. These tasks therefore make it possible to reveal the impact of differences in encoding between languages on the way people speak about the same events. For example, von Stutterheim and Nüse (2003) compared the way in which English and German speakers and narrators retell the events of a short, seven-minute silent film, while it is being played. The study showed that English speakers divided the action of the film into many very specific events, whereas German speakers divided the sequences into fewer events, of a more global nature. Thus, for the same linguistic input, the division into events seems to be done differently between the two languages. In addition, many differences were also observed in the description of the same event. While in English, verbs alone were very common (he falls, he jumps, etc.); in German, the point of arrival or the direction of the action was mentioned more often (he jumps from the cliff, he falls to the ground, etc.). This elicitation task thus made it possible to show that people speaking close languages, such as German and English, divide the flow of visual information according to different criteria, on the basis of varying elements in the events of their story, and that they provide different time perspectives for such events. These conclusions could be drawn thanks to the possibility offered by free elicitation in collecting natural language under similar conditions of production, between different languages. Thus, free elicitation is a technique particularly suitable for collecting data at the discourse level, as participants can choose how to order their stories by themselves.

To summarize, free elicitation tasks make it possible to generate the production of infrequent elements, to test fine semantic distinctions, as well as to assess the lexical or syntactic level of productivity between different groups of participants, of varying levels, ages or languages. These tasks can either involve identical production conditions, or manipulate the production context in order to study the role of the different contexts on speakers' productions. Free elicitation tasks are also not very repetitive and can be used on several occasions without producing a training effect or tiredness. However, their low level of control does not ensure that a specific linguistic structure or vocabulary will be produced.

3.4. Constrained elicitation tasks

The constrained elicitation tasks we describe in this section are used for quantitatively studying the ability of different groups of people to produce certain linguistic elements, while systematically controlling the different factors involved in such productions. These methods involve the use of a specific protocol so as to prevent examples being given of the structure or word in question. Unlike the repetition tasks we will present in the following section, these tasks do not provide a model of the structures to be produced but work only as incentives for producing such structures.

For example, in one of the classic experiments on lexicon acquisition, Berko (1958) showed a picture of a small invented animal to children, telling them that it was a *wug*. She then showed them another image in which two of these animals were drawn, saying to them: “Now there is another one. There are two of them. There are two…”. The children's task was to complete her sentence. Most children aged between 4 and 7 years gave the right answer: *wugs*. In a very ingenious way, this experience made it possible to show that young English-speaking children are already able to use the morphology of their mother tongue in a productive way, with words never encountered before.

Other elicitation experiments aim to test the mental lexicon of adults. For example, these experiments can involve measuring the latency necessary for producing different words. This method has made it possible to show that words constructed morphologically do not always take longer to be produced than words with the same frequency, having the same number of syllables, but not constructed morphologically (Bonin 2013). These experiments show that all morphologically constructed words are not assembled on the spur of the moment, at least when they are derivational suffixes of frequent words.

Many other constrained elicitation experiments relate to the field of syntax, since it is at this linguistic level that such tasks become most interesting. Some experiments produce a certain grammatical form as a prompt, for example a passive sentence, in order to determine whether this form will prompt speakers to use passive constructions spontaneously for describing other events, more often than they would do if there had been no prompt. These experiments made it possible to show that the speakers have a tendency to reproduce recently heard syntactic forms, whether in their mother tongue or in another language. For example, Kim and McDonough (2008) asked university students with Korean as their mother tongue and different levels of English proficiency, to interact in English with an English speaker. During these interactions, participants had to describe a series of cards. The English-speaking person's cards contained 20 sentences with passive verbs and 20 sentences with active verbs. The participants' cards contained only verbs, half of which were the same verbs as those appearing on the cards of the English-speaking participant and the other half were different. The results indicated that learners produce more passive sentences when using a verb that has just been used in the passive voice, confirming their sensitivity to this priming effect, as with native speakers (Branigan *et al.* 1995).

Constrained elicitation experiments are also used for testing the development of many complex syntactic structures (see McDaniel *et al.* (1998) for a review). For example, it is possible to test children's ability to produce certain structures by asking them to transform a sentence into a question. In this type of experiment, however, it is important for the context to be plausible and to help children understand the need to produce the expected form. For example, in some experimental paradigms, children are asked to act as an intermediary between a person and a doll who cannot hear properly. The experimenter stands in a different place in the room, in relation to the child and the doll and asks the child to help her talk with the doll, by asking her questions. An example of a question elicitation task could be as follows:

(4) I don't know whether she likes eating French fries. Ask her.

In order to elicit the production of negative sentences, one possibility would be to ask the children to say the opposite of what the experimenter said. Whatever the format chosen, we should ensure that the instructions provided encourage the participants to produce the expected form. This can be achieved by means of a pretest with older children or with adults.

Indeed, it is possible that participants, children and adults, may prefer an alternative strategy. For example, when participants have to say the opposite of sentence (5), the answer given could imply a lexical opposite (6), rather than the expected negative sentence (7):

(5) Julian is kind.

(6) Julian is mean.

(7) Julian is not kind.

In this case, it would be wrong to conclude that children are not capable of producing negations. This problem is all the more significant since, in an elicitation task, it is absolutely necessary to avoid providing a model of the expected answer, as in the case of repetition tasks. It is therefore not possible to provide a first example of morphological negation. To avoid this type of problem, one solution would be to choose adjectives that have no salient lexical opposite, such as *gifted*, or qualifiers having no antonyms, such as *American*.

In addition to their usefulness for testing the mastery of syntax by children or by non-native speakers, elicitation tasks can also be used for studying discursive and sociolinguistic phenomena in native speakers. For example, Kehler and Rohde (2013) tested the links between the type of coherence relation uniting sentences, such as causality, goal or temporality, and the type of referential expression chosen for designating a discourse referent. To do this, participants had to insert an argument after sentences, as in (8) and (9):

(8) Luke lent Peter a book. He ____________________.

(9) Luke lent Peter a book.______________________.

This experiment enabled them to observe that the presence of a pronoun influenced the choice of coherence relation used for continuing discourse. In fact, when a pronoun was present, the majority of participants chose to continue discourse with a causal relation (“he wanted to read it”) or an elaboration relation (“he often liked the same books as him”). On the other hand, when the pronoun was not imposed, as in (9), and the participants chose to use a full noun phrase (Luke or Peter), the sentence they produced involved a different discourse relation, implying either a result (“Peter loved it”) or a goal (“Peter used it for preparing a presentation”). This elicitation task made it possible to bring to light the constraints which associate the different aspects of textual coherence (referential expressions and coherence relations).

In the field of sociolinguistics, constrained elicitation experiments are useful for determining the dissemination of linguistic traits. This method was used, in particular, by Avanzi *et al.* (2016) for mapping the dissemination areas of lexical and grammatical regionalisms of French spoken in Europe. Thanks to the use of an online questionnaire, data from more than 10,000 French-speaking Europeans were collected. This questionnaire contained a task that represented a form of constrained elicitation. Indeed, for every word tested, the participants read a definition of the word (10) or contextual information (11) associated with an image:

(10) What do you call this object, on which clothes are dried?

(11) In winter, in order to keep our feet warm, we put on our ________?

In the case of words for which several regional variants were documented in dictionaries, participants were asked to choose from a word list. If none of the suggested words seemed to suit them, it was possible to check an "other" box and insert their own word. For other words with a supposedly more general distribution, participants had to indicate the frequency with which they used them on a scale from 0 (never) to 10 (very often). In the case of syntactic expressions, they had to choose, from a closed list, the expression they would use the most spontaneously in such a situation. This method of eliciting production from a closed list of possibilities made it possible to show that certain regionalisms, listed in dictionaries, are used progressively less and less, whereas certain words presented as regionalisms in dictionaries have an area of such wide dissemination that the qualifier of regionalism is no longer appropriate.

In sum, in this section, we have seen that the constrained elicitation method makes it possible to obtain targeted linguistic productions, in order to answer research questions in fields as varied as lexicon, syntax, discourse and sociolinguistics. The main advantage of this method is that the experimental context attached to it ensures a sufficient number of linguistic productions for carrying out a quantitative analysis of data. Furthermore, this method makes it possible to manipulate independent variables and to avoid interference of confounding variables, which sometimes obscure corpus data. However, we should be careful to make sure that the material used for these tasks does not include other factors of complexity, apart from the one being tested (excessively long sentences, rare words, etc.). Its main drawback, which applies to all experimental methods, is the unnatural nature of production contexts, which do not always reflect what people do spontaneously. In the context of sociolinguistic research in particular, the participants might be tempted to answer by following conventions rather than in relation to their actual practices, which often escape consciousness. Furthermore, constrained elicitation requires a certain level of linguistic competence, making this method inapplicable to children younger than 3 years (Eisenbeiss 2010). In general, children and learners obtain lower linguistic development scores in constrained elicitation tests compared to spontaneous production data. It is therefore necessary to compare different contexts of production, as much as possible, by combining different methods in order that the analyses accurately reflect the real linguistic competence of speakers.

3.5. Repetition tasks

To conclude the presentation of production tasks, in this section, we introduce the method displaying the highest degree of linguistic constraint on production: the repetition task. As its name suggests, the repetition task involves asking participants to repeat either a word or a sentence after it has been presented. These tasks are based on the observation that linguistic repetition is not a simple imitation task, but requires the ability to process the stimulus. In the case of sentences, numerous studies in the field of language acquisition have shown that it is not possible for young children to properly repeat sentences which are not yet part of their grammatical system (e.g. Bloom *et al.* 1974), that is, sentences that they would not be able to produce by themselves. This inability is illustrated in an amusing way in the following dialogue between a father and his child, retold by Pinker (1994, p. 281):

> Child: want other one spoon, Daddy.
> Father: you mean you want the other spoon?
> Child: yes, want other one spoon, please, Daddy.
> Father: can you say “the other spoon”?
> Child: other… one… spoon.
> Father: say “other”.
> Child: other.
> Father: “spoon”.

Child: spoon.
Father: “other… spoon.”
Child: other… spoon. Now give me other one spoon.

Sentence repetition tasks make it possible to accurately test which elements are still problematic for children. For example, it is possible to test the role of semantic and syntactic representations of children in their ability to interpret relative clauses, by modifying its lexical head (McDaniel *et al.* 1998 p. 57). In example (12), the lexical head has a precise semantic meaning, whereas this is not the case in (13). The role of syntax can be tested by alternating a sentence with a lexical head such as in (13) and without it, such as in (14).

(12) Max bought the toy Paul chose.

(13) Max bought the thing Paul chose.

(14) Max bought what Max chose.

A comparison of children’s repetition abilities makes it possible to determine whether it is the semantic or syntactic factors that appear to cause problems for young speakers, while in the phase of acquiring relative clauses.

Repetition tasks can be used for testing many aspects of syntax, such as constituency structure, as in the example above, as well as constraints associated with word order. For example, Lust and Wakayama (1989, cited by McDaniel *et al.* 1998) used this method with Japanese children to test the repetition of sentences with an unmarked SOV order in Japanese (15) and a right-dislocated order (16). Most mistakes made by these children, in repeating sentences with right dislocation, corresponded to an attempt to restore the canonical order of words. This experiment shows that young children already integrate the constraints of syntactic linearity of their mother tongue:

(15) Rion-to tora-gahashiru (“Lion(s) (and) tiger(s) run”).

(16) Hashiru-yousagi-to kame-ga (“Run, rabbit (and) tortoise”).

In some cases, it is also possible to provoke repetition of incorrect sentences, in order to determine whether children and learners are already sensitive to certain aspects of lexicon and syntax. Children and adults tend to correct mistakes when repeating a sentence. This paradigm can be used for testing irregular inflected forms ("*you goed*" instead of "*you went*") as well as agreement mistakes ("*two big horse*" instead of "*two big horses*"). This type of paradigm has also made it possible to show that children distinguish words that are mistakenly repeated due to a fluency problem, as in (17), from words having an intentional repetition, as in (18):

(17) He is, he is very kind.

(18) He is very, very kind.

In addition to syntactic structures, repetition tasks can focus on words. An example of a widely used paradigm is the repetition of non-words, that is, words which could exist, according to morphophonological rules in a language, but which do not exist in the lexicon, such as *degate* or *galpin* in English. This task tests the ability of people to process the phonological component of words. It is often used in research on language impairments, since the inability to repeat non-words, which reflects a deficit in phonological working memory, is one of the linguistic markers typical of specific language impairments (e.g. Coady and Evans (2008)).

In other experiments, children are asked to repeat the last word they heard while reading sentences with regular interruptions. This type of paradigm makes it possible to determine which linguistic elements young children consider to be words, without having to resort to a metalinguistic task requiring an explicit reasoning on language. Through this type of task, Karmiloff-Smith and her team (Karmiloff and Karmiloff-Smith 2003) showed that by the age of 4 years, the majority of children already consider both content and functional words as words, and that this rate comprises almost all children by the age of 5 years. Yet, at this age, children are unable to answer explicit questions about what a word is.

We should also point out that certain word or sentence reading tasks may resemble a form of repetition of written prompts. However, reading tasks are mainly used for testing elements related to the linguistic signals produced (such as phonology or speech prosody), rather than as a way of indirectly studying the linguistic representations of speakers. These tasks are used, in particular, to accurately determine the different pronunciations of a phoneme depending on the speaker's geographic region. This method has the advantage of making it possible to control the effects of the phonological environment on pronunciation, for example, by testing oppositions between phonemes placed at the beginning of a word, between open or closed syllables, etc. Schwab and Avanzi (2015), for example, sought to determine whether French speakers from French-speaking Switzerland and Belgium had a slower articulation speed than French speakers from France. Speech excerpts from two speech contexts were compared. The first context was a reading task, whereas the second one was a conversation excerpt. Results showed that articulation speed varied significantly from region to region (people living in French-speaking Switzerland tend to have a slower articulation speed), as well as the fact that the speech context played an important role, since the articulation speed of syllables was faster in conversations than in reading. This study thus provides an additional illustration of the need for combining data from elicitation tasks with natural data.

In summary, repetition tasks can be used as tools to assess the linguistic representations of speakers. They are valid at the lexical and syntactic levels. Due to the limitations of working memory, it is difficult to use this method for testing elements beyond the sentence level. This type of task also has the advantage of being applicable to children between 1 and 2 years old (McDaniel *et al*. 1998) since the latter develop imitation abilities from a very early age. On the other hand, studies have shown that learners do not correctly imitate sentences that they are able to spontaneously produce correctly (e.g. Bernstein Ratner (2000)). The main difficulty for applying a repetition task is finding the right level to test the skills of a certain group of speakers. If the task is too simple, all sentences will be repeated correctly. If the sentences are too complex, the task will no longer make it possible to draw a distinction between the different structures or words tested. As with any experimental task, use of a repetition task also requires rigorous control of the experimental material. Just to give an example, it is necessary to ensure that the different sentences are equivalent in terms of the number of syllables they contain, word frequency, etc. Finally, we should bear in mind that sometimes, when the participants do not repeat a sentence correctly, it is not always easy to explain such mistakes. Indeed, repetition mistakes do not necessarily imply a lack of competence, but can sometimes reflect a limitation in the ability to process information, which may lead to replacing a certain structure by a simpler one. In this case, the incorrect repetition would reflect a problem of performance more than of linguistic competence. Again, to limit this bias, it is necessary to diversify the research strategies, in order to benefit from the advantages of each of them.

3.6. Conclusion

In this chapter, we started by introducing the differences between language comprehension and language production, and argued that these two components of linguistic ability should be investigated in parallel. We saw that language production skills are often more limited than comprehension skills in all groups of speakers but the reverse asymmetry also exists. We then focused on the important methodological difference between the observation of production in a corpus and experimentally elicited production. We have seen that corpus data have the advantage of being natural and able to contain very large samples of language, but that the data resulting from elicitation tasks are more suitable for studying rarer linguistic phenomena, syntactic differences or subtle differences in meaning.

Among the various tasks that can be used to experimentally provoke linguistic productions, we presented free elicitation, constrained elicitation and repetition tasks. We established that free elicitation tasks can be used in addition to corpus data to increase the number of occurrences of a certain linguistic phenomenon, as well as for testing the role context plays in production. This method is particularly suitable for testing discursive phenomena. Constrained elicitation makes it possible to test the ability to produce precise words or syntactic structures in a quantitative manner, and within a controlled context. Finally, repetition can also be applied to words and sentences, making it possible to indirectly assess the way in which speakers process and understand these elements.

3.7. Revision questions and answer key

3.7.1. *Questions*

1) List three arguments that justify the need to study language production and comprehension separately.

2) What are the main advantages of experimenting with language productions rather than observing them in a corpus?

3) What would be the most appropriate experimental method for eliciting productions in the following research questions?

a) What are the syntactic and semantic constraints at work in the acquisition of relative clauses?

b) Are learners able to formulate indirect speech acts in a foreign language?

c) Are adults able to produce different types of relative clauses?

4) What are the common points between free elicitation and the observation of productions in a corpus?

5) List and explain three methodological constraints related to the development of a constrained elicitation task.

6) What are the main advantages and disadvantages of a repetition task, compared to a constrained elicitation task?

3.7.2. *Answer key*

1) A first argument showing that these two components of the language faculty are dissociated and should be studied separately is that children and learners do not develop linguistic competences at the same rate in the field of comprehension and language production. A second argument would state that language impairments may affect the competences of speakers in one area while preserving another, such as Broca's aphasia, which essentially affects language production. Finally, a third argument comes from corpus studies and experiments showing that the size of the mental lexicon for language production and language comprehension is very different.

2) One of the great advantages of elicitation tasks is that they make it possible to control the context in which language productions take place. As we saw in the chapter, context often has great importance on the quantity and quality of the language produced by participants. Furthermore, elicitation tasks make it possible to collect a large number of occurrences of rare linguistic phenomena by forcing people to produce the targeted elements. These phenomena cannot be analyzed on the basis of corpus data, due to the small number of occurrences found there. Finally, elicitation tasks allow you to have a better grip on what the participants' intentions are when they produce certain words or sentences. Indeed, in these tasks, participants must produce words or sentences for describing an image or a video. It is thus possible to check that the words or sentences are used in an appropriate way. On the other hand, in a corpus, it is not always possible to determine what a speaker intended to communicate.

3) a) For this study, a repetition task seems the most appropriate, since it would make it possible to accurately test whether very fine syntactic or semantic variations in stimuli have an impact on the way in which the participants reproduce them.

b) This study would require the use of a free elicitation task in order to give participants enough freedom to use various structures for expressing requests. Indeed, a constrained elicitation task would bias the results, by pushing the participants to use certain formulations, which might not correspond to the way in which the requests are spontaneously produced.

c) This study could be carried out with a constrained elicitation methodology, for example, by presenting the beginning of a sentence until the relative pronoun ("I like the man who…" or "I like the car that…"), and then asking the participants finish the sentence. These different prompts would make it possible to compare the way in which the participants complete relative clauses with a subject pronoun (who) and with an object pronoun (which/that). Another study, in which the relative pronoun is not included, would make it possible to determine whether the participants prefer to complete a sentence with a subject or an object relative clause.

4) Like corpus data, free elicitation tasks have the advantage of providing relatively natural outputs, since the interference of researchers remains very low. The latter consists only of placing the participants in a certain linguistic situation. This is why these two methods are well suited to the study of spontaneous speech but more limited for eliciting repeated productions of a specific element. Rare elements often cannot be studied quantitatively using these methods.

5) First, constrained elicitation tasks involve the need to find a context in which the targeted production is mandatory. For example, to trigger the production of a relative clause, it is not enough to start a sentence with a noun phrase and ask the participants to complete it (e.g. "the little girl…"), since other options, simpler than the relative clause, are possible and will probably be chosen (e.g. "the little girl with red hair" rather than "the little girl who has red hair".) It is also essential to check that the targeted productions are those which are produced spontaneously, by carrying out a pretest with other populations than with the one that will be tested, for example, native speakers in the case of studies on learners, or adults in the case of studies on children. Second, these tasks involve checking that all the elements of the prompts have a suitable level of difficulty. This level must

also be constant between the different experimental items. For example, word frequency of the words used and sentence length should enable participants to understand the task and what is expected of them. Third, for these tasks to be valid, it is advisable to avoid modeling the expected answer. It is therefore not possible to give examples of the expected structure, which limits the use of this task with some populations, such as young children.

6) The advantages are that repetition tasks make it possible to test younger children and learners at a less advanced level than constrained elicitation tasks, because these are simpler. In addition, they make it possible to test very fine grained factors, such as the semantic and syntactic differences in sentences, or the alternation between phonemes. Their disadvantages are that item difficulty must be very well calibrated. If they are too simple, the participants will easily reach a maximum score, and if they are too complex, the participants will fail for the wrong reasons. Furthermore, mistakes made during repetition tasks are not always easy to interpret. Finally, these tasks are not at all natural and do not provide information on what the participants would spontaneously produce.

3.8. Further reading

The different types of elicitation tasks are presented in detail in the book by Gass and Mackey (2007), which places them in the context of research in language acquisition and learning. Menn and Bernstein Ratner (2000) also provides very complete references on the different methods for studying the linguistic productions of different populations. Eisenbeiss (2010) provides a more concise introduction to the analysis of elicited and spontaneous productions and clearly presents their advantages and limitations. In the field of language acquisition, the work of McDaniel *et al.* (1998) contains several chapters dedicated to methods for testing the syntactic productions of children. These methods are presented in a very concrete way with lots of methodological advice. In the field of sociolinguistics, Schilling (2013) discusses the methodological aspects related to the creation and analysis of surveys.

4

Offline Methods for Studying Language Comprehension

In this chapter, we begin with the presentation of experimental methods which can be applied to the study of comprehension, while focusing on so-called *offline* methods. These methods are typically used for observing the result of the comprehension process once it has been achieved, but do not provide access to the comprehension process itself; this is why they are called offline methods. We first discuss what are known as explicit tasks, where participants are asked to consciously assess certain aspects of the language, or certain linguistic stimuli. We then describe so-called implicit tasks which assess the comprehension process indirectly. As we see throughout this chapter, the distinction between explicit and implicit tasks appears more as a continuum rather than as two clearly distinct categories. Since language comprehension covers a broad range of processes, from word recognition to discourse comprehension, we won't deal with each of these areas in particular. Instead, we make a general presentation of the different types of tasks that can be implemented and illustrate them with the help of studies devoted to these different fields. We also see that it is possible to use different techniques in parallel in order to collect indicators which signal complementary comprehension processes.

4.1. Explicit tasks

In an explicit task, participants must consciously use their language skills to judge stimuli, such as the grammaticality of a sentence. In this case, the purpose of the study is not concealed, as it explicitly deals with language

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

comprehension. Explicit tasks often rely on the metalinguistic abilities of the individuals tested, on their capacity to consciously reflect about language and its uses, and to report their linguistic knowledge or intuitions. For example, this would be the case with a task requiring judgment of whether a certain formulation of a speech act is polite or not.

We will present five types of explicit tasks: metalinguistic tasks, acceptability judgments, questionnaires, interpretation tasks and comprehension tests. As we will see in the course of the chapter, explicit tasks have limitations due to the fact that they crucially depend on the level of linguistic competences or metalinguistic abilities of the participants. This limitation is particularly problematic for the application of such tasks to populations whose metalinguistic abilities or linguistic competences are restricted or weaker, such as children or people suffering from language impairments.

4.1.1. *Metalinguistic tasks*

Along the continuum between explicit and implicit, metalinguistic tasks can be considered as the most explicit ones. In this type of task, participants are asked to consciously reflect about language as an object of study rather than as a means of communication. In practice, metalinguistic tasks may deal with all the aspects of language, such as phonological awareness, syntax or the comprehension of conversational implicatures. These tasks may be framed within different methodological paradigms, depending on the metalinguistic abilities investigated. To study phonological awareness, for example, one possibility would be to ask the participants to split words into syllables. To study syntax, one possibility would be to ask the participants to explain a grammatical rule accounting for the grammaticality or ungrammaticality of a series of sentences. Finally, to study the comprehension of conversational implicatures, one possibility would be to ask participants to explain the difference between literal meaning and contextually communicated meaning for a number of statements. In all cases, metalinguistic tasks are based on the ability of the individuals to reflect upon language and to report the product of their reflection.

Our first example of a metalinguistic task comes from the study by Colston and Gibbs (2002), concerning the comprehension of irony and the metaphor. According to their hypothesis, understanding irony requires second-order inferences (such as “X believes that Y believes that Z”), related to the interlocutor’s intentions and beliefs. However, these are not necessary for understanding the metaphor, which only requires first-order inferences (“X believes that Y”). In their study, participants had to read small scenarios and adopt the viewpoint of one of the characters. The last sentence of each scenario conveyed either a metaphor or irony. An example of such a scenario for the metaphorical condition involved a teacher talking about a student and ending the comment with “this one is really sharp”, an expression describing somebody as witty. In this case, through the use of a metaphor, the teacher was referring to a real virtue in the student (the fact of being intelligent). In the ironic condition, the scenario changed to a teacher looking for a pair of scissors, and – unable to find one that really worked – uttering “this one is really sharp”. In this case, the teacher was referring to an instrument in need of sharpening, but mentioned this property ironically, as the instrument was unsharpened.

Colston and Gibbs (2002) used different methods for assessing the comprehension of metaphorical and ironic sentences; here, we will discuss the second experience of their article, using a metalinguistic task. After the presentation of each scenario, the participants (university students) had to indicate their agreement with five statements aimed at assessing different metalinguistic skills necessary to understand metaphors or irony. Such statements targeted the comprehension of the speaker’s intention as in (1a) and (1b), the fact that this person did not actually think what they are saying, as in (2a) and (2b), the reference to this person’s personal beliefs as in (3a) and (3b), the existence of several possible beliefs as in (4a) and (4b), and finally the fact that the ironic sentences made fun of previous beliefs, as in (5a) and (5b):

> (1a) The teacher’s remark reflects her current belief that the student is smart. (metaphor).
>
> (1b) The teacher’s remark reflects her current belief that the scissors are not sharp. (irony).
>
> (2a) The teacher’s remark reflects the fact that she is only pretending that the student is a cutting instrument. (metaphor).
>
> (2b) The teacher’s remark reflects the fact that she is only pretending that the scissors are sharp. (irony).

> (3a) The teacher's remark refers to her prior belief (meaning her belief about the student before the conversation) that the student should be smart. (metaphor).
>
> (3b) The teacher's remark refers to her prior belief (meaning her belief about the scissors before the conversation) that the scissors should be sharp. (irony).
>
> (4a) The teacher's remark reflects her multiple beliefs, in that she is both referring to her present belief that the student is smart and her prior belief that the student should be smart. (metaphor).
>
> (4b) The teacher's remark reflects her multiple beliefs, in that she is both referring to her present belief that the scissors are not sharp and her prior belief that the scissors should be sharp. (irony).
>
> (5a) The reason that the teacher possibly refers to her prior belief that the student should be smart is to mock this expectation given that the student is smart. (metaphor).
>
> (5b) The reason that the teacher possibly refers to her prior belief that the scissors should be sharp is to mock this expectation given that the scissors are not sharp. (irony).

The results indicated a similar level of agreement for the statement concerning the speaker's intention, between the irony and the metaphorical condition. For all the other assertions, agreement was always higher in the ironic condition than in the metaphorical condition. According to the authors, these results show that people are aware of the essential difference between irony and metaphor, particularly in relation to the necessary metalinguistic reasoning for understanding irony.

Another example on how to implement a metalinguistic task comes from Borghi *et al.*'s (2016) study, which delved into the meaning of abstract concepts. The aim of the study was to determine to what extent the comprehension of abstract concepts depends on linguistic and contextual elements. There are different ways of conceptualizing the representation of concrete and abstract concepts. A first approach would be to consider that representations, and cognition in general, are built on the basis of abstract symbols. In such a conception, the sentence “Laura passes the yellow ball to John” would simply be a proposition based on the predicate TO PASS, the subject LAURA, the object BALL, modified by the adjective YELLOW, to the recipient JOHN. A second approach, following the perspective of embodied – or grounded – cognition, would be to consider that representations and cognition are not only based on abstract symbols, but also on experience. From this perspective, the meaning of a linguistic stimulus emerges via a simulation process, based on experience, in order to build a representation going beyond the stimulus itself. Understanding the sentence “Laura passes the yellow ball to John” could, for example, involve a simulation of Laura’s movement for passing the ball. It could also activate the image of a tennis ball that matches the color of the ball in the sentence, or adopt Laura’s perspective rather than John’s.

Borghi *et al*. (2016) hypothesized that if comprehension was grounded in experience, then abstract concepts such as *risk*, *danger* and *prevention* should be understood differently by people having a different relationship with these notions in everyday life. To test this hypothesis, the researchers chose four groups of people with different theoretical and practical expertise in the field of safety and security (S&S) at the workplace. There were managers with first-class theoretical knowledge but no practical experience, security technicians with both theoretical and practical knowledge, trade union delegates specifically trained on safety issues, and workers lacking theoretical knowledge, but with everyday practical experience in these issues. The authors asked the participants to define the terms *risk*, *danger* and *prevention,* transcribed and then coded the definitions. For the coding phase, they made a distinction between four types of components which emerged from the replies: situational, introspective, taxonomic and attributive components. They then analyzed which types of components were related to the three concepts, as well as the types of components reported by the different groups of participants.

The results of the study showed that, for the three concepts, the participants mainly reported situational components, related to their own experience, supporting the hypothesis that the comprehension of an abstract concept is grounded in experience. By comparing the different definitions reported by the four groups, Borghi *et al.* (2016) found that workers provided the greatest number of components. Next, came the security technicians, the managers and then the union delegates. Experience seemed to play a role in the conceptualization of abstract concepts, since the participants with the most practical experience – the workers – offered the definitions containing the most components. By looking more closely at the types of components mentioned by each group, the researchers were also able to show that while all the groups mainly relied on situational components, the groups with the most practical experience offered more introspective elements than the others. The groups with the most theoretical experience, for their part, focused on taxonomic and attributive components. The results of this study support the hypothesis that representations and knowledge are not only symbolic, but, most importantly, are also grounded in experience.

The examples described above illustrate the advantage of metalinguistic tasks for reaching the conscious reflection processes required for language comprehension. They also illustrate the fact that such tasks have two important prerequisites: that the people tested have conscious access to their intuitions about language, and that they have the ability to report such intuitions. Some demographics, such as young children or people suffering from language impairments, cannot be tested with metalinguistic tasks. Furthermore, these tasks can quickly become complex. Going back to the example of Colston and Gibbs (2002), we can observe that the statements to be evaluated are complex sentences, some of which involve several subordinates. In order to be able to carry out this metalinguistic task, it is necessary for the participants to have good linguistic competences. These cannot be taken for granted, even in a population of adults, since there are multiple individual differences in language proficiency (Kidd *et al.* 2018; Zufferey and Gygax 2020).

4.1.2. *Acceptability judgments*

As their name implies, acceptability judgments are tasks in which people are asked to judge the acceptability of a sentence or a linguistic form. These judgments can be assessed by different means, either through binary evaluations (yes/no) or via an acceptability scale, offering several levels of acceptability. This second possibility makes it possible to qualify the responses, by means of a scale comprising intermediate levels ranging from *not at all acceptable* to *totally acceptable*. In this case, the measurement corresponds to the perceived degree of acceptability. Instead of measuring the acceptability of an isolated stimulus, it is possible to simultaneously present two stimuli and have the participants choose the most acceptable one. This procedure makes it possible to directly compare two linguistic stimuli. A final way of collecting acceptability judgments is to present a reference stimulus that is associated with a certain degree of acceptability and to have the degree of acceptability of other stimuli evaluated in comparison with the former. This method, drawn from psychophysics, is called magnitude estimation and was borrowed by linguistics (Bard *et al.* 1996). Despite their differences, it is interesting to observe that these methods all seem to be informative in their own right (Weskott and Fanselow 2011).

Acceptability judgments have mainly been used in the fields of syntax and semantics. Linguists, particularly in the Chomskyan tradition, think that acceptability judgments shed light on the structure of the knowledge of language and represent a direct reflection of it (Chomsky 1986). However, certain studies having used this method resort to a methodology which does not correspond to experimental methodological standards. As a matter of fact, these studies were often carried out on a small number of participants, linguists themselves, with few stimuli and few response options, which permitted only basic analyses of data. In the scientific literature, there is a lively debate on the merits of these studies and on the possibility of drawing reliable conclusions on the basis of only a few stimuli evaluated only by experts. Various studies have shown that the responses of naive subjects differ from those of experts, which calls into question resorting to the latter (Gordon and Hendrick 1997; Dabrowska 2010).

When used in compliance with the principles of experimental methodology, by presenting numerous items to many naive subjects, acceptability judgments can provide quality information. Below, we will illustrate different ways of implementing a study based on such judgments.

An example of an acceptability task can be found in Zufferey *et al.* (2015b), who investigated the influence of L1 transfer effects on the comprehension of discourse connectives in a second language. According to Zufferey *et al.*, transfer-based errors in production are often mirrored by the use of connectives among learners of a second language. For example, French-speaking learners of English tend to misuse the connective *if* (*si* in French) for conveying contrastive relations (instead of the connective *while*), whereas, in English, it cannot fulfill this function. This can be explained by the fact that in French, the connective *si* can convey a conditional relation, as in the sentence "*si elle ne vient pas demain, je lui téléphonerai*" (if she doesn't come tomorrow, I'll call her) or a contrastive relation, as in "*si en Belgique ce groupe a beaucoup de succès, il est encore inconnu en France*" (while this group is very successful in Belgium, it is still unknown in France). Likewise, it has been shown that Dutch-speaking learners of English tend to use the connective *when* for conveying conditional relations instead of the connective *if*. Again, this can be explained by L1 transfer effects.

In their study, Zufferey *et al*. (2015b) tested French-speaking and Dutch-speaking learners of English, as well as native English speakers. Sixteen sentences requiring the use of the connective *if* (6) were created, as well as 16 sentences requiring the use of the connective *while* (7). In order to build an incorrect version of each sentence, the connective *if* was replaced by *when* (8), and the connective *while* was replaced by *if* (9):

> (6) The kids don't look very tired today. *If* they don't take a nap now, we can go out for a walk.
>
> (7) The admission policy for foreign students is variable across universities. *While* in some of them all students can enroll, in others there is an entrance examination.
>
> (8) The kids don't look very tired today. *When* they don't take a nap now, we can go out for a walk.
>
> (9) The admission policy for foreign students is variable across universities. *If* in some of them all students can enroll, in others there is an entrance examination.

The study included a reading task, which we will not detail, as well as a sentence judgment task, which we will focus on. In this last task, the different sentences contained either the correct connective or the wrong one. In addition to these sentences, there were filler sentences, aiming to hide the purpose of the experimental manipulation. We will go back to the notion of fillers in Chapter 6, but it is already useful to say that, in an experiment, the experimental material is hidden within the non-experimental material so that the central manipulation cannot be detected. In this experiment, the filler sentences corresponded to sentences containing the connective *when*, conveying a temporal relation (the correct version), sentences containing relative propositions introduced by a correct or an incorrect relative pronoun, as well as sentences containing obvious mistakes (such as subject and verb agreement mistakes), in order to verify the participants' level of attention.

In the judgment task, the participants received the different experimental and filler items in writing and had to indicate for each sentence whether it was correct or not. If the sentence was judged as incorrect, the participants had to circle the mistake in order to check that the connective or the filler mistakes were the source of their response. The number of correct answers was then compared among the different conditions.

The results showed that the incorrect use of *when* (conditional relation instead of temporal) was less detected by Dutch speakers than by French or English speakers. Likewise, the incorrect use of *if* (contrastive relation instead of conditional) was less detected by French speakers than by Dutch or English speakers. These results clearly support the hypothesis of a transfer effect in the ability of learners to detect the misuse of connectives in their second language. It is important to note that the results of the online task performed on the same material suggested that incorrect uses of connectives were detected at the reading stage, even if these were not necessarily consciously reported later. This indicates that it is often necessary to test comprehension in different ways, in order to have a more global picture of the processes involved and about the suitability of the different tasks for detecting them.

Acceptability judgments can also be made using a variety of materials, including linguistic and visual cues. For example, Coventry *et al.* (2001, experiment 1) studied the influence of certain components of a visual scene on the comprehension of spatial prepositions in English: over, under, above and below. Numerous studies have been carried out on spatial prepositions, in order to better grasp the relations underlying each of them (including, for example, Carslon-Radvansky and Radvansky (1996) and Logan and Sadler (1996)), showing, for instance, that *above* differs from *over*, in that it defines a higher point which is not in direct contact with the reference object.

It has also been shown that other variables influence the use of these spatial prepositions, such as the frame of reference (Levinson 1996) or the presence of a functional relationship between the elements described in the visual scene (Carslon-Radvensky and Radvansky 1996). In the study we will discuss here, Coventry *et al.* manipulated two variables: the geometric relationships between the different objects in the scene and their function.

The geometric relationship was operationalized as the position of the object in relation to the ground: the canonical orientation (usual position), an angle of 45º or an angle of 90º. The function of the objects was manipulated in the following way: an element in the scene was missing for the function of the object to be fulfilled, the object fulfilled its function or it didn't. The combination of the modalities of each variable led to nine possible images for each scene, as illustrated in Figure 4.1 for a rain scene. There were eight different scenes in total, making a total of 72 possible images.

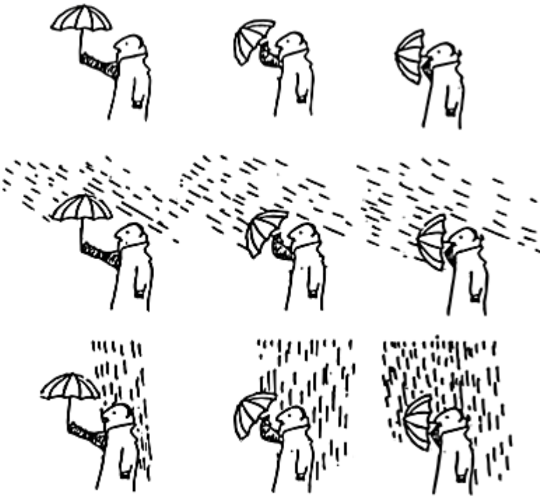


Figure 4.1. *Examples of situations presented in Coventry* et al. *(2001)*

Each image was associated with two pairs of sentences to be evaluated: one pair containing the prepositions *over* and *under*, and another pair containing the prepositions *above* and *below*. Participants received a booklet containing the different images associated with different pairs of sentences, presented in a random order. Their task was to assess the acceptability of each sentence for each image, on a scale ranging from 1 (*totally unacceptable*) to 7 (*totally acceptable*). The results showed that the two independent variables examined played a role in the choice of the appropriate spatial preposition. As in previous studies, the acceptability of the different prepositions was maximum when the object was in its canonical orientation, and decreased when the angle increased. In relation to the object function, the acceptability of the prepositions was higher when the object fulfilled its function and lower when it didn't fulfill it. Furthermore, the function influenced the acceptability of the prepositions only in the case where the angle's position was not standard. Interestingly, the results also revealed that the acceptability of the different pairs of prepositions was not influenced in the same way by the variables studied. The pair *over/under* was more prone to the influence of function than the pair *above/below*, whereas the opposite was apparent for the influence of angle.

There is also a particular case of acceptability judgments, known as lexical decision tasks. In these tasks, participants have to decide whether a linguistic stimulus is a real word in their language or not. Lexical decision tasks are usually associated with online measurements, as the time required for decision-making is often recorded and analyzed. For this reason, this type of task will be presented in the next chapter devoted to online measurements. However, they can also be applied without measuring the response time, in order to collect the participants' answers. For example, Lemhöfer and Broersma (2012) developed an instrument to test proficiency in English as a second language based on the answers to a lexical decision task, including only 60 items. This questionnaire, *Lextale*, was later adapted to other languages such as French (Brysbaert 2013), Dutch and German.

In summary, acceptability judgments have several advantages. To begin with, they offer the possibility of studying forms which have never been pronounced or are impossible to find in a corpus. They are also easily set up, since they can be carried out with simple means (paper and pencil tasks or online questionnaires). Finally, they can be combined with metalinguistic response justification tasks, in order to answer specific research questions. The main limitation of these tasks is that the quality of acceptability judgments is not always optimal. First, acceptability judgments may not reliably reflect the structure of the language, as participants are sometimes influenced by other factors, such as the overall meaning of the sentence. Therefore, conclusions drawn from acceptability judgments often have to be corroborated by other methods. We will see that this limitation may also apply to other tasks that we will develop later. In general, every method has its limitations and conclusions should be based on results drawn from several studies implementing different tasks. In the example by Zufferey *et al.* (2015b) presented above, combining the results of online and offline studies made it possible to hypothesize that learners can unconsciously detect mistakes which are not consciously reported when the task requires explicit reflection upon linguistic rules.

A different yet related point is that the validity of the measurement resulting from judgment tasks can be called into question. Indeed, it is possible that the participants do not share the same opinion as linguists about what is (grammatically) acceptable and what is not. Similarly, the acceptability of stimuli may not depend exclusively on the variable under investigation, but on other aspects, such as the difficulty of processing them or their improbability (Branigan and Pickering 2017). The validity of acceptability judgments has often been questioned, since these vary significantly from individual to individual, and even between test phases (Gibson and Fedorenko 2010, 2013; Schütze 2016). A final limitation related to acceptability judgments is the necessity they imply, for the interviewees to have a certain metalinguistic competence. Thus, the above-mentioned limitations concerning metalinguistic tasks can also be applied to these tasks.

4.1.3. *Questionnaires*

Rather than measuring the acceptability of a linguistic stimulus, it is possible to set up questionnaires for testing comprehension indirectly, how a statement was perceived or whether an argument was found convincing or not. For example, this is the case for the study by Schumann *et al*. (2019) on the factors influencing the effectiveness of a particular type of fallacious argument called *the straw man*. The straw man fallacy corresponds to an exaggeration of the original argument advanced by an opponent, in order to present it as unacceptable and easily refute it. Schumann *et al*. identified three linguistic variables which could influence the acceptance of straw man fallacies, and separately tested them in three experiments. In each experience, dialogues were presented, including the intervention of a first speaker presenting a point of view followed by an argument (10), then the response of a second speaker, which could contain *straw man-type* (11) fallacious argument, or not:

(10) Barbara: it is essential to support young parents because having a child means a lot of financial charges.

(11) Alexander: let’s increase family allowances since it is only about money.

These fallacious arguments were constructed so as to test three variables of interest. Here, we will only present one of them by way of illustration. One of the hypotheses of the study was that a fallacious argument introduced by a connective like *puisque* (closely related to the English *since*) – associated with subjectivity – might trigger the fallacious character of the argument, compared to an argument introduced without using such a connective. Following the presentation of a dialogue, the participants had to respond to four questions relating to the arguments presented in order to – indirectly – assess the effect of the straw man features on the acceptance of such arguments. The first question aimed to evaluate whether the characteristic exaggeration of the straw man was detectable. The second question assessed whether the logical link was perceived to be deficient when the straw man argument was present. The third question assessed the participants' degree of agreement with the person who had employed a straw man argument. Finally, the fourth question evaluated the participants' degree of agreement with the first speaker's initial affirmation. The results particularly revealed that fallacious straw man arguments were better accepted when the argument was not introduced by a connective rather than when it was preceded by the connective *puisque*. Therefore, some connectives such as *puisque* alert participants about the subjective and potentially fallacious nature of an argument. In a second set of experiments, they replicated these results in English with *since*, demonstrating that this effect is not specifically linked to one French connective (Schumann and Zufferey 2020).

The type of questionnaire used by Schumann *et al*. (2019) differs from the metalinguistic tasks described above in the sense that language is considered as a vector for communication. The comprehension element under investigation does not relate to the linguistic structure itself, but to the participants' perception about the merit of an argument. Questionnaires make it possible to build items to specifically match the interests of every study, and applicable to all areas of linguistics. However, when building a questionnaire, it is necessary to carefully ponder the many aspects involved, such as the number of questions, the wording of the questions, the answer options, etc. These aspects will not be developed in this book, but we offer some references tackling them at the end of the chapter.

4.1.4. *Forced-choice preference tasks*

Another means to explicitly assess comprehension are preference tasks, where participants have to choose one from among several answers on the basis of instructions related to certain linguistic properties. These tasks can also take the form of matching tasks between linguistic stimuli and images, in which the stimuli can be words or sentences. In these tasks, participants are asked to choose one image from among many that best represents the stimulus, or to choose a word or phrase that best suits the stimulus presented in the form of images.

For example, Colonna *et al.* (2012, experiment 1) studied anaphora resolution in French and German speakers, specifically focusing on the influence of subject or object topicalization in sentences. Previous studies had revealed a difference between French and German as to the resolution of ambiguous pronouns, like in (12). While German speakers preferred to associate the pronoun with the first noun mentioned (the postman), French speakers matched it with the second noun (the sweeper):

(12) The postman met the street-sweeper before he went home.

In this study, the researchers hypothesized that these differences in language preferences could be found in ambiguous sentences such as (13). They also wished to find out whether these preferences could be modified by topicalizing either the subject (14) or the object (15) of the sentence. They therefore created items in German and in French, similar to the examples below, which they presented in three versions: an ambiguous version, a version with subject topicalization and a version with object topicalization. Every sentence was followed by a proposition containing the information provided in the subordinate, except for the subject, which the participants had to complete with the name they found appropriate (16):

(13) Peter slapped John when he was young.

(14) As for Peter, he slapped John when he was young.

(15) As for Peter, John slapped him when he was young.

(16) …was young.

The results of this study showed that in the ambiguous condition (13), the choice of the first referent (Peter) was much more frequent among German speakers (69.4%) than among French speakers (38.5%). This was expected, based on the results obtained with sentences like (12). Subject or object topicalization produced different results depending on the language of the participants. Among German speakers, subject topicalization did not increase the proportion of people choosing the first referent. This can be explained by the fact that the first referent is the standard choice in this language. Object topicalization, on the other hand, entailed a decrease in the proportion of people choosing the first referent, indicating that topicalizing makes it possible to influence anaphora resolution, but only when it brings an unusual order to light. Among French speakers, subject topicalization had no effect, and the object was always preferred as a referent. On the other hand, object topicalization increased the preference for the subject. In the latter, we can find the influence of topicalization once more. These results show that topicalization influences anaphora resolution but that it is difficult to reach a general conclusion regarding this effect.

Preference tasks can be adapted to children or to populations with language impairments. It is possible to construct such tasks using non-linguistic stimuli and to ask the subjects to point to the image corresponding to their interpretation. In this case, we speak of *pointing* tasks. For example, Bernicot *et al*. (2007) were interested in the way children understand various forms of nonliteral language uses. In order to investigate this question, they examined children between 6 and 10 years old, dividing them into three groups: 6, 8 and 10 years old. They operationalized the difficulty of the nonliteral form by choosing to observe various forms requiring different types of inferences: indirect requests as in (17), idioms as in (18), semantic-inference implicatures as in (19) and conversational implicatures requiring a sarcastic inference as in (20):

(17) Cold is coming through the window. (Close the window.)

(18) Change your tune. (Change the subject.)

(19) Should I mow the lawn? The nephews are sleeping in their room. (No.)

(20) Should I open the umbrella? No, I really like getting sunburnt. (Yes.)

For every type of nonliteral form, researchers created four short stories presented as images, which were then shown to the children on a screen, in the form of a video game. The children had to choose the last picture in the story, which represented an action verifying whether the nonliteral form had been understood or not. For example, in the case of idioms, the image represented either a character changing the subject or a character changing a record. The data collected made it possible to quantify the number of children nonliteral language uses and to analyze the influence of the variables *age* and *nonliteral form* on the number of correct answers given. The results showed that age influenced the comprehension of nonliteral forms, 10-year-olds giving more correct answers than 8-year-olds, the latter giving more correct answers than 6-year-olds, except for indirect requests, for which all age groups obtained similar scores. At the same time, the difficulty of the inferences necessary for understanding the nonliteral form also influenced their comprehension. Children gave more correct answers for semantic-inference conversational implicatures than for indirect requests, then for idioms, and, finally, for sarcastic-inference implicatures. This study also showed that the mastery of each type of form is reached at different ages. At the age of 6, children generally understand semantic-inference implicatures and are close to mastering indirect requests; at the age of 8, they generally master indirect requests, and by the age of 10, they master idioms. The mastery of sarcastic-inference implicatures is not reached at the age of 10 years, and does not seem to appear until later in development.

However, these results do not provide all the information one might desire to collect about the acquisition of nonliteral forms by children. While they properly show to what extent children can understand these forms, it is not clear whether they understand *the reason why* these forms do not convey the literal meaning of what is stated. In order to answer this second question, researchers added a metalinguistic task at the end of each comprehension test, inviting the children to explain why they had chosen a certain image. The answers were then evaluated by researchers, and classified into three categories. The first category corresponded to irrelevant explanations which were simple descriptions or reformulations of what was happening in the chosen image. The second category corresponded to simple explanations, based on the context in which the nonliteral form appeared. The third category included elaborate explanations, in which the children proved their ability to distinguish what was said from what was meant in the nonliteral form. By analyzing the responses, researchers found that the understanding of nonliteral forms develops before the metalinguistic skills associated with such forms. In their study, children were first able to explain idioms (at 8 years old), then sarcastic-inference implicatures and semantic-inference conversational implicatures (at around 10 years old). However, none of the groups was able to explain indirect requests. To sum up, this study made it possible to show that the mastery of nonliteral forms was not related to the metalinguistic abilities displayed by children.

This example illustrates the fact that different methods offer different and complementary insights on the same process, and that each of them has advantages and disadvantages. Preference tasks have the advantage of not being based on metalinguistic abilities, unlike the tasks presented above. Indeed, respondents do not need to explain their choices and can simply follow their intuition. These tasks also make it possible to determine the interpretation these individuals prefer among the ones proposed to them. However, they must be combined with other methods for understanding the processes that underlie the comprehension of a stimulus.

4.1.5. *Comprehension tests*

The last type of explicit tasks we will present corresponds to comprehension tests. In these tests, linguistic stimuli – generally in the form of sentences or texts – are presented, followed by one or more questions relating to their content. From a formal point of view, this type of test is very similar to the questionnaire described above. However, we will present it separately, since it aims to measure the comprehension of linguistic stimuli in a more explicit and profound manner than the questionnaire. In fact, comprehension tests make it possible to collect two different types of indicators. Firstly, they can be used to determine what is inferred from sentences or texts. Secondly, they provide a way for measuring comprehension, in terms of answer accuracy. In some cases, open-ended questions may also be offered, in order to observe how responses may vary depending on the variables investigated in the study.

Even though they have been put aside for a while and replaced by online tasks (see Chapter 5), offline comprehension tests are highly informative, in that they allow access to the product of comprehension or mental representations resulting from linguistic stimuli (see Ferreira and Yang

(2019), for an in-depth discussion on the difficulty of assessing comprehension). For example, comprehension tests have shown that mental representations constructed during the reading of a sentence are not automatically correct and complete, but often are simply *good enough representations* (Ferreira *et al.* 2002; Sanford and Graesser 2005). Indeed, various studies have shown that readers do not necessarily process every linguistic stimulus in depth. For example, the question "how many animals of each kind did Moses take on the ark?" is often answered as *two* (Erickson and Mattson 1981), without noticing the fact that it is Noah and not Moses who is supposed to have built an ark.

In order to better understand the conditions that may provoke a relatively superficial processing of sentences during comprehension, Ferreira (2003) examined various conditions. She was interested in the influence of an unusual structure, cleft sentences, as well as the active or passive mode of the sentence. In the rest of this section, we will describe the experiment testing this last variable in detail, in which experimental items describing simple transitive events were presented. Every item was written either in the active or the passive form, and every argument of the verb could either appear as the agent of the action, or as its theme or subject. A third of these items were symmetrical, that is, the relationship between the arguments was plausible in both directions, as in "The man kissed the woman/The woman kissed the man". Another third of the items were reversible, meaning that one arrangement was more plausible than the other as in "The dog bit the child/The child bit the dog". The last third of the items were asymmetrical, that is, the inversion of the elements led to a loss of meaning, as in "The mouse ate the cheese/The cheese ate the mouse".

Ferreira (2003) asked university students to perform a comprehension task, in which every item was presented orally. For each item, the participants had to indicate either the agent or the theme (or subject) of the sentence.

The analysis of the results showed that the participants gave more correct answers when it came to determining the agent of the sentence rather than its theme. Better performances were visible when the sentences were in the active rather than in the passive voice. These results were valid for symmetrical sentences as well as for reversible sentences or asymmetrical sentences. In the case of asymmetrical sentences, performance was also better when sentences described plausible events. These results show that active sentences are easier to understand than passive ones, and that for the latter, it seems more difficult to separate the information coming from the syntactic form than from thematic roles. In addition, these results show that when the sentence conveys improbable content (as in the case of the cheese eating the mouse), people rely on their knowledge of the world rather than on the content of the sentence in order to completely understand it. This can explain why discourse comprehension is not always optimal.

As we saw in the previous example, it is possible to build a comprehension test for investigating a specific research question by developing items that can help us to manipulate the variables of interest, and then to ask questions relating to these items. In order to obtain a measure of general comprehension, it is possible to turn to tests which have already been constructed and validated, that is, tested by other people and discussed in one or several scientific publications. Such tests make it possible to get to know the standards in which the expected results are placed. There are many standardized tests, generally accessible directly via the people who developed them, or through a test library. Examples of such tests are the developmental reading assessment (Beaver and Carter 2019) or the Peabody picture vocabulary test (Dunn and Dunn 2007).

4.2. Implicit tasks

We will now turn to implicit tasks for measuring offline comprehension. Even if some of the above-mentioned tasks can also be considered to be implicit, the tasks we will now present are special, in that (a) they do not directly ask an opinion of the persons tested, and (b) try to access representations or mental processes which cannot necessarily be approached by means of explicit tasks. Implicit tasks make it possible to circumvent some limitations inherent in explicit tasks, such as their strong dependence on the (meta)linguistic abilities of the participants, and the difficulty in explicitly accessing certain processes. For example, this is the case of the processes underlying the organization of the mental lexicon. Implicit tasks are generally associated with the online study of comprehension, as we will see in the next chapter, but they can also be applied offline, in particular, via action tasks, in which the behavior resulting from a linguistic stimulus is observed, or by recall or recognition tasks, as we will see below.

4.2.1. *Action tasks*

In action tasks, the participants have to perform an action on the basis of a linguistic stimulus. This action often involves playing with figurines in order to reconstruct a scene described orally or in writing. This type of task is particularly well suited to demographics such as children and people with language impairments, since comprehension can be measured without the participants having to use language or provide metalinguistic explanations.

An example of such a task is presented in the study by Chan *et al*. (2010) on the acquisition of the canonical subject–verb–object (SVO) transitive word order in children. Chan *et al*. tested three groups of English-speaking children aged approximately 2 years, 2 years and 9 months and 3 years and 5 months old, using an intermodal preferential looking online task, which we will not describe here, and an act-out task. The material for the experiment comprised 10 pairs of plastic animals and six verbs, two of which were familiar verbs (*kick* and *push*) and four were invented verbs (*meek*, *pilk*, *gorp* and *tam*), each describing a specific action. For example, the verb *tam* corresponded to the action of swinging. Before the task, one of the experimenters presented the child with the animal figurines and made sure that the child knew each animal well. Following this, the task itself began. It consisted of six trial sets (corresponding to the different verbs) and each trial set consisted of three phases. During the demonstration phase, the experimenter took a pair of animals and made one animal act on the other by saying "Look! This is VERB-ing!". It is important to observe that during the demonstration phase, the verbs were presented in isolation, the above sentence indicating neither subject nor object. During the training phase, the experimenter then gave the animals to the child and asked them to perform the same action by repeating "Yes, this is VERB-ing!". Then, the experimenter reversed the animals' roles, and the demonstration and training phases were repeated. The last phase corresponded to the testing phase, in which the experimenter gave two new animals to the child and said "Look, the A is VERB-ing the B! Can you do it?" and then waited for the child to complete the action.

The results of the study showed that the older the children, the higher the number of correct answers. The type of verb also played a role, as familiar verbs gave rise to a higher number of correct answers than invented verbs. Analyzing the results per age category, Chan *et al*. (2010) found that children over 3 years old obtained similar results for the two types of verbs and, most of the time, they represented the scene correctly. 2-year-olds, on the other hand, mostly failed to represent the scene (between 32% and 39% of correct answers) for both types of verbs. The intermediate group succeeded in correctly describing the scene 80% of the times when familiar verbs were involved, and 63% of the times with invented verbs.

As we can see in this example, action tasks are implicit in the sense that participants are not directly asked to assess their understanding or to explain a linguistic stimulus. However, the instructions and the procedure are not necessarily implicit: the task of the previous study, as well as the explanations and encouragement by the experimenter, placed the emphasis on a very specific action.

It is nonetheless also possible to manipulate the implicit or explicit nature of instructions in an action task. As for Kissine *et al.* (2015), they studied the comprehension of indirect requests by children with Autism Spectrum Disorder (ASD). Children with ASD have often been said to present a global pragmatic impairment. However, studies have shown that some pragmatic skills are preserved among children with ASD, which suggests that these skills might be related to two different processes: one based on the theory of mind (the ability to draw inferences as to the intentions of others) and the other based on contextual indicators, which can forego this type of inference. As indirect requests highly depend on the context in which they are issued, Kissine *et al.* hypothesized that these should be understood even by people who lack functional theory of mind skills.

In their study, Kissine *et al.* (2015) tested children with ASD aged between 7 and 12 years, and 3-year-old neurotypical children. The control group was chosen so that the children would have similar skills to ASD children, in terms of linguistic development and theory of mind. The experiment in which the children took part was as follows. Every child sat in a quiet room with two experimenters, one of whom interacted with the child, while the other pretended to read a magazine. The first experimenter presented the child with four copies of Mr. Potato Head, a toy made up of a head attached to feet and which can be decorated with various elements such as a nose, eyes, glasses, hat, etc. After presenting the figurine to the child, the latter could add the elements he or she wished. After a certain time and according to the elements already attached by the child, the first experimenter pronounced the sentence “Oh, he doesn’t have a hat/glasses!” In this context, this statement was an indirect request to add a hat or glasses to the toy, whereby the addition of the accessory contained in the request represented a correct answer.

Besides, in order to verify that the child's action corresponded to their comprehension of an indirect request and not to an automatic action based on a linguistic stimulus, two other target sentences were later presented in the experiment. Once the first Mr. Potato Head was assembled, the first experimenter invited the child to create a second one. During the assembly of the second toy, the second experimenter, ostensibly looking at her magazine, repeated the same sentence as the previous one, "Oh, he doesn't have a hat/glasses!" At that time, the first experimenter moved near the second one, looked at the magazine and repeated the sentence again. In these two cases, the sentence should not be interpreted as an indirect request, since it was not directly addressed to the child.

Kissine *et al.* (2015) then coded the actions performed by the children following the different sentences and compared the results between the two groups of children. Children with ASD all responded correctly to the indirect request, whereas the children in the control group were less likely to do so. In addition, children with ASD also adopted proper behavior in response to the second and the third appearance of the target sentence, by not adding an accessory to their toy. The children in the control group had more difficulty not reacting when the sentence was spoken for the third time, which suggests that the task was more difficult for them due to their young age. These results support the idea that children with ASD understand indirect requests based on contextual cues. They also enable us to revisit previous results based on metalinguistic tasks suggesting that the comprehension of such requests was not yet acquired by these children. Once again, this study reveals that different tasks are based on different processes and abilities, and that it is therefore always advisable to combine various approaches.

In addition to not depending on metalinguistic skills, action tasks have the advantage of having good ecological validity, by making it possible to keep the experimental situation as close as possible to a daily-life situation, and to reduce the stress or the apprehension the participants could feel. However, they have an important limitation as to the cognitive skills they require. Since it is necessary to keep the stimulus in mind in order to prepare and to perform the action resulting from it, they can pose problems for populations with memory or executive function deficits, such as people with aphasia, for example. Their use in children can also be problematic for the same reasons. In the experiment conducted by Chan *et al.* (2010), for example, the results obtained in the action task did not show a difference between the conditions in 2-year-old children, whereas such differences did in fact emerge in the online task. As we can see once more, varying the methods seems to be the best solution for overcoming these limitations.

4.2.2. Recall tasks and recognition tasks

Recall tasks and recognition tasks provide access to the mental representations that people construct during language processing. They are based on the assumption that when something is understood on the basis of a linguistic stimulus, this element is encoded and stored in memory. Testing the participants' memory, after reading or listening to a linguistic stimulus, allows us to access their linguistic representations, since the linguistic processing has already been carried out. In this type of task, not only are the correct answers interesting, but so are the errors made. These reveal the similarity between the stimuli or the processes underlying comprehension, as we will see later in the examples.

A recognition or recall task generally takes place in three phases: (a) a first learning phase, in which the linguistic items to be remembered are presented; (b) a break or another task, in order to "empty" the short-term memory; and (c) a test phase, in which the previous linguistic items and other items are presented. During this last phase, the participants have to decide whether the items presented are the same as the ones presented in the first phase, or simply recall the items presented in the first phase. In order to analyze the results, it suffices to determine the number of correct answers for the recognition task. As regards the recall task, it is necessary to decide which answers are considered correct among those produced by the participants.

A classic example of a recognition task can be found in the study by Bransford *et al.* (1972), aimed at determining whether mental representations are exclusively representations of the propositional structure of the sentences, or whether they contain information going beyond this structure. To do this, the authors created 14 scenarios for which it was possible to construct two different situations by manipulating an adverb and a pronoun. The best known example is the one presented below, featuring turtles, a fish and a log. For each scenario, a pair of sentences described the same situation, as in (21) and (22), and a pair of sentences described different situations, as in (23) and (24):

(21) Three turtles rested *on* a floating log, and a fish swam beneath *them*.

(22) Three turtles rested *on* a floating log, and a fish swam beneath *it*.

(23) Three turtles rested *beside* a floating log, and a fish swam beneath *them*.

(24) Three turtles rested *beside* a floating log, and a fish swam beneath *it*.

In the learning phase, a sentence relating to each scenario was presented orally, either (21) or (23). In the test phase, the researchers presented the sentences heard and additional sentences to the participants, who had to indicate which sentence had been presented before, as well as their degree of certainty about their response. When sentence (21) was presented in the first phase, (21) and (22) were presented in the test phase. When sentence (23) was presented in the first phase, (23) and (24) were presented in the test phase. The main point of the researchers' manipulation was the fact that the second sentence of the pair was different from the first sentence at the propositional level, since the final pronoun was modified. This modification of the pronoun, however, only resulted in a modification of the situation described for the pair (23) and (24), but not for the pair (21) and (22). If we build our representations on a propositional basis, we should generally be able to distinguish the sentences presented during the training phase from those added at the testing phase, and no difference should appear between the types of pairs during the recall phase. On the other hand, if we build representations going beyond the text, we can expect more recognition errors between (21) and (22), which describe similar situations, than between (23) and (24), describing different situations. The results corroborated the latter scenario, supporting the hypothesis that our mental representations go beyond the simple content of text or discourse.

Another example, this time relating to a recall task, is a classic study aimed at showing the importance of context for reading comprehension. In this study, Bransford and Johnson (1972) made their participants listen to short texts like the one presented below and asked them to retain as much information as possible from the text so that it could be recalled at the end of it:

> "The procedure is actually quite simple. First, you arrange things into different groups depending on their makeup. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities, that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important, but complications can easily arise. A mistake can be expensive as well. The manipulation of the appropriate mechanisms should be self-explanatory, and we need not dwell on it here. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one never can tell." (Bransford and Johnson 1972, p. 722)

At first glance, this text is difficult to follow and it is difficult to remember it all at once at the end. Now, imagine that before your reading, you received the information that the passage would be about washing clothes. In this case, the text becomes much easier to understand, because your general knowledge helps you to build a context in which to interpret the sentences as you hear them. In one experiment, Bransford and Johnson (1972, experiment 2) separated their participants into three groups. The first group received an indication of the context before listening to the passage, the second group received this information after listening to the passage, whereas the third group received no indication at all. As expected, the indication before listening to the text facilitated the correct recall of the elements.

These examples show that recognition or recall tasks provide implicit access to mental representations as well as to the variables that can influence such representations. It is, however, necessary to observe that the measurement made during such tasks depends not only on the comprehension process, but also on the processes involved within the task itself. For recall tasks, in particular, it has been shown that the first and last pieces of information presented are generally better remembered (Potter and Levy 1969). In this case, it is essential to take this phenomenon into account when setting up the presentation order of the items, by randomizing it, for instance (see Chapter 6 for more information).

4.3. Conclusion

In this chapter, we have presented various offline methods which can be applied to the study of language comprehension. We have shown that these tasks lie on a continuum between explicit and implicit. While explicit tasks test comprehension in a more direct way, implicit tasks do so more indirectly. We have also seen that explicit tasks are often based on the evaluation of linguistic stimuli (e.g. the use of specific questions), most of which require the use of specific metalinguistic abilities, which can hinder their use with certain populations. Implicit tasks provide a way of circumventing this problem, by examining comprehension in a roundabout way. Throughout this chapter, we have also argued that different methods can lead to different results, and that it is necessary to combine different methods in order to benefit from the advantages of each one, while going beyond their limitations.

4.4. Revision questions and answer key

4.4.1. *Questions*

1) If you had to choose between a metalinguistic task and an action task, in your opinion, which method would be the most suitable for studying the comprehension of syntactically complex sentences in people suffering from aphasia?

2) What is the difference between a recognition task and a recall task?

3) Two researchers want to set up an acceptability task. The first one wishes to use a YES/NO scale, whereas the second one wishes to use a scale from 0 (not at all acceptable) to 5 (completely acceptable). What could be the arguments by each of them?

4) Researchers wish to study the comprehension of irony in L2 in beginner learners. Which of the methods presented in this chapter do you find most appropriate?

5) What is the point of offline comprehension tests?

6) How could the comprehension test carried out by Ferreira (2003) (see section 4.1.5) be applied to preschool children?

4.4.2. *Answer key*

1) The different forms of aphasia are characterized by a difficulty in producing language. This can be almost complete in cases of global aphasia or limited to certain aspects of language production in other cases. Different processes are required from participants when carrying out metalinguistic tasks, such as consciously accessing their intuitions about language and succeeding in expressing them. People with aphasia suffer from certain deficits, which can greatly interfere with the processes required for carrying out metalinguistic tasks. For this reason, it may be more suitable to test them using action tasks, where there is no need to resort to language. On the other hand, action tasks require good memory abilities, such as working memory, which can sometimes also be affected in people with aphasia. It would therefore be necessary to make sure that these abilities are properly preserved in the participants before implementing an experiment based on an action task.

2) Recognition and recall tasks both examine the mental representations developed during language processing, by testing the memory of the participants after a learning phase. The difference between these two methods lies on how the memory is tested. During a recognition task, the stimuli of the learning task are presented in parallel with new stimuli, and participants have to make a distinction between the stimuli already presented and the new stimuli. In a recall task, no stimulus is presented, and participants have to recall as many elements as possible.

3) The first researcher, wishing to use a binary YES/NO scale, could argue that the different acceptability scales are all as informative as one another. On this basis, the participants' evaluation task could be simplified by offering them only two choices, acceptable and not acceptable. The researcher could add that it is difficult to evaluate a degree of acceptability,

in the sense that a statement contains a grammatical error or not and is semantically relevant or not. Finally, they could question the usefulness of a six-category scale, from which definitive conclusions cannot necessarily be drawn based on the difference between categories.

The second researcher, who wishes to use a six-point scale, could argue that there may be different degrees of perceived acceptability, depending on the importance of the linguistic aspects manipulated in the experience and the participants' intuitions. For example, a simple sentence containing a grammatical error could be considered completely unacceptable, whereas a complex sentence containing the same error could be considered partially unacceptable (since there would be proportionately more correct elements in the sentence). This researcher could also argue that reducing the participants' response to YES/NO could force participants choices, whereas their responses may be more nuanced. As regards the number of points used on the scale, the second researcher could agree with the opinion of the first one and propose a scale containing more points, which could subsequently be considered as a ratio scale for the analyses.

4) The answer to this question depends on the research question examined, which, in this case, is the comprehension of irony in a second language. An adequate task for this type of question could be a metalinguistic task, in which the participants should report their interpretation of an utterance. Another possible task could be a preference task, where the participants could show their comprehension of the irony of the utterance by choosing a response as in the experiment by Bernicot *et al.* (2007) (see section 4.1.4). An action task based on the utterance could also be implemented. The second element to take into account when answering the question is the language in which the study is carried out. If this is done in the language that people are learning, it is necessary to take into account that it can be difficult for them to express themselves in this language. It is also possible that their present knowledge of the language does not yet allow them to understand or to convey complex ideas. In this case, the use of certain metalinguistic tasks is compromised, and it would be more appropriate to turn to preference or to action tasks. If the experiment is run in the participants' mother tongue, then it would be possible to examine their comprehension of irony by means of metalinguistic tasks.

5) Offline comprehension tests provide access to the content of mental representations that people build during language comprehension. In other

words, they make it possible to observe the elements that people retain and what they have really understood when processing a linguistic stimulus. The use of such tasks enables examination of the influence of specific variables on comprehension. Not only do they contribute to the evaluation of comprehension skills in general, but also to the definition of groups of people with good or not so good comprehension skills, on the basis of standards validated by other researchers.

6) In her study, Ferreira (2003) examined the influence of the active or passive mode of the sentence on comprehension. Each sentence had two elements, one appearing as an agent and the other as a theme. After each sentence, the participants had to indicate either the agent or the theme of the sentence. In order to adapt this experience to preschoolers, rather than asking them to indicate the agent or the theme of the sentence, it might be more appropriate to implement a preference task. In this case, we would present them with two images for each sentence, staging its elements, and whose arrangement might correspond to the statement in question or not. For example, for the sentence *The cat is chased by the mouse*, one image could show a cat chased by a mouse, whereas the other image could show a mouse being chased by a cat. Children should then simply indicate which image corresponds to the statement. It would also be possible to implement an action task, by asking the children to reproduce the content of the statements presented to them with the help of figurines.

4.5. Further reading

For a detailed presentation of acceptability judgments, see Schütze and Sprouse (2013). Rasinger (2010) and Wagner (2015) develop different aspects related to the creation and use of questionnaires. The book by Gonzalez-Marquez *et al.* (2007b) offers numerous examples of the application of the techniques discussed in this chapter to the field of cognitive linguistics. For a comparison of the pros and cons of offline and online measurements for testing comprehension, see Ferreira and Yang (2019).

5

Online Methods for Studying Language Comprehension

In this chapter, we review a second type of methods used for the study of language comprehension, based on *online* measurements. These methods provide access to the processes involved in language comprehension the very moment they take place. As comprehension processes are not directly observable, it is necessary to rely on indirect indicators for measuring them. Such indicators can be obtained through tasks where participants are asked to report a piece of information concerning the linguistic stimuli in process, such as verbalizing their thoughts while they process a stimulus. However, verbalization tasks have important limitations, which is why most online methods tend to be based on the time required for completing certain tasks. We describe how time can be used for signaling comprehension, and then we present various online tasks for which reading time is the central measurement.

5.1. Think-aloud protocols

A *think-aloud protocol* is an introspective method in which participants are asked to report their thoughts, either as they unfold, or after reading or listening to a text. Verbalization may refer to spontaneously developed thoughts by participants, or to the justifications or explanations requested by researchers. Verbalization can thus be used for assessing the metalinguistic or non-metalinguistic aspects of comprehension, depending on the instructions given. It is also possible to categorize reported thoughts following previously defined criteria, in order to transform them into

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

quantitatively measurable data. For this classification to be feasible, it is necessary to give precise instructions to the participants about the type of thoughts that should be reported, or the moment when these thoughts need to be expressed.

For example, Blanc *et al.* (2008) studied the conditions in which readers update the mental representations constructed during reading, as new information is provided in the text, sometimes contradicting what has already been presented. As we have already discussed in Chapter 4, readers build the comprehension of a text by developing a mental representation which contains not only the information transmitted by the text itself, but also inferences. These are deductions based notably on world knowledge. In order to adequately reflect the content of the text, mental representations have to be continuously updated, as new pieces of information are continually brought in by the text. This update may correspond to the addition of information to an existing mental representation, but in other cases, it may require revising already formed mental representations and inferences, as is the case when contradictory information emerges.

Blanc *et al.* (2008) investigated the updating of mental representations constructed while reading newspaper articles reporting dramatic events. They put together six experimental articles, all following the same structure and presenting two plausible causes for the dramatic event. In each text, a critical sentence provided elements either in favor of the first cause or in favor of the second one, or else, neutral elements in relation to these causes. Six other items were similarly constructed, but presented only one plausible cause for the dramatic event. The participants of the experiment had to read the 12 articles, sentence-by-sentence, and answer a question orally: “At this moment, what comes to your mind about the information you have just read?” This question arose every two sentences of the text, as well as after the presentation of the critical sentence, and then again at the end of the text.

The responses were coded depending on whether or not readers mentioned one of the two causes appearing in the text as a reason for the occurrence of the dramatic event. When both causes were mentioned, the existence of a relationship between the causes (or not) was coded as well. Finally, any mentions of comprehension difficulties were also recorded. The analysis of the responses given after the presentation of the second cause showed that the two causes presented in the text were generally taken into account. In addition, the participants were aware that the causes changed as the text progressed, that is, the event was eventually explained by the latter cause, but had earlier been explained by a different cause. Furthermore, when an alternative cause was presented, participants mentioned that they had probably misunderstood information presented earlier in the text, or expressed criticisms as to the way the report had been written. These reactions showed that the emergence of a second cause prompted them to update their mental representation of the text.

An analysis was then carried out on the answers given directly after the critical sentence, which strengthened one cause or the other, or did not strengthen any cause, as well as on the answers given at the end of the article. When the critical sentence strengthened the first cause or transmitted neutral content, the participants included the two causes in their responses, showing that the two had been activated in parallel in their mental representation. On the other hand, when the critical sentence strengthened the second cause, the participants mentioned this cause more than the first one in their responses. These results confirm that the order in which information is presented, rather than the number of times presented, influences the mental representations of readers. Actually, when the first cause was strengthened, the two causes were activated in the memory of the participants, whereas when the second cause was strengthened, only the latter was activated in the end. This can be explained by the fact that the strengthening of the second cause took place directly after its presentation, encouraging participants to forget about the first cause and to accept the second one. This explanation could be corroborated by observing the difficulties reported by the participants. They found the texts more difficult to follow when the critical sentence strengthened the first cause or none of the causes, than when it strengthened the second one.

This example illustrates the interest of think-aloud protocols for the study of processes taking place during comprehension. By observing the thoughts transmitted by the readers as they progressively read a text, it is possible to observe the steps followed by the participants, as well as the changes in their representations. It would not be possible to access these processes only by observing the final representation, as with offline tasks discussed in the previous chapter, or through the use of other online tasks that we will present in the rest of this chapter.

Think-aloud protocols have, however, significant limitations. First, verbalizing our thoughts whilst reading requires high cognitive skills, in order to be able to become aware of our train of thought and to verbalize it. Moreover, we cannot dismiss the fact that the very nature of this task – having to consciously access our thoughts and report them – influences the natural reading process or interferes with text comprehension. Indeed, expressing our thoughts while reading involves distancing ourselves from the text during the time required for expression, before returning to the text. This distancing can compromise the construction of the representation or can add elements which would not have been present if the participants had only read the text. Finally, this task is only appropriate to study those processes which are accessible to consciousness and can be reported on a voluntary basis. For all these reasons, think-aloud tasks are not the best method for studying online processes, and their use has been marginalized since the development of other time-based methods. The rest of this chapter is devoted to them.

5.2. Using time as an indicator of comprehension

The cognitive processes involved in language processing are extremely fast and most of the time, inaccessible to consciousness. In order to study them, it is necessary to access them in an indirect way, by observing indicators or signals of the processes, as explained in Chapter 2. A very commonly used indicator in experimental linguistics – and in cognitive science in general – is the time required to complete a task. The use of this indicator is based on the idea that the time required for processing a linguistic stimulus reflects its degree of difficulty: the more complex the stimulus and/or its processing, the longer processing time it requires.

The term "processing" refers here to the different stages and sub-processes involved in comprehension or production of a linguistic stimulus. This term is deliberately vague, because it depends on the specific aspects of language processing targeted by the different tasks in which time is the dependent variable. Such tasks can be divided into two broad categories. A first category groups the tasks in which the participants must react to a stimulus. In this case, reaction time is measured. A second category concerns the reading tasks themselves. Here, the focus is placed on the time it takes for text segments to be read. Importantly, the time measured does not reflect the same type of processes, depending on the task involved. For example, the reading time of a simple naming task could reflect the time required for deciphering the word, building its phonological representation and then being able to pronounce it aloud. The reading time of a sentence placed in the middle of a text reflects the time needed to decipher every different word, to connect words and to add incoming information to the mental representation already constructed on the basis of previous sentences. Reading this new sentence also potentially involves the derivation of inferences or the need to revise a prior mental representation.

Differences in response times have been documented by studying multiple variables, such as the length of the linguistic stimulus, for example: short words are processed faster than longer words. Another factor is word frequency: frequent words are processed faster than rarer words. Syntactic complexity also seems to play a role: simple sentences are processed more quickly than complex ones (Just and Carpenter 1980; Rayner 1998; Smith and Levy 2013). The processing time of a stimulus makes it possible to deduce information not only about the complexity of the linguistic stimulus itself, but also about the number and the dynamic organization of stages involved in the processing of the stimulus.

In a typical experiment aimed at measuring reaction or reading times, the participants see linguistic stimuli, namely words, sentence fragments or complete sentences, and must perform an action based on these stimuli. As we will see later, these actions may be varied such as deciding whether a string of letters corresponds to a possible word, saying whether a certain word was present in a sentence or even simply reading a sentence. Stimuli are presented on a computer screen by means of experimental software and reading time is measured by asking the participants to indicate their response by pressing a key on a keyboard. In order to get a more precise measurement, it is possible to use a *button box*, which allows us to measure the response time with a millisecond accuracy. It is also possible to ask participants to respond orally by using a microphone and a voice key for collecting the audio signal and automatically recording the response and the response time. Response time generally corresponds to the time between the initiation of the stimulus and the participants' response.

The use of response time as a dependent variable requires following certain methodological principles. We will approach these from a theoretical point of view, before taking them up again in the form of practical advice in Chapter 6.

In general, in a task measuring response times, participants can choose between two possible answers: YES and NO, for deciding, for example, whether a string of characters corresponds to an existing word or not. The processes underlying YES and NO responses are different, and for this reason, it is necessary to associate them with different motor responses (see, for example, Rossi (2008)). Typically, YES responses are associated with the participants' dominant hand, whereas NO responses are associated with the other hand. In order to accurately measure the response time, it is also necessary for the participants to constantly keep their fingers on the response keys during the experiment, so as not to add extra time for identifying the keys. Participants are generally invited to sit in front of a screen, their fingers placed on the keys, and are asked to remain in this position throughout the experiment.

Another important methodological point to understand is that response time *per se* provides little information about the processes underlying comprehension. Its contribution to the study of an independent variable requires a comparison of at least two conditions, one where the independent variable is present and one where it is absent. In this case, we refer to a subtractive method, in which the difference in the response time between the two conditions reflects the extra time needed for processing a particular type of stimulus or for carrying out a certain process. For example, measuring the time a person takes to read the pronoun *she* in a sentence would not contribute to drawing any conclusion. In contrast, comparing the reaction time for this pronoun in two different contexts, for example, with reference to *secretary* and to *astronaut*, could shed some light on the processes involved when reading the sentence. Here, the reaction time for *she* would probably take longer after *astronaut* than after *secretary*, showing that readers deduce gender-information about a character based on stereotypes pervading society. Firstly, this highlights the need to clearly define the process examined, so as to construct experimental conditions effectively isolating such process. Secondly, it shows that it is essential to choose the appropriate task for measuring the process as directly as possible.

Let us first focus on the need to isolate the process we want to study. Comparing response times between two conditions implies that such conditions must be similar in all respects, apart from the manipulation of the independent variable. It is essential to compare only that which is comparable. Let us go back to the effect of word frequency on processing time, already presented in Chapter 2. There, we saw that it was necessary for the frequent and the less frequent words used in the experiment not to differ on other criteria, such as their length, in order to reliably assess the effect of frequency itself. Thus, the most favorable situation would involve having the same linguistic stimuli repeated across different conditions. In this case, this would be unattainable, since a word cannot simultaneously be very frequent and infrequent. Actually, the different stimuli should be similar on as many points as possible in order to prevent confused variables from jeopardizing the validity of the experiment. All the variables susceptible of influencing response times, such as word length, word frequency, their concreteness, their grammatical category, their position in the sentence and their contextual predictability, should be kept equivalent across the different conditions.

Let us now turn to the choice of the task used to collect response times. This choice is essential, since the processes involved in the different conditions should allow us to reveal the influence of the independent variable. For example, let us imagine a study seeking to determine whether it is faster to indicate an answer on the keyboard using the dominant hand or the other hand. For this study, it would be necessary to build an experiment in which the participants answer half of the time with their right hand, and the other half of the time, with their left hand. For instance, the task could involve verifying simple operations by indicating whether the result is correct (using the forefinger) or incorrect (using the middle finger). In this case, the task of the participants would comprise several stages, such as deciphering the figures and symbols appearing on the screen, resolving the operation, comparing the result with the one shown on the screen, deciding the answer, choosing the finger for pressing the corresponding key and finally pressing the key. Reaction time would also reflect all of these steps. This is why we speak of choice reaction time (or complex reaction time) in cases like this: the response depends on a choice made by the participant. Using choice reaction time for the question we are analyzing might pose two potential problems.

The first problem relates to the fact that the different stages involved in the response should be equivalent for both conditions (right hand or left hand), especially on how difficult the operations are. In order not to threaten the validity of the experiment, all operations should be kept even in terms of the difficulty level. The second problem stems from the large number of steps involved in the task, which can prevent the detection of the desired effect. Indeed, the contribution of the left hand/right hand activation to the reaction time is moderate (see Figure 5.1 for an illustration), as it only represents one stage among others, and is not the longest one to complete. In addition, the hypothetical time difference between the answers given by the right hand and those given by the left hand is probably small. By using this type of task, there is a risk of drowning the effect in the combination of processes involved in the task. In order to study the reaction speed of the two hands as directly as possible, we should aim for the simplest possible experiment, in which the number of processes involved should be kept to a minimum. For example, we could simply present dots on the screen and ask participants to press a single key as quickly as they can when a dot appears. This simple task would make it possible to obtain a more direct measurement of the influence of the hand employed over the reaction time, without tainting the response with incidental processes. As we can see in Figure 5.1, the process of pressing the key would have a more important weight over the reaction time. In this case, we would speak of simple reaction time, since the task would only aim at giving one single response when a stimulus appears.

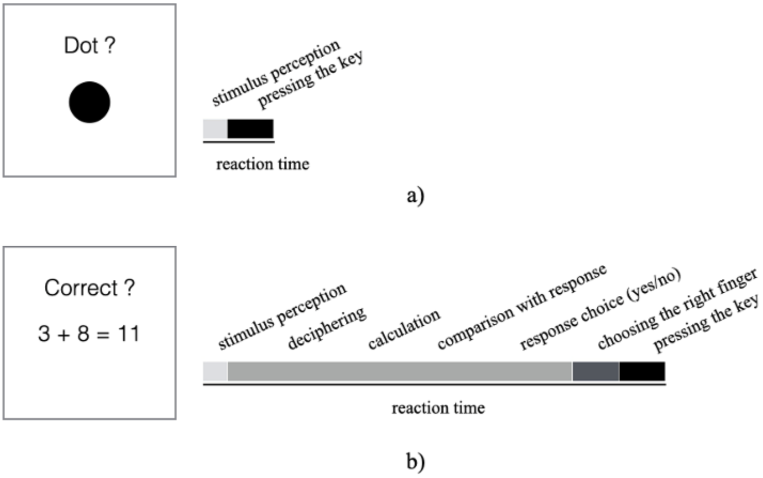


Figure 5.1. *Example of fictitious steps involved in simple (a) or choice (b) reaction times*

Unlike the fictitious tasks presented above, the tasks that we will describe in the remainder of the chapter aim to ensure that the participants cannot guess the purpose of the experiment, so as not to compromise its validity. There is always the risk that the participants' responses can become unnatural as soon as they guess or believe they guess the hypothesis under examination.

Using time as a dependent variable could apply to most offline tasks described in the previous chapter. A possibility would be to measure the time needed for providing the answers in the different tasks we presented, without changing the instructions given to the participants. Measuring time would offer additional information to that already offered by the responses themselves. For example, in an acceptability judgment task, it would be possible to observe that people give similar acceptability scores to different stimuli, but that evaluating some of these stimuli may require a longer amount of time than others. In this case, it could mean that some of the stimuli were more complex. The reason for this greater complexity has yet to be defined, but the response time could act as an indicator that the stimuli or the processes involved in their comprehension differ, despite the similarity of their acceptability score.

In the rest of this chapter, we present a series of tasks aimed at studying online comprehension, for which the response time is at the center of the paradigm, as much as the content of the answer provided. In these tasks, a certain time constraint is placed on the participants, unlike the above-mentioned tasks. This time constraint aims to ensure that the desired processes are measured, regardless of other processes the participants could engage in. To do this, the participants are instructed to respond as rapidly and as accurately as possible. Indeed, without any time constraints, we could not rule out that the participants maximize the number of correct answers by taking their time to answer, or on the contrary, minimize their response time by not paying attention to the accuracy of their answers (speed-accuracy trade-off). By asking the participants to focus on both speed and accuracy, we try to avoid these phenomena, and encourage the participants to respond correctly, and this, in the shortest possible time. Online tasks often rely on a priming mechanism, which is described in section 5.3.

5.3. Priming

The priming effect occurs when a first stimulus (called the prime) is presented and influences the processing of a subsequent stimulus, the target. Priming is said to be positive when the presentation of the prime decreases the target's processing time, and is said to be negative when the presentation of the prime increases the target's processing time.

Priming can be explained by the fact that concepts are stored in networks within the memory, on the basis of shared properties. Words, for example, are concepts that are connected by their semantic and phonological properties within the mental lexicon. The presentation of a prime activates all of the properties associated with it, and this activation spreads to other concepts sharing similar properties. When the target appears, some of its properties are pre-activated, which makes it easier to process. For example, seeing a prime picture of a swan facilitates access to the word *duck*, compared to the word *dog*, since there are more similarities between a duck and a swan, than between a dog and a swan.

Priming effects are investigated by means of tasks presenting a prime and then a target for which the participants have to provide an answer. In general, the target is shown until the answer is given, but the maximum duration of its presentation time can also be defined beforehand. The answers can be of different types, depending on the task the participants have to carry out. The most common tasks are evaluation, lexical decision or naming tasks, which are presented later in this chapter. Primes can be shown in such a way that participants can see them and process them consciously, or on the contrary, they can be presented for a very short amount of time, so that they are only perceived subliminally. It is also possible to control the capacity of processing the primes by presenting them either in an isolated manner or preceded or followed by a mask (i.e. visual noise, e.g. #####, that blocks their processing). Finally, it is possible to vary the time between the presentation of the prime and that of the target, in order to study the time frame related to the processing of the primes and its influence on the processing of the targets.

The central manipulation in priming tasks is the relation between the prime and the target. This relation can be semantic, phonological, syntactic or even affective. In all cases, it is necessary for the different types of primes to differ only in terms of the variable examined. For this reason, rather than comparing the presence or the absence of a prime on a process, it is recommended to compare the presence of one type of prime and its absence (but a prime, altogether). Later on, we will provide illustrations for this. If a priming effect is present, we can expect to obtain faster response times for the targets in the priming condition than in the control condition.

The priming effect is central in experimental linguistics, because it can shed light on many processes. For example, this effect has allowed us to better understand the organization of the mental lexicon, as well as the way in which we access words during language comprehension and production. Numerous studies have indeed shown priming effects for semantically, orthographically or phonologically related word pairs (e.g. Ferrand and Grainger (1992, 1993) and Dell'Acqua and Grainger (1999)).

We will now turn to the different tasks for testing the priming effect, or more generally, involving a priming mechanism.

5.4. Lexical decision tasks

In a typical lexical decision task, combinations of letters corresponding to words and non-words (sometimes called pseudo-words) are presented, and participants are asked to indicate whether the stimuli are words or not, as rapidly and as accurately as possible. These answers are given by pressing one of two keys at their disposal, representing YES or NO responses. This type of task implies that half of the items presented are words and the other half are non-words.

Lexical decision tasks can be used to study a wide variety of processes. In the field of word recognition, for example, these tasks allow for the manipulation of many variables, such as their visual, phonological or semantic properties, in order to determine the importance of such properties in the recognition process. Lexical decision tasks have notably revealed one of the most robust effects in the field of reading, namely the frequency effect. The frequency of a word corresponds to an estimate of the number of times a person has encountered this word, and is calculated on the basis of the number of occurrences of this word in a corpus. The higher the word frequency, the easier it is to categorize it as a word (e.g. Baayen *et al.* (2006)).

Most lexical decision tasks involve a priming process. Through this method, it has been possible to demonstrate the phonological priming effect which we have already mentioned. As an example, we will present the study by Carreiras *et al.* (2005) that revealed certain characteristics of phonological activation during silent reading and whilst reading aloud. Numerous studies have shown that the phonological information associated with words is activated during silent reading (Ferrand and Grainger 1993; Ziegler *et al.* 2000; Drieghe and Brysbaert 2002) thanks to tasks that combine monosyllabic primes and targets. Since they contain only one syllable, these primes and targets can be phonologically very similar. Carreiras *et al.* extended the study of the role of phonology to the reading of bisyllabic words, in order to examine two central questions. First, determining whether phonemes and syllables are decoded sequentially or in parallel. Second, exploring whether it is possible to obtain phonological priming only when the overlap between primes and targets is partial.

To answer these questions, Carreiras *et al.* (2005) chose 120 French words as well as 120 bisyllabic non-words (five to eight letters long) as targets. Then, they created non-words which would serve as primes for the two types of targets. In relation to the first question, the focus was placed on priming the first syllable for half of the targets, and the second syllable for the other half. In order to answer the second question, primes were divided into three categories. The first category corresponded to a prime containing a phonologically similar but orthographically different syllable from the one examined in the target. For example, *fomie* appeared before *faucon* in the case of the first syllable, and *retôt* appeared before *gâteau* in the case of the second syllable. The second category of primes contained the same first phoneme (and grapheme) as the target syllable of interest (*fémie* before *faucon* in the case of the first syllable, and *retin* before *gâteau* in the case of the second syllable). The third category was neither phonologically nor orthographically related to the target (*pémie* before *faucon* and *redin* before *gâteau*).

Each target, whether a word or a non-word, and whether a first-syllable or a second-syllable target, was associated with a prime in each category. Each participant saw 20 targets of each type in each priming condition, in order to ensure that every participant saw all items and all conditions without being presented with the same item more than once. We will return to this notion of groups of items, also called lists, in Chapter 6.

Each target was presented in a random order, and the same procedure was applied to all the different items. Participants were instructed to indicate as rapidly and as accurately as possible whether or not the letter string was a French word. For each of the 240 tests, a mask (XXXXXXXXXXX) appeared on the screen and remained there for 500 milliseconds. Immediately afterwards, the mask was replaced by the non-word prime which remained on the screen for 59 ms, too short a time-lapse for the participants to process the prime consciously. Then, at the end of the 59 ms, the prime was immediately replaced by the target, which remained on the screen until the participant responded.

Analyses were performed on the correct response times for the target words, depending on whether the first or second syllable had been primed. For first-syllable targets, reaction times were faster in the phonological-priming condition than in the first phoneme and unrelated priming conditions. These last two conditions did not differ from each other. No effect, however, appeared for second-syllable targets.

These results showed that – at least in a lexical decision task – phonological priming occurs when the primes have only a partial phonological overlap and the orthographic overlap with the target is minimal. Moreover, the fact that phonological priming only appeared for first-syllable targets supports the idea that phonological processing during reading is sequential.

Lexical decision tasks can also be used for studying the content of mental representations constructed while reading, especially the inferences generated by readers. In this case too, studies are based on the presence of a priming effect, more specifically on the fact that the activation of a concept in the readers' memory should be transmitted to the associated concepts and make them more accessible. On this basis, it is possible to present texts requiring the derivation of inferences and to test whether an inference has actually been generated by making it appear at different places in the text, in the form of a lexical decision task. The necessary time for responding is supposed to reflect the activation that the concept received while reading the text.

De Vega *et al.* (1997) used such a lexical decision task to study the ability to infer a character's emotion by adopting their perspective. The authors created short stories in which the main character was aware of a piece of information (or not) which should have influenced their emotional state. For example, one of the stories featured a woman waiting for her partner who was very late for an appointment. The story elaborated that her partner was inundated with work and that she would not want to put even more pressure on him. In any case, the woman finally decided to call him at home and she was told (in the informed condition) or wasn't told (in the uninformed condition) that he was playing poker with his friends. The story continued with a neutral sentence explaining that the woman was thinking about her partner and ended with a concluding sentence.

Such a story can activate different representations about the protagonist's emotion. On reading the first sentences, readers will likely infer that the woman feels sympathy towards her partner. He works a lot, and she doesn't want to put any additional pressure on him. The protagonist's emotion should remain the same in the uninformed condition, as she is unaware of the fact that he is not meeting her because he is spending time with his friends while she believes him to be at work. However, in the informed condition, the protagonist's emotion should no longer be sympathy but a shift to anger, as soon as she realizes the real reason for her partner's absence.

If the readers infer the characters' emotions by taking their perspective, then the concept of *sympathy* should be activated in their mental representation in the uninformed version, and *fury* should be activated in the informed version. In order to test this, de Vega *et al.* (1997) asked participants to read the stories sentence-by-sentence and to complete a lexical decision task. This contained an adjective describing the initial emotion (*sympathetic*) in one experiment and the other emotion (*furious*) in another experiment. Target words appeared at the end of the neutral sentence. The results showed faster decision times for the initial emotion than for the alternative emotion in the uninformed condition, and the opposite results for the informed condition. These results support the initial hypothesis that readers adopt the characters' perspective during reading.

In summary, lexical decision tasks have the advantage of being very easy to implement thanks to the use of experimental software that allows the presentation of stimuli, the recording of responses and of reaction times. However, they present a significant limitation, in that the decision process underlying responses can be assimilated to a categorization process (*yes, it's a word* or *no, it's not a word*) along a lexical familiarity continuum (Ferrand 2001). In a simplified manner, the participants in a lexical decision task can set up strategies in order to classify the strings of letters presented. These strategies depend on various variables, such as the familiarity of the words and the non-words shown, their phonological or orthographic characteristics, or even the instructions given. As a consequence, they may influence the responses and reaction times obtained in the experiment, which may actually depend on different factors from the ones we wish to investigate. It is therefore important to take this limitation into account when choosing the items for a lexical decision task.

5.5. Naming tasks

Naming tasks offer an alternative to the inherent limitations of the categorization process involved in lexical decision tasks. These tasks are extremely simple: participants have to pronounce a word, either immediately after its appearance or in a delayed manner. The word to be pronounced is presented on a screen, and the naming time (i.e. the time between the presentation of the stimulus and the start of the response) as well as the response itself are recorded. This method makes it possible to calculate the correct/incorrect response rate, which can then be analyzed in parallel with response times.

Naming tasks are based on perceptual and on production processes. This is why they measure various processes related not only to silent reading, but also to the production and articulation of words. For this reason, the risk of this method is that is does not allow one to distinguish the role played by those different processes in the response time. A first possibility to solve this problem would be to set up a delayed naming task, in which the response does not immediately take place after the presentation of the word, but after a certain delay. By comparing the latencies obtained from a delayed task with those drawn from an immediate task, we can determine whether the differences between conditions stem from processes related to reading or to naming.

A second possibility would be to add another task to the naming task, in order to confirm the effects using a different technique that does not present the same advantages and limitations. In the literature, we often find a naming task combined with a lexical decision task. As an example, let us refer back to Carreiras *et al.* (2005) and their study on phonological priming. In their paper, they presented a second experiment, similar in all respects to the one presented above, but this time using a naming task. The results confirmed those obtained with the lexical decision task and showed an additional effect, namely a shorter response time when the target started with the same grapheme as the primer. According to the authors, this new effect can be explained by the fact that the naming task involves an articulatory process which is not present in the lexical decision task. This example illustrates the complementarity of approaches, which we have already emphasized several times in this book, as well as the need to verify the results by using different methods. The particularities of each method can reveal effects which would not emerge using other techniques.

Another example of the application of naming tasks concerns the field of discourse and inferences. One type of inference, predictive inference, involves the activation of information about predictions that can be made based on the text, such as the consequences of an event described in the text, or about future events. Lassonde and O'Brien (2009) investigated the specificity of predictive inferences depending on the contextual support conveyed by the text. According to their hypothesis, the more information a text contains about a specific inference, the higher the probability of observing such an inference, and of increasing its specificity. The specificity of the inference was operationalized as the number of activated lexical items following the reading of a text. If an inference is specific, few items should match it. On the contrary, if the inference is rather general, then more items should be activated.

In order to test this idea, Lassonde and O'Brien (2009) developed three experiments. The first one aimed to test the assumption that predictive inferences are generally not specific, and can include different lexical items. In order to investigate this assumption, participants had to read short stories and perform a naming task at the end of each story. The stories were presented in such a way as to trigger the development of a predictive inference, or not. For example, one story described a young boy, Jimmy, playing with local children throwing rocks at a target. Then, the story continued with the manipulated sentence, which could either activate a predictive inference, "Jimmy missed the target and he accidentally hit the door of a new car", or transmit some content not activating this type of inference: "a dog came racing across the street and distracted Jimmy from his throw". Immediately after this sentence, participants saw one of the two possible target words (*dent* or *damage*) appear on the screen, and had to pronounce it. Target words were determined before the experiment so as to be the most likely to reflect predictive inferences related to the text. We can nonetheless appreciate that the second target word is less specific than the first one. The results showed that for the two target words, the naming times were faster in the inference condition than in the control condition.

Another experiment aimed to verify that as contextual support increases, the predictive inference becomes more specific and the number of lexical items activated diminishes (Lassonde and O'Brien 2009, experiment 3). The same texts were used after their introduction had been slightly modified, and this time contained information aiming to encourage one of the two target items, compared to one another.

For example, Jimmy's story stressed the fact that the brand new car of his friends' family did not have any scratches or blemishes, and that children should be careful not to damage it. The task was the same as in the first experiment, as were the target words presented. The results showed there was only a difference in response times between the inference and the control conditions for the most specific target word (*dent*), and that there was no difference for the other word (*damage*). This result suggests that contextual support influences the specificity of predictive inferences. These inferences are captured by fewer lexical items, as the information transmitted by the text progressively constrains them.

5.6. Stroop task

The Stroop task draws its name from the *Stroop effect* (Stroop 1935), well known in psychology, which illustrates the phenomenon of automatic access to meaning while reading a word. In a classic Stroop task, participants must name the color of the ink that is used to write down the name of a color, such as, for example, the word *blue* being written in blue or in red. Naming times are slower when the color of the word is incongruent with the name of the color, thus reflecting the impossibility of preventing reading a word and accessing its meaning. In other applications of this task, the words presented in color may refer to concepts other than a color itself (see the example below), or to meaningless strings of letters or symbols. In all cases, naming times in the incongruent condition (*sun* written in blue, for example) are slower than in control conditions (*xxx* written in blue), which correspond to an *interference* effect associated with the Stroop task. But this effect is not the only one at work in a Stroop task. There is also a *facilitation* effect, whereby naming times in the congruent condition (*sun* written in yellow) are also faster than in the control condition (*xxx* written in yellow) (Augustinova *et al*. 2016). From a methodological point of view, it is therefore very important to choose the relevant control items.

The Stroop effect is useful for studying different aspects of language. Regarding the lexical access while reading, various studies have questioned the automatic semantic encoding postulated by the results of the classic Stroop task described above. By modifying the types of items presented, Besner *et al.* (1997) showed that semantic activation is not automatic during a Stroop task. As a matter of fact, the interference effect decreased when only one letter of the color name was colored. This effect also diminished if the control words were non-words (in this case, the reading of the words becomes less relevant for the participants) and even disappeared completely when a single letter of the non-words was colored. These results led to the conclusion that the automatic processing of words and access to meaning depend on specific conditions, and that it is possible to not activate such processes.

Another possible application of a Stroop task is found in the study by Eilola *et al.* (2007), who investigated the processing of emotional words in the mother tongue (L1) and the second language (L2) of bilingual participants. In order to study this question, an emotional Stroop task was implemented. In this task, positive, negative and neutral emotional words are presented in colors. Participants simply had to indicate the color of the word. This task has generally shown that negative and taboo, emotional words are treated differently from neutral words, as their response times are generally longer (McKenna and Sharma 1995; Williams *et al.* 1996). The mechanism underlying the interference observed in the emotional Stroop task is not yet clear, and different contending explanations have been proposed in the literature (see, for example, Algom *et al.* (2004) and MacKay and Ahmetzanov (2005)).

In order to extend these findings to the different languages of bilingual subjects, Eilola *et al.* (2007) recruited advanced Finnish–English bilinguals and presented them with 80 words (20 positive, 20 negative, 20 taboo and 20 neutral) written in four colors, namely yellow, red, blue and green (once per color), in the two languages spoken by the participants. The items were presented in blocks, grouping the items of a single condition and in a single language. Participants thus saw eight blocks, one for each type of item, in one language (L1) and then in the other (L2). We will return to this notion of a block in Chapter 6. The results showed that the participants named the color of negative and taboo words more slowly than that of neutral words. Contrary to what had previously been shown at the discourse level, the effects were similar in L2 and L1, suggesting that the processing of emotionally negative words in L2 does not differ from that in L1, at least among proficient bilinguals.

5.7. Verification task

We now turn to a task specially designed for studying the comprehension of sentences and texts: the verification task. In this task, participants are presented with sentences or short passages, and then asked to indicate whether an item was present in the text or not. The item in question can be of different types, such as a written word or a picture, with specific properties. The nature of the item depends on the research question, as is illustrated below. The response and the response time are recorded in the same way as in the tasks previously discussed.

This type of paradigm makes it possible to investigate the nature of representations constructed while reading. As we have already discussed in this book, mental representations have perceptual properties. Therefore, they are not simply conceptual, but related to our experience. This connection can be revealed, for example, through the use of tasks combining texts and pictures, as proposed by Madden and Zwaan (2003), for exploring the influence of tense categories on mental representations. Madden and Zwaan focused on verb aspect, comparing the perfective aspect, which conveys the fact that an action has been completed, and the imperfective aspect, which conveys the fact that an action is ongoing. For example, in English, the past tense (Eva wrote a book) is perfective, whereas the past progressive (Eva was writing a book) is imperfective.

Madden and Zwaan (2003) constructed 26 experimental sentences describing a character involved in an action that implied duration and obligatory end points. For example, painting your house or going to work implies a certain result (the house is painted, the person has arrived at work). Every chosen action was described in the past tense and in the past progressive. For each sentence, two pictures, one corresponding to the completed action and the other to the action in progress, were created. The sentences and pictures were combined to create four possible conditions: past tense and completed action, past tense and action in progress, past progressive and completed action, past progressive and action in progress. Four lists of items were constructed, each containing a sentence–picture combination, so that every participant saw all the conditions throughout the experiment, but saw every item only once. The task of the participants in this experiment was simply to read each sentence and then to decide, when the picture appeared on the screen, whether or not it represented the scene described in the sentence. Replies and response times were recorded.

A very interesting element of this experiment lies in the fact that the two pictures were compatible with the situation described in the sentence even though one of the two pictures was more suitable than the other, in terms of the verb aspect. All experimental items were intended to elicit a YES decision. The authors' hypothesis specifically concerned response times, namely that these times would be shorter when the picture was consistent with the aspectual information conveyed by the verb. In order to offer participants the possibility of responding NO, and thereby accomplishing the task, 26 filler sentences were constructed with the same structure as the experimental sentences, followed by a picture which did not correspond to the scene at all.

An analysis of the YES responses showed that, in general, responses were faster in the congruent condition, but that this effect was mainly driven by the perfective sentences. In other words, when the sentence conveyed the idea of a completed action, participants were faster to respond YES to the pictures containing the completed action compared to the ongoing action. The research hypothesis was therefore confirmed, but only partially. Madden and Zwaan (2003) suggested two explanations for these results. The first one argued that it was possible that when reading the description of an action in progress, the participants represented such an action at different stages, and that the picture representing the action in progress did not really correspond to any of these stages. The second explanation suggested that the imperfective aspect of the verb could have encouraged the participants to represent all the stages of the action, leading them to accept both pictures as representing the situation.

5.8. The self-paced reading paradigm

In this section, we will discuss the paradigms that are specifically related to the reading process. The first of these, the self-paced reading paradigm, invites participants to read sentences or texts, either word-by-word, or in segments, or sentence-by-sentence. The participants indicate that the word, the segment or the sentence has been read by pressing a key, which makes the text disappear and the next element appear. In this task, participants can read at their own pace. The instructions generally encourage them to do so in the most naturally possible way, for them to properly understand the text presented.

In order to ensure that the participants read the texts properly, it is necessary to include comprehension questions about the elements that have just been read. These questions are generally associated with filler items, aimed at concealing the experimental manipulation from the participants. It is important to construct suitable questions, neither too simple nor too complicated, since the characteristics of these questions may influence the strategies implemented by participants while reading (Jegerski 2014). Comprehension questions generally appear for only a number of the items.

The advantages of the self-paced reading paradigm are firstly the possibility of getting access to an online measurement of comprehension and secondly the collection of reading times related to segments of text, or even to every word in a text. In addition, this technique is non-invasive, and relatively simple to set up and use.

The disadvantages of this method are mainly related to the fact that the words or segments of text already processed disappear as the text unfolds This does not allow the normal text processing involved in natural reading, which we discuss in further detail in the next section. For the moment, suffice it to say that readers sometimes go back to the text in order to verify information, or to read certain passages over again. By preventing them from going backwards, the method involved in the self-paced reading paradigm requires larger memory capacities than those involved in natural reading. For this reason, the reading times collected in experiments are generally slower than those one would observe in natural reading.

The characteristics of reading should also be taken into account when creating experimental items used in a self-paced reading experiment. For example, it is known that the last segments of a text are read more slowly than the others, due to the finalization of the mental representation. At this point, the different pieces of information included in the representation are linked to one another in order to create a thorough representation. This is why critical elements should not be placed at the end of an item, so as not to confuse the potentially obtained effects with integration effects. Spill-over effects are also common in this type of paradigm. These effects correspond to the fact that the processing of a word or of a segment is not always finished when the person goes ahead with the text, and can continue while processing the next word or segment. For this reason, it can be useful to analyze not only the times related to the critical word or segment, but also those relating to the words or segments directly following the critical sections.

In order to illustrate the use of this method, we will present the study by Kelter *et al.* (2004, experiment 1), dealing with the influence of the recency of an event on its activation in the mental representation of the text. Various studies have indeed shown that it is not the event's recent mention in the text that plays a role in its representation, but its recency regarding the current situation described in the text. For example, in a verification task, Carreiras *et al.* (1997) showed that readers more quickly recognized a role name such as the *baker* or the *teacher* introduced in a story when the role was associated with the protagonist in the present rather than in the past. In their research, Kelter *et al.* examined the activation of past events depending on the time elapsed between the event and the present.

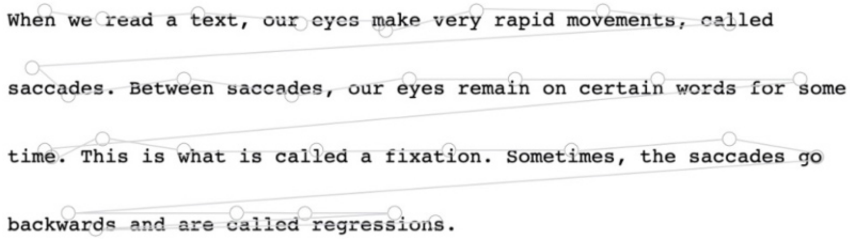
In order to do this, short stories introduced a situation before presenting a first event. Then, a second event was described, which was either long or short. Finally, a third event was presented before the story mentioned a sentence referring back to the first event. The story ended with one or two neutral sentences. For example, one of the stories described a couple getting ready to celebrate Christmas. The husband informed his wife that he disagreed with her choice of Christmas decorations. She got angry. Then, she went into the kitchen and, in the short-term condition, put some cookies on a plate, whereas in the long-term condition, she baked these same cookies. The story continued with a description of the smell of cookies and the Christmas spirit that emanated from them. The target sentence then appeared, referring to the anger experienced earlier by the woman: “At that time, she regretted her anger”.

Participants simply had to read the stories sentence-by-sentence, and the reading time for each sentence was recorded. In order to examine the potential spill-over effects of the experimental sentences on the following reading times, analyses were carried out not only on the target sentences, but also on the filler sentences that followed the experimental ones. While reading times for the filler sentences did not differ between conditions, reading times for the target sentence were slower when a long-term event was described in comparison with a short-term event, as expected. These results support the idea that the organization (and accessibility) of consecutive events in the mental representation of readers reflects the temporal references transmitted by the text.

5.9. Eye-tracking

The eye-tracking technique is similar to the previous paradigm, in that the participants have to read words, sentences or text excerpts. The difference lies in the type of measurement used: this time, this is the participants' gazes' direction while reading that is observed. To do this, the participants have to look at a screen on which the stimuli are displayed. Eye movements are recorded by a camera placed in front of them while they are processing the stimuli. A light source illuminates the eye, causing reflections in the pupil and on the cornea, which are detected by the camera. On the basis of the reflections, it is possible to infer the direction of the gaze with impressive spatial (0.5 degree) and temporal accuracy, a measurement being made every millisecond.

When we read, our eyes move forwards and backwards, and dwell for a longer or a shorter period of time on certain words. Contrary to popular belief, eyes do not scan the text in a regular manner while reading. They dwell on certain words and then quickly move on to other words. We refer to saccades when speaking about eye movements, and to fixations when the eyes remain motionless for a short time. During fixation, which lasts approximately 200–300 milliseconds, information can be retrieved and processed, and information processing goes on during the following saccade. The reason for this succession of fixations and saccades is that the acuity of our visual field is high in the central area of the eye, the *fovea*, but decreases very quickly in the parafoveal and peripheral regions. In order to process words, it is therefore necessary to bring them to the center of the fovea. Moreover, not all words are fixated during reading. Content words are fixated 85% of the time, whereas function words are fixated only 35% of the time, and short words are often skipped (Rayner 2009).



When we read a text, our eyes make very rapid movements, called saccades. Between saccades, our eyes remain on certain words for some time. This is what is called a fixation. Sometimes, the saccades go backwards and are called regressions.

Figure 5.2. *Illustrations of eye movements during reading. The circles correspond to fixations and the lines to saccades*

Eye-tracking can be done in different ways. The most natural way is to present the text and to record the eye movements made over it. It is also possible to adapt the text presentation to the person's eye movements, using the moving window technique (McConkie and Rayner 1975). In this case, only a portion of the text is presented, whose center corresponds to the fixation point and whose width (a certain number of characters) is manipulated in the experiment. The rest of the text is replaced either by the same character (an X, for example) or by other characters such as letters which may be visually similar to the original letters, or not. This technique made it possible to define the size of the perceptual span, which is three to four letters to the left and 14–15 letters to the right of the fixation point, for languages such as English and French. We can see that the perceptual span is rather narrow and that it is asymmetrical, that is, that it is larger towards the side where the eye naturally goes while reading. Another technique used in eye-tracking corresponds to the foveal mask (Rayner and Bertera 1979). In this case, the portion of the text around the fixation point is hidden, whereas the rest of the text remains visible.

On the basis of eye movements, it is possible to study the number and duration of fixations, the size of saccades and (backtracking) regressions, as well as their starting and finishing points, among other measures eye-tracking can offer. When using this method, the sentences or texts are often divided into areas of interest, grouping together the segments manipulated during the experiment that will be compared between the conditions. The variables frequently used for the aggregation of fixation points into reading measurements are numerous. For example, the first fixation duration corresponds to the duration of the first fixation on a word or area of interest. The first-pass reading time or first-run dwell time is the total time spent on a word or area of interest before the gaze goes to the right or to the left. The regression path duration corresponds to the time spent on a word or area of interest before leaving it to the right and includes the time spent re-reading previous portions of the text. The total reading time or dwell time is the sum of all fixations made on a word or area of interest, including regressions from other portions of the text. These variables give different clues as to the processes involved in reading. Some of them, such as the first fixation durations or the first-pass reading time, are considered as early processing measurements, whereas others, such as the total reading time, correspond to late measurements reflecting more elaborate processes (Staub and Rayner 2007).

Measuring eye movements is based on the idea that the time spent on a word corresponds to the time needed for processing this word (Just and Carpenter 1980). This idea is nonetheless questioned by the fact that eye movements can hardly account for the cognitive processes involved in the interpretation of the information just read. As a matter of fact, they partly depend on uncontrollable processes, such as the size or the speed of saccades that lead the eye to land more or less accurately on a given word. Skipped words are also processed, probably in a peripheral way, even if there is no fixation to objectively attest for this. We will not discuss these limitations in further detail, since they go beyond the scope of an introductory book. What is important to remember is that there is no perfect concordance between fixation time and processing time for a word or group of words.

Analyzing eye movements is useful for studying a wide variety of processes. As we saw above, it can contribute to a better comprehension of the natural reading process, as well as the basic characteristics of eye movements. The influence of many variables on the reading process, such as word frequency, contextual predictability, word length or polysemy, could also be investigated thanks to this method. The analysis of eye movements is also interesting for other levels of language processing, such as sentence or text comprehension, the development of mental representations or even the pragmatic processes involved in discourse comprehension.

Measuring eye movements also has the great advantage of being much closer to natural reading than self-paced reading. It also offers the possibility of obtaining fine measurements of the time spent on words or segments, as well as an indication of the processes underlying comprehension while reading, thanks to the observation of regressions, which are not accessible through other methods.

From a technical point of view, measuring eye movements is complex. It requires mastery of the necessary tools, as well as great accuracy in measurement. Accuracy can be granted thanks to prior instrument calibration, but the presence of the experimenter is required at all times to verify the quality of the measurements as the experiment progresses.

From a methodological point of view, the characteristics of natural eye movements while reading should be taken into account while creating the linguistic material. It is essential to verify that the target interest areas are comparable, in particular in terms of positioning on the screen, and that they do not appear at the start of the line, which is where the gaze position is generally adjusted.

The main disadvantage of this measurement is the cost of the equipment required, as well as the time cost, as it is only possible to test one person at a time. Once the data have been acquired, their processing also requires advanced technical and statistical knowledge. The large amount of measures recorded in an eye-tracking experiment may also become a disadvantage in some cases, since the different measurements sometimes offer different results which are not easy to interpret. During the analysis of numerous indicators, it is also likely that one or the other may look different from condition to condition. It is therefore highly advisable to define which indicators will be observed beforehand, as well as the specific hypotheses related to such different indicators. Without doing this, there is a risk of finding spurious results, stemming from the accumulation of statistical tests which increases the probability of finding a difference that does not reflect a real effect.

As an example of the application of the eye-tracking technique, we will present the study by Gordon *et al.* (2006, experiment 1) on how complex sentences are processed while reading. Their research focused on the comprehension of complex syntactic structures, requiring the simultaneous activation of two noun phrases (NP), before being able to associate them with the different expressions of the sentence. Gordon *et al.*'s specific hypothesis concerned the similarity between these NPs, the idea being that NPs of the same type could cause interference in their processing. A previous study measuring reaction times demonstrated that retrieving the object NP as in (2) leads to longer reading times than retrieving the subject

NP, as in (1). This effect was also larger when the subject and the object were semantically similar (barber-tailor, John-Bill) (Gordon *et al*. 2001):

(1) It was the barber/John who saw the lawyer/Bill in the parking lot.

(2) It was the barber/John who the lawyer/Bill saw in the parking lot.

In order to study the comprehension of such structures in a more natural way than in a self-paced reading task, and to determine at which point in the sentence readers find it difficult to deal with such structures, Gordon *et al.* (2006) used the eye-tracking methodology.

Their participants had to read sentences containing a relative clause (RC) associated with the subject of the sentence and whose relative pronoun was either in the subject position (3) or in the object position (4) of the clause. The RC also contained either a proper name or a role name. Each sentence could thus appear in one of the four versions, two of which contained the same type of NPs (role names) and two contained different types of NPs (role name and proper name). Each sentence appeared isolated on the screen, and the participants' eye movements were recorded until they indicated that the sentence had been read, by pressing a button. At that point, a comprehension question could appear (in 15% of cases), which the participants simply had to answer orally with YES or NO:

(3) The banker that praised the barber/Sophie climbed the mountain just outside of town.

(4) The banker that the barber/Sophie praised climbed the mountain just outside of town.

The eye movements associated with the area of the relative clause (from the relative pronoun until the main verb, without including it), as well as those associated with the verb of the main clause, were analyzed. As it had been previously demonstrated in the literature, the results confirmed that relative clauses with object NPs were read more slowly than those with subject NPs. Likewise, the reading time for the verb of the main clause was longer in the object condition than in the subject condition. The difference between these two conditions was itself larger when the NP of the relative clause was a role name (similar to that of the main clause), than when it was a proper name. These effects emerged in early processing measures (first fixation duration and first-pass reading time), suggesting that the type of NP influences its processing, as well as its integration into the main sentence, as soon as it appears. Later processing measures showed that the rereading of the target areas was also influenced by the variables examined.

In summary, eye-tracking, as well as self-paced reading, enables us to study many aspects of reading, from the simplest level of word processing to the more complex level of discourse comprehension. Given the large variety of measures that eye-tracking allows us to collect, this methodology could be interpreted as being more advantageous and interesting, *prima facie*. The choice to use eye-tracking rather than self-paced reading should, however, be made on the basis of the processes we want to observe and the possibility of investigating such processes that each methodology offers. In cases where the experimental design makes it possible to investigate a question using the self-paced reading paradigm, resorting to the eye-tracking method – with all the technical difficulties it entails – could end up being superfluous.

5.10. The visual world paradigm

To conclude, we present an experimental technique that makes it possible to study comprehension of spoken language, by means of the visual world paradigm. In this paradigm, participants listen to linguistic stimuli while looking at a scene, objects or words on a screen, while their eye movements are recorded. The participants' task may simply be to listen to a sentence or text while watching a scene, and then to attach any of the objects or words, according to the instructions received. This paradigm is based on the assumption that when we process speech, at the same time as observing a scene or pictures, we tend to relate what we hear to what we see. The eye movements of the participants involved in a visual world task reflect the attention devoted to the different objects or parts of the scene, according to the linguistic content heard. Somehow, this paradigm makes it possible to observe how people interpret the flow of discourse and to establish what they predict on the basis of what they hear.

In this paradigm, the most frequently used measurements are related to the specific regions of the screen participants look at during the task, specifically after listening to a target word. Common measurements are the fixation proportions on areas of interest or the number of saccades directed towards them. Of course, the time window in which fixations, saccades or regressions are analyzed must be defined depending on the research question and the process investigated.

The visual world paradigm makes it possible to study language comprehension at all levels. At the level of word comprehension, research has been carried out, for example, on the areas of phonological processing, word recognition by bilingual speakers, and the effects of context on word recognition. At the discourse level, this paradigm has made it possible to better understand the role of lexical and structural constraints in sentence comprehension. It has also been useful for examining questions related to pragmatics or to linguistic relativity (for a review, see Huettig *et al*. (2011)).

The advantage of this method is that it allows us to assess language comprehension without requiring reading skills, the use of written material or even metalinguistic abilities. Participants simply look at a screen while words or sentences are presented to them. For this reason, this method is very useful for studying children or people with written language impairments, such as illiterate people.

As an example, we present the study by Engelen *et al*. (2014) on the resolution of anaphora in children's narrative comprehension. This study specifically investigated the ability of children, aged 6–11 years, to determine and follow the character being referred to as they listened to a story. It also had the particularity of adopting a natural approach, presenting a story which lasted almost eight minutes, rather than many unrelated items, as is common in most experimental studies. Furthermore, children were split into groups on the basis of their comprehension of the story, which assessed their memorization of literal information and also inference-based information.

The story, told in Dutch, contained four characters (a hedgehog, a rabbit, a squirrel and a mouse), presented simultaneously on a screen in front of the children. The characters had human-like attributes, such as being able to talk. They were all masculine, so that the grammatical gender of a pronoun could not be used as a cue, and the anaphoric pronoun could refer to any of them. In the story, the characters were introduced and then started to interact with one another. A portion of the story is reproduced below (the words in italics are those for which eye movements were analyzed):

> "Meanwhile at the lake, the rabbit and the squirrel were sitting on a log. The *squirrel* wanted to do a running contest with the *rabbit*. 'I'm sure that from here I can run to the giant rock and back faster than you', he said. 'Well, let's see', the rabbit said.

> 'Okay', the squirrel said. 'I'll count to three and at three we run'. They both got on their marks. The *squirrel* started counting: 'One… two… three!' The *rabbit* dashed away with great speed. But what did the *squirrel* do? He just stayed there. The *rabbit* didn't notice anything and rushed on. The *squirrel* lay down on the log in the sun. *He* thought it was a good joke and knew what *he*'d say when his friend would come back." (Retrieved from Engelen *et al.* (2014))

All of the referential expressions in the text could not be analyzed due to the varying levels of difficulty they presented. For this reason, Engelen *et al.* (2014) chose 42 expressions (28 names and 14 anaphoric pronouns), for which eye movements were analyzed and compared between the groups of children. The main hypothesis was that comprehension of anaphoric pronouns and the ability to follow the protagonist of a story depend not only on literal, but mostly on inferential skills. It was expected that children with good skills in these two areas would look towards the picture representing the character in question when this was referred to by a name or an anaphoric pronoun. For children in the middle group, the glances towards the corresponding picture should be more numerous when the character was designated by a name, rather than by an anaphoric pronoun. Finally, for children with poor comprehension skills, it was expected that they would generally look less at the pictures related to the characters than the other two groups.

When analyzing the results, only two groups of participants could be constructed on the basis of their answers to the comprehension test: a group with good literal and inferential skills and a group with poor skills. As no child revealed skills for being placed in the intermediate group, eye movements were eventually compared between these two groups.

Eye movement analyses were performed on two-second time windows from the appearance of the name or the anaphora. Compared to children with poor comprehension skills, the group with good skills looked at the picture associated with the reference character more, either after hearing names or anaphoras. Interestingly, this difference did not result from the adjustment of the gaze following the hearing of the name or the anaphora, but from the probability of fixating the target picture at the time of hearing it. In other words, the group with good comprehension skills was more inclined to make fixations on the target picture, in advance, and even more when this picture was referenced by an anaphora. According to the authors, and based on other observations (Barr *et al.* 2011), this reflects the expectations of good comprehension, in terms of the unfolding of a text. This study therefore suggests that children with good comprehension skills are those who can anticipate the content of the text.

5.11. Conclusion

In this chapter, we reviewed the different methods for studying online comprehension. We first presented think-aloud protocols which allow us to access people's thoughts and reflections during the comprehension of a text. This method is particularly useful for identifying the stages of comprehension, but it has many limitations, in particular due to unnatural protocols. We also discussed the interest of measuring the response time as an indicator of comprehension, especially when studying processes or representations that are inaccessible to consciousness, or when we wish not to draw the attention of participants to the object of study. Response time can be collected while performing various tasks, specifically assessing one or many of the processes involved in language comprehension, and thereby leading to different conclusions being drawn. Most of the tasks used in experimental linguistics have in common the activation of certain characteristics which are present in linguistic stimuli, and which can be shown thanks to priming and interference effects.

5.12. Revision questions and answer key

5.12.1. *Questions*

1) What are the main characteristics of the online tasks presented in this chapter compared to the offline tasks presented in Chapter 4?

2) What is the difference between simple reaction time and choice reaction time? Find an example of a task in which a word is presented and simple reaction time can be recorded, and another example where choice reaction time can be recorded.

3) Which online task do you consider most suitable for studying:

a) the influence of connectives on text comprehension?

b) the influence of the number of orthographic neighbors on word recognition?

4) Which technique, between self-paced reading and eye-tracking, would you choose for studying:

a) the influence of instructions (reading for pleasure vs. learning) on reading a 10-sentence passage?

b) the influence of font on reading speed?

5) How would you use the visual world paradigm to study the lexicon of bilingual people?

6) How would you use the verification paradigm to test whether people include a representation of color in their mental representation when understanding a sentence such as: *Chloe chose the ripest tomato and crunched it*?

5.12.2. *Answer key*

1) Online methods aim to study the processes involved in comprehension, whereas offline methods make it possible to assess the results of comprehension. Besides, online methods are implicit, that is, they attempt to indirectly access indicators reflecting the processes involved in comprehension. Since these processes are for the most part inaccessible to consciousness, online methods are generally based on the processing time of a linguistic stimulus, in order to be able to study its complexity or the processing it requires.

2) The time between the presentation of a stimulus and the response given by a person corresponds to the reaction time. This is referred to as simple when the response is triggered automatically by the appearance of the stimulus, for example, when participants simply press a button when something appears on the screen. On the other hand, reaction time can correspond to choice reaction time, when the processing of the stimulus presented is required to be able to provide an answer. For instance, this can be the case when it is necessary to answer only whether the stimulus presented is of a certain color or if the answer requires a choice (YES/NO, for example). In this case, it is not only the appearance of the stimulus that triggers the response. Going back to our question, simple reaction time following the appearance of a word can be measured when the task requires the pressing of the key as soon as a stimulus appears on the screen. Choice reaction time can be measured when the task requires the pressing of the key as soon as the stimulus corresponding to a word appears (as opposed to a

non-word), when it corresponds to a word in French (vs. another language), or when it has certain properties (e.g. when it matches a certain grammatical category). Reaction times measured in recognition or in reading tasks are yet other examples of choice reaction times.

3) a) To study this question, it is necessary to turn to a method that makes it possible to evaluate text processing in real time, such as the self-paced reading paradigm, or eye-tracking.

b) Here, a lexical decision task would be appropriate, because this task makes it possible to evaluate the time necessary to categorize a string of letters as a word, which gives an evaluation of the time necessary for the recognition of a word.

4) a) The instructions received before reading a text influence the depth of processing dedicated to the content of the text. In the case of reading for pleasure, we can assume that the participants read the text naturally, with the sole aim of understanding it, but without putting any particular effort into retaining the content. In the case of reading in order to memorize the content of the text, we may assume that participants set up strategies which differ from the ones used in natural reading. In order to verify this, the most suitable method would be to measure eye movement, since it contributes to measuring not only the reading time for different words or segments, but also to observing the regressions performed while reading, that is, the eye movements aimed at re-reading certain sections in the text. This last possibility is particularly useful for this research question.

b) This research question aims to study how different fonts can influence reading speed. The self-paced reading paradigm is suitable for studying such a question, since reading speed is the dependent variable measured. It is therefore unnecessary to resort to complex measures such as eye-tracking.

5) The organization of the lexicon in bilingual speakers can be approached in different ways. First, we could hypothesize that the lexicon of every language is independently organized from other languages, as if the words in each language were stored in closed sets. We could also hypothesize that the two lexicons are stored jointly; that is, that the words in the two languages are stored in the same place. Finally, we could imagine an intermediate version, in which the lexicons are separated at a certain level (e.g. at the phonological level), but connected at another (e.g. the semantic level). In order to verify a possible interconnection using the visual world paradigm, it would first be necessary to determine the linguistic variable we

wish to examine (e.g. phonology). Once this variable has been chosen, we could choose words in L1 and L2 that may or may not be similar regarding this variable. For example, if we are interested in the phonology of words, we could choose words in L1 and L2 sharing their first phoneme, or not. We could then present L1 words orally to participants, while simultaneously presenting them with pictures corresponding to the selected words (in L1 and L2) and filler words. By observing their eye movements, it would be possible to know whether listening to a word in L1 activates the representation of a phonologically similar word in L2. For example, such a study was carried out by Spivey and Marian (1999).

6) In order to test this question using the verification paradigm, it would be appropriate to present sentences containing different colored objects, such as a tomato, which is generally red, but which can also be green when it is not ripe. Other examples of different colored objects are bananas, traffic lights or the sky. Following the reading of each sentence, a picture of the object described in the sentence could be presented. This could either be the same color as the one described above (red) or another possible color (green). The task of the participants would be to indicate whether the object presented on the picture corresponds to an object contained in the sentence. If the participants include a color in their mental representation, objects presented in the same color as the one described in the sentence should be recognized more quickly than those presented in another color. Research similar to this proposal was carried out by Connell (2007) and then replicated by Hoeben Mannaert *et al*. (2017).

5.13. Further reading

For more developments on reading time based methods and visual attention based methods, the reader may refer to Kaiser (2013) and Jegerski (2014). Rayner (1998) and Clifton *et al*. (2007) are references in relation to eye-tracking, its use, as well as the results obtained in the fields of word recognition and sentence comprehension. Staub and Rayner (2007) offer a more accessible text for beginners. For more developments on the visual world paradigm, see Huettig *et al*. (2011). Finally, we recommend reading the book by Gonzalez-Marquez *et al*. (2007b) which offers numerous examples of the application of the techniques discussed in this chapter, in the field of cognitive linguistics. For a contribution comparing the interests of online and offline measurements in the study of comprehension, see Ferreira and Yang (2019).

6

Practical Aspects for Designing an Experiment

In this chapter, we take you step by step through the different practical aspects of designing an experiment, as well as the resources needed for every stage. We first see how to look for scientific sources and access the bibliographic resources required for developing the research question. We then return to the conceptualization and formulation of the research question and the operational hypotheses. The different stages involved in building the experiment itself will then be described one after the other. We primarily address the choice of experimental design and the constraints linked to the different types of design, before discussing the key aspects of experiments in linguistics: the linguistic items used in the experiment. We then describe the different stages which mark out the typical course of an experiment, and discuss the ethical principles that have to be respected while conducting experiments on human participants.

6.1. Searching scientific literature and getting access to bibliographic resources

The first crucial step in the implementation of experimental research is the definition of the research question. It is from this definition that all the subsequent stages will ensue: the choice of method, the observed indicators, the experimental design, the linguistic material employed and finally statistical analyses. This is why it is necessary to devote time and reflection

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

to it, in order to arrive at a well formulated and clearly delimited research hypothesis, which will in turn lead to a well-designed experiment. Many problems may arise from an incomplete or ungrounded research hypothesis, and these can be avoided by careful work prior to the implementation of the experiment itself.

The first steps in research are often guided by a general problem which is somehow related to the researcher's personal interest, for example, the perception of different accents, language acquisition in bilingual children or the connection between language and thought. Sometimes, it is also possible to formulate a specific question intuitively, based on prior knowledge or daily observations. In these different cases, before embarking on experimental research, it is necessary to perform a thorough review of existing literature, going through the studies and the accumulated knowledge in relation to the topic of interest. For researchers who are at their beginning, the literature review phase should make it possible to get a general idea of a specific research domain, in order to delimit the research topic to a specific hypothesis, which can be investigated experimentally.

In order to carry out a review of the scientific literature, different types of sources can be consulted. These can be monographs, that is, scientific works produced by a single researcher, textbooks, such as this one, or collaborative works, in which every chapter has been written by an expert or a group of experts on the subject. Put together, these sources provide an overview of the research problem. Scientific articles are the sources that generally describe specific experimental research in the most detail, since book chapters and monographs tend to focus on offering a general overview of a field. This is why scientific articles represent eminently useful sources for preparing an experimental study.

Nevertheless, it is advisable to start with the general works and articles, which present the different aspects of a topic and summarize the knowledge acquired so far, in order to clearly define the particular aspect that will be investigated. Thereafter, once the aspect that will be examined is set, the literature review may enter a new phase, in which we go through scientific articles more specifically related to the chosen aspect.

Relevant scientific literature can be identified through specific search engines, such as *Google Scholar*[1], or bibliographic databases such as *Web of Science*[2], *PsycINFO*[3], *JSTOR*[4], *Linguistics and Language Behavior Abstracts/LLBA*[5], or *Bibliography of Linguistic Literature Database/BLLDB*[6].

The first queries, performed using general keywords, often result in a considerable number of scientific articles likely to be relevant. Let us take the example of the tip-of-the-tongue (TOT) phenomenon. When typing *tip-of-the-tongue* into a search engine like *Google Scholar*, more than 560,000 entries are retrieved. It is therefore essential to quickly determine other keywords, making it possible to narrow the query to a more specific problem. For example, we could add the keyword *bilingualism* in order to limit our query to TOT phenomena among bilingual speakers. Despite this addition, the results are still numerous: around 20,000 in this case. It is possible to further restrict the results by carrying out an advanced search, in which various fields can be selected, such as the general subject, title, language, date of publication or even the type of publication, which can also be combined with the Boolean operators AND, OR or NOT (if we wish to exclude some keywords).

It is very useful to quickly get acquainted with the different search engines, in order to use their properties effectively. Most of these allow you to look for a specific expression, by phrasing it between quotes. In this case, rather than getting access to all the sources including the different keywords of the expression, we only obtain those sources where the expression itself appears. For example, for the expression “Tip-of-the-tongue”, the query includes all the sources including such an expression, but leaves out the sources with words *tip* or *tongue* when these appear isolated. A query carried out on the titles containing the keywords “tip-of-the-tongue” AND “bilingual” in Web of Science only yields five entries and thus makes it possible to target only the most relevant sources for the subject.

[1] http://scholar.google.com.

[2] http://apps.webofknowledge.com.

[3] http://psycnet.apa.org/search/basic.

[4] htttp://www.jstor.org.

[5] https://search.proquest.com/llba.

[6] http://www.blldb-online.de.

The access to bibliographic databases and to scientific articles is often restricted to people or institutions with a paid subscription. In order to access them, it is necessary to identify ourselves as a member of a university benefiting from the required subscription. When we do not have this type of affiliation, different solutions exist to still have access to a source. The preliminary query by searching for keywords can be performed on *Google Scholar,* which shows the links to the documents associated with the source and can be accessed without a subscription where these exist. An alternative solution is to turn to *Unpaywall*[7], a project linking the original publications to their open access versions where these exist. By October 2019, *Unpaywall* had a database of more than 24 million scientific articles for free access, which could either be browsed through the database search or by adding an extension to *Firefox* or *Chrome* browsers. If, despite this, the sources are still not accessible, another possibility is to consult the personal web pages of the authors, or their *ResearchGate*[8] or *Academia*[9] profiles, on which the articles are sometimes made available. Finally, it is also possible to contact the authors directly to request a copy of their publication. *ResearchGate* also allows you to request private versions of the articles via the website's interface.

6.2. Conceptualizing and formulating the research hypothesis

A helpful literature review should not only lead to an overall vision of the problem studied, but also offer a good understanding of the methods used when investigating it. Particular attention should be devoted to the dependent and independent variables observed, as well as to the manner in which these variables have been operationalized.

It is quite interesting to observe that when screening existing literature in a certain field, it can either simplify the problem considered or, on the contrary, make it more complex. Depending on the case, the information acquired during the documentation phase can easily be brought together in order to build a precise research problem, or lead to completely reformulating the initial hypothesis, instead. Generally, different sources offer various insights into the same problem, and it quickly becomes

[7] https://unpaywall.org.

[8] https://www.researchgate.net.

[9] https://www.academia.edu.

necessary to try to narrow the complexity inherent in a research field, to a specific aspect which seems to be preeminent and that can be studied in an experiment. It is indeed impossible to exhaustively study a whole subject, even a specific one, in a single study. However, the accumulation of studies, each focusing on a specific facet of the problem, makes it possible to construct a comprehensive view of the subject.

Once a specific problem has been identified on the basis of the literature, different scenarios may arise. First of all, it is possible that the literature review gives rise to a new idea, which has not yet been studied by empirical research. It is also possible that different studies which resorted to multiple methods ended up revealing conflicting results. In this case, the research question may aim to understand the cause of these conflicting results, for example, by suggesting studying it by means of a new method. It could also be that an explanatory variable has not yet been examined, and gives rise to a research question based on this new variable. Finally, the validity of existing research can be called into question by new knowledge. This could lead to an attempt to replicate known results in order to verify their quality.

Whatever the initial situation and the reasons for research, the next step will consist of formulating the research question, as well as the hypothesis based on the literature review. We have already described the notion of research hypothesis in Chapter 1. We stressed the fact that it has to be empirically testable. In other words, every research hypothesis has to propose a directional relationship between an independent variable and a dependent variable, at least. It also has to operationalize these variables, by specifying the indicators used for measuring them.

The operationalization of the variables aims to ensure the validity and the reliability of the experiment, two concepts presented and discussed in detail in Chapter 2. At the operationalization stage, we have to maximize the chances of the chosen dependent variable to help us measure the process we want to observe and to reveal the connections between independent and dependent variables. At this stage, one way to do so is to rely on measurements used in previous studies. However, there are cases where the new study seeks to call into question the results found in the literature. In this type of situation, the use of a different type of measurement is evidently necessary. The choice of the new type of measurement should nonetheless be based on accumulated knowledge from existing literature. The first possibility would be to turn to a type of measurement whose effectiveness has been proven for testing similar phenomena to the one that will be examined. For instance, if the study aims to criticize the use of acceptability judgments, for evaluating the way in which different speech acts (e.g. requests, promises) are acquired, we could replace this type of measurement by a more implicit one. In order to assess the acquisition of pragmatic skills in children, we could resort to action-based tasks, which do not pose the same constraints as acceptability judgments, but whose effectiveness has already been demonstrated (Pouscoulous *et al.* 2007). A second possibility would be to choose a type of measurement that has not yet been applied to the phenomenon studied, but which seems appropriate in virtue of the processes it aims to shed light on.

Let us now turn more specifically to the definition of independent variables. As a reminder, independent variables correspond to the causes that will be manipulated in an experiment, in order to observe their effects on the dependent variable(s). For these variables, it is not only the type of measurement, but also the different conditions – or modalities – that need to be defined. As we have seen in Chapter 2, these conditions have to differ by the presence and the absence of the independent variable. They also have to make it possible to maximize the probability of observing the expected effect. For example, to test the hypothesis that frequent words are processed more quickly than less frequent words in a lexical decision task (and imagining that this effect has not already been widely demonstrated in the literature!), it might be necessary to build groups of words differing widely in terms of frequency. Then, if the effect is confirmed using these groups of words, it could be a good idea to refine the study by reducing the frequency difference between the groups of words, in order to offer a more accurate vision of the frequency effect.

Once the operational variables and the modalities of the independent variables have been defined, the hypothesis can be formulated clearly and precisely. At this point, it is appropriate to think about the different external variables, not examined directly in the experiment, but which could influence the independent and the dependent variables (see section 2.8). These external variables may be related to participants or to items. In order to illustrate the research hypothesis and to identify the external variables that need to be controlled, it may be useful to draw a diagram of the dependent and independent variables, on how to operationalize them and of the potential external variables (see Figure 6.1).

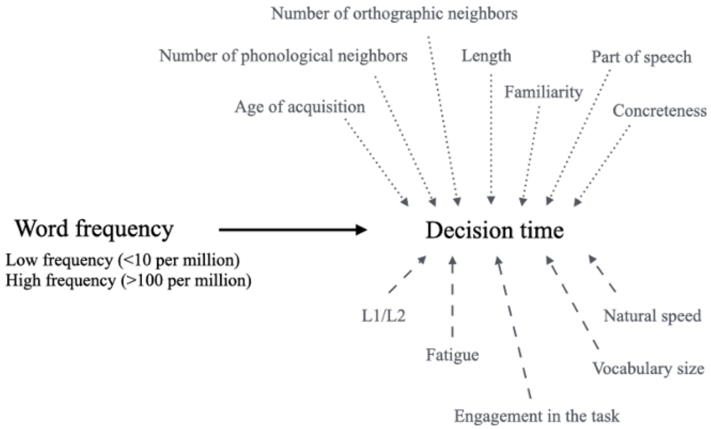


Figure 6.1. *Diagram of the operational hypothesis and examples of external variables which could influence the dependent variable. The variables above the dependent variable are related to the experiment items, and the ones below to the participants*[10]

On the basis of a list of external variables, it is then possible to determine those that will be considered as confounding and those that could vary randomly. As a reminder, a confounding variable is a variable whose modalities vary systematically with those of the independent variable. These variables must necessarily be controlled in order to ensure the internal validity of the experiment. If we go back to the example above, the word frequency is most likely related to word length (long words are less frequent than short words), and word length is likely to influence the lexical decision time. This is why word length is probably a confounding variable. In order to control it, we have two possibilities. We can either present only same-length words throughout the experiment or balance the conditions in terms of word length. In this case, it would be necessary to include, in each frequency condition, the same number of different-length words (e.g. 10 bisyllabic words, 10 three-syllable words and 10 four-syllable words). Likewise, other lexical characteristics could be related to frequency and may have an impact on the response time. Among other things, we could consider the word part of speech, or its number of phonological and orthographic neighbors. Other external variables, this time concerning the participants, could also come into play, such as their vocabulary size or their decision-making speed.

[10] The idea for this diagram comes from Pascal Gygax's Research Methodology course at the University of Fribourg (Switzerland).

In the same way as the variables investigated in the experiment, confounding variables are sometimes identified on the basis of intuition, but most of the time by consulting the literature. Such information can be found not only in the general results of the studies, but also in the *Method* section of scientific articles, where an exhaustive description of the material and control checks carried out on the variables is provided. Besides, it can often be useful to discuss research hypotheses, their operationalization and the variables to control with other people, either experts or less experts in the field, in order to detect potential problems, before implementing the experiment.

6.3. Choosing the experimental design

Once the research hypothesis has been formulated, the experimental design can be defined. Experimental designs can be categorized following different criteria: the number of independent or dependent variables and the assignment of participants to the different conditions of the independent variable. The number of dependent variables mainly stems from the method. Offline methods generally result in only one dependent variable, such as the score on an acceptability scale, or the number of correct answers in a recall task. On the other hand, in addition to the response itself, online methods measure the response time, thus introducing two dependent variables. These dependent variables can then be treated independently from one another, or one of the two variables can be considered as independent in the analyses. For example, it would be possible to create two conditions, one for YES replies and one for NO replies, and to analyze the differences in reaction times between these conditions. We will not discuss this possibility in further detail, since the points we will develop for univariate designs (having only one dependent variable) can be easily transferred to multivariate designs, as long as each dependent variable is analyzed separately.

6.3.1. *One independent variable*

When a single independent variable is examined, there are two main types of experimental design, depending on whether the participants take part in one or in all of the conditions. In a between-subject design (or independent measures), every person takes part in only one condition, whereas in a within-subject design (repeated-measures design), every participant takes part in all the conditions.

The choice of the experimental design depends first on the type of independent variable examined. When the independent variable cannot be manipulated by researchers and corresponds to an inherent characteristic of the individual, such as mother tongue or intellectual abilities, we can only use a between-subject design. A person's characteristic can correspond to only one of the modalities of the variable. When the independent variable can be manipulated by the researcher, it is possible to implement either a between-subject or a within-subject design.

Different factors come into play when choosing a between-subject or a within-subject design. The first important factor is the control of external variables. In a between-subject design, where the conditions do not contain the same individuals, it is necessary to ensure that the characteristics of the individuals that may influence the dependent variable, are distributed evenly between the conditions. Imagine an experiment comparing sentence comprehension in a spoken or a written modality. An external variable could be the intellectual level of the participants, measured through their IQ score. If the participants in the spoken condition have a higher IQ than those in the written condition, then the results of the experiment could be influenced by the IQ level in addition to the modality of presentation.

In order to control external variables, one solution could be to assign participants to the conditions randomly. For an assignment to qualify as random, every person should have the same chance of being assigned to a condition and that assignment should not depend on the characteristics of other participants. In this case, the assignment criterion should be the chance factor, so that every time a person showed up, a coin is tossed to decide which condition they should be included in. In our example, a person with a high IQ would have the same chance as a person with a lower IQ of belonging to each condition. In real life, such a random assignment is difficult to implement, since it could lead to a highly unequal number of participants between conditions. It is therefore preferable to use block randomization, in which all the conditions appear once within a block, before moving on to the following block. Within each block, the order of the conditions is random. For example, for a design with three conditions, the first three participants could be assigned to conditions 1, then 3, then 2, the next three to conditions 2, 3, then 1, and so on. We cannot rule out the fact that even by using such randomization, the participants of the different groups may systematically differ on some points. However, this method is effective, particularly for large samples.

Another way of controlling external variables, in the case of a between-subject design, would be to assign similar individuals to each condition. In this case, the relevant characteristics have to be defined beforehand, so that the groups of participants are equivalent. Going back to our example, it would be desirable for the two groups of participants not to differ in terms of IQ. Of course, it would be impossible to only recruit individuals with the same IQ score. What would be more feasible, is to make sure that the IQ scores of the two groups belong to a similar range and *on average*, individuals in the first group have a similar IQ to those in the second group. This type of assignment is only doable when the criteria that need to be taken into account are few and easily measurable. When the number of criteria increases, it becomes very difficult, not to say impossible, to recruit similar participants. For example, we can easily imagine the difficulty of finding homogeneous groups of 20-year-old bilingual French–Portuguese participants with a similar IQ. When multiple constraints must be met, it is necessary to turn to a within-subject design.

In a within-subject design, participants work as their own control, since their characteristics follow them from one condition to another. In our example, IQ would no longer be a problem, since a person with a high or a low IQ would be tested in both conditions of modality presentation. As a consequence, the effects of modality presentation could no longer be attributed to the characteristics of the participants. In this respect, within-subject designs resolve the difficulties associated with building even groups of participants. On the other hand, they show a risk of spill-over effects, that is, participating in one condition may influence the responses given in a condition presented afterwards.

There are different kinds of spill-over effects. One of them, the order effect correponds to the fact that the order of participation in the different conditions could become an involuntary factor within the experiment. In short, depending on the order in which the conditions are presented, the results may differ. For example, this effect was shown in the case of scalar implicatures. Children derive more scalar implicatures linked to words such as *some*, namely *some but not all*, when the conditions with the words *some* and *all* are alternated, producing a contrast effect which encourages the derivation of the implicature, when compared to an experiment in which all the items with *some* are presented in a block, followed by another block containing all the items with *all* (Skordos and Papafragou 2012)[11]. Another spill-over effect is that of learning, in the specific case where participation in one condition improves performance in the other condition. In the example we are analyzing, this would mean that simply carrying out the comprehension task in one modality condition, could improve the performance in the second condition. This would, of course, be the case if the sentences were the same in the two conditions. It is for this very reason that different sentences should be presented. We will return to this in the section on experimental material. This could also be the case if the participants developed specific strategies in the first condition, which could then be reused in the second condition. The effect of fatigue is yet another example of a spill-over effect producing opposite results to those from the learning effect. As conditions go by, performance levels decrease, due to the fatigue or weariness participants start to experience.

In order to overcome these different spill-over effects, two counterbalancing solutions can be implemented. The first way of counterbalancing would imply randomly modifying the order of presentation of the different items in the experiment, all conditions combined. The second way of counterbalancing would be to modify the order of presentation of the items within the conditions, and to modify the order of presentation of the conditions themselves. For an independent variable with two modalities, half of the participants would first take part in condition 1, then in condition 2, whereas the other half would take part in condition 2, then in condition 1. For example, half of the participants would complete the task in the spoken condition before moving on to the written condition. As the number of modalities for the independent variable increases, it very quickly becomes difficult to counterbalance the conditions. With three modalities, there would be six possible orders, and 24 possible orders for four modalities. It is therefore necessary to resort to partial counterbalancing, also known as *Latin square design*. In this type of design, rather than presenting all the possible combinations, we choose a portion of these so that each condition appears in a possible position, as illustrated below.

[11] In this study, the order of presentation of the conditions was controlled in order to assess the spill-over effect. In this sense, it was an independent variable in the experiment. Had this not been the case, and had the items only been presented in one of the order conditions, the conclusions might not have reflected reality.

Condition 1	Condition 2	Condition 3	Condition 4
A	B	C	D
D	A	B	C
C	D	A	B
B	C	D	A

Table 6.1. *Combination possibilities for 4 conditions in a Latin square design*

We should nonetheless note that there are situations where counterbalancing conditions is not convenient. For example, this would be the case for an independent variable involving a process difficult to cancel once activated. For example, Gillioz *et al.* (2012) studied the influence of different factors on the construction of emotional inferences. One of these factors corresponded to the simulation process, by which the readers imagined being characters of a story, in order to understand it from the inside. In this experiment, the simulation was manipulated through a within-subject design, by giving no specific instructions to the participants in the first part of the experiment, before specifically asking them to follow a simulation strategy in the second part. Here, a counterbalancing of the simulation conditions was not feasible, since adopting a simulation strategy during reading is difficult to cancel at a simple request. Therefore, a decision was made to present the conditions in the same order, at the risk of unintentionally inducing order effects into the results. Another solution would have been to manipulate this variable between the participants, while being careful to build even groups, as previously explained.

In summary, it is possible to build between-subject or within-subject designs. In order to control the external variables, within-subject designs should be used instead of between-subject designs whenever this is possible.

6.3.2. *Several independent variables: factorial designs*

An experiment can also be used to simultaneously test the role of several independent variables, by using a factorial design. This type of design makes it possible to test the hypotheses related to each independent variable, as well as to observe the joint influence of the independent variables, for instance, to determine whether the influence of a variable depends on the modality of another variable.

The simplest factorial design contains two variables, with two modalities each. It can be presented in a simplified way as a 2x2 design. In this type of design, the modalities of the two variables are combined to produce four conditions. Imagine that you want to extend the study about the modality of presentation (spoken or written) on sentence comprehension, by adding another independent variable, such as sentence complexity. The experiment should present spoken and written sentences in order to study the first variable. In addition, complex sentences and simple sentences should be used for studying the second variable. Combining these variables would result in the four conditions described below:

	Simple sentences	**Complex sentences**
Spoken modality	Simple sentences presented orally	Complex sentences presented orally
Written modality	Simple sentences presented in writing	Complex sentences presented in writing

Table 6.2. *Combination of independant variable modalities for creating conditions*

In a factorial design, every independent variable can correspond to a between-subject or within-subject measurement. To continue with our example, one possibility would be to follow a 2x2 design with independent measurements, in which participants only take part in one condition. In this case, the two independent variables would be between-subjects. Alternatively, it would be possible to follow a 2x2 repeated-measures design, in which the participants take part in all the conditions. Here, the independent variables would both be within-subjects. Finally, it would be possible to set up a mixed 2x2 design, in which one of the variables is between-subject and the other within-subject. In this case, participants would see two of the four conditions presented above, either a single type of sentence presented in spoken and written modalities, or the two types of sentences presented in a single modality.

When using factorial designs, different effects can be observed: the main effects and the interaction effects. The main effects correspond to the general effect of an independent variable on the dependent variable. In the example above, since there are two variables, two main effects can be observed. The first main effect would correspond to the presentation modality (spoken vs. written) effect on sentence comprehension, regardless of the type of sentences involved. For example, it may be that in general, participants better understand written sentences. The second main effect would correspond to the type of sentence effect on comprehension, regardless of the presentation modality. It is probable that simple sentences are generally better understood than complex ones. Finally, the interaction effect corresponds to the effect of one variable depending on the modality of the other variable. In the case of our example, a possible interaction effect would be that simple sentences are equally well understood in the spoken and written modalities, whereas complex sentences are better understood in writing than when spoken. Thus, the presentation modality effect would depend on the complexity of the sentence, since it would only be observed in the case of complex sentences.

In a factorial design, the number of modalities for every variable may vary, as well as the number of independent variables examined. For example, in a 2x3 design, two variables would be manipulated, one with two modalities and the second with three modalities. A 2x2x2 design would include three variables with two modalities each. Manipulating more than three variables in the same experiment is, however, not recommended, since the effects of interactions from such designs can become very complex to interpret.

6.4. Building the experimental material

The experimental material is a key element in every experiment. In the field of linguistics, given the variety of phenomena and processes investigated, the material sometimes corresponds to words, sentences or short texts whose length may vary. In addition, when the research hypothesis concerns individual differences such as the language level, for instance, it is necessary to measure such individual characteristics using questionnaires or specific tasks. Numerous resources concerning these questionnaires or these tasks, their validations, as well as their use for different purposes, are provided in the scientific literature. For this reason, we will not develop these aspects here, but will focus instead on the essential elements to be taken into account when developing the linguistic material used in the experiment. In this section, we will first present the general characteristics of the experimental material, as well as some useful resources for creating experimental material in linguistics. We will then approach the notions of experimental items and filler items. We will finally discuss the notion of lists, which have central importance in many linguistic experiments.

6.4.1. *Experimental items*

The nature of the experimental material developed for an experiment depends on the independent variables formulated in the research hypothesis, as well as on the task chosen. In most of the examples discussed in the previous chapters, we have seen that the participants had to judge, recall, read or even react to linguistic stimuli belonging to one and/or the other conditions of the experiment. These linguistic stimuli, called *items*, are selected so as to manipulate the independent variable and to control the external variables.

Before going further, it is important to understand the concept of an item. An item is an element for which a response is recorded in an experiment. For example, in a lexical decision task, an item corresponds to a series of letters that the participants have to categorize either as a word or a non-word. In an acceptability judgment task, an item could be a sentence whose acceptability participants have to judge on a scale; in a comprehension, reading or recall task, an item may correspond to a sentence or a short passage. Experiments contain many items, in order to collect a reliable measurement of the process investigated. Testing multiple items, as well as testing multiple participants, decreases the portion of error that is attributed to the specific characteristics of the items or the participants.

Experimental items are constructed so as to vary the properties investigated, while keeping the other properties as stable as possible. For example, an experiment on scalar implicatures could contrast two conditions: on one hand, the sentence “some kids like chocolate”, which gives rise to the implicature, and, on the other hand, the sentence “all kids like chocolate”, represents another condition. In the same way that the use of a repeated-measures design decreases the biases related to participants, repeating an item through all the conditions may decrease the biases this could entail. For this reason, items are developed so that they can appear in the different conditions whenever possible. Later, we will discuss how to distribute the items across the different conditions, but it is already useful to take this point into account when building the experimental material.

To illustrate these principles, let us go back to the example on the influence of presentation modality and complexity on sentence comprehension. Sentence complexity could be operationalized by the presence of a relative clause with a subject pronoun (less complex), or an object pronoun (more complex), for instance. In order to manipulate the presentation modality, half of the sentences would have to be presented orally and the other half in writing. The dependent variables of this experiment could be the number of correct answers given to verification questions, following the presentation of the sentence, as well as the response time needed to provide such answers.

Following these criteria, the experimental items of this experiment could take the form:

(1a) The woman who follows the man carries an umbrella.

(1b) The woman whom the man follows carries an umbrella.

(2a) Elegantly, the courtier who adores his sweetheart picks her up to kiss her.

(2b) Elegantly, the courtier whom his sweetheart adores, picks her up to kiss her.

The two sentences in each pair clearly differ on the role of the relative pronoun. However, the two pairs of sentences also differ in structure, which can be problematic. The second pair of sentences is more syntactically complex than the first one. It also contains less frequent vocabulary, which could entail comprehension difficulties. In this case, the items would vary in relation to other complexity aspects than those investigated (the role of the relative pronoun), which increases the possibility of external variable involvement. In order to avoid the intrusion of unwanted external variables, it is preferable to build a homogeneous set of items, by establishing criteria relating to their structure, their style, language register, etc. applicable to all items.

In experiments where the material corresponds to sentences or texts, compliance with these criteria can be checked by means of a pretest, where people are asked to give their opinion about some of the material's features. These people can be colleagues or people with the same general characteristics as the participants in the experiment. Performing a pretest may sometimes seem superfluous depending on the criteria applied to the items, but it is important to keep in mind that the judgment of researchers, albeit informed and justified, is not always shared by others, especially by the participants. For example, imagine an experiment on the influence of the emotional connotation of a text on the type of information drawn from it. In order to carry out this study, it would be necessary to choose or build texts conveying positive or negative emotions, as well as texts conveying neutral information. The judgment of a single person would be problematic in this case, because many parameters can influence the evaluation of the emotions conveyed by a text. These parameters may differ from respondent to respondent, but they may also differ in their relative importance regarding the attribution of emotions to texts. In this case, it would be compulsory to pretest the items and to obtain the evaluations of different people, in order to make sure that the experimental material is valid.

When the experimental material is a word list, numerous databases are available for obtaining the different relevant characteristics of the words. Different researchers and research groups have drawn up lists of existing databases for preparing experimental material. For example, this is the case for websites such as *The Language Goldmine*[12], *Experimental Linguistics in the Field*[13], the *Postdam Research Institute for Multilingualism*[14] or *OpenLexicon*[15], which we encourage you to consult, in order to become familiar with the various types of available information for creating experimental material. These websites also list resources specifically tailored to the choice of suitable non-words or standardized images following different criteria.

Lexical databases list many objective word characteristics, such as the number of syllables, phonemes and orthographic neighbors, or their frequency in the language. Different frequency indicators are often accessible, one being calculated on the basis of a corpus of texts and the other on the basis of a corpus of film subtitles, which better reflect word frequency in the spoken modality. Studies have shown that the frequencies drawn from the subtitle corpus predicted word reading time better than their written frequency (New *et al*. 2007; Brysbaert and New 2009).

[12] http://languagegoldmine.com/.

[13] https://experimentalfieldlinguistics.wordpress.com/experimental-materials/.

[14] https://www.uni-potsdam.de/en/prim/labs-experiments/resources-software-databases/online-databases.html.

[15] https://chrplr.github.io/openlexicon/ and http://www.lexique.org/shiny/openlexicon/ for online research.

It is also possible to access more subjective data, such as the age of word acquisition, their familiarity, their emotional valence, their concrete nature, their imagery value or their subjective frequency. These different characteristics are assessed on the basis of judgments performed by native speakers. As a result, this data is only available for a limited word sample in certain languages.

In English, we can turn to databases such as *CELEX*[16] (Baayen *et al.* 1995) or the *MRC Psycholinguistic Database*[17] (Coltheart 1981). The latter gathers information drawn from different sources, in order to provide data relating to 26 different linguistic properties. *Subtlex-US*[18] (Brysbaert and New 2009) and *Subtlex-UK*[19] (van Heuven *et al.* 2014) contain frequencies based on film subtitles. Subtlex databases also exist for Dutch (Keuleers *et al.* 2010), Chinese (Cai and Brysbaert 2010), German (Brysbaert *et al.* 2011), Greek (Dimitropoulou *et al.* 2010), Spanish (Cuetos *et al.* 2011), Italian (Crepaldi *et al.* 2015), Portuguese (Soares *et al.* 2015), Polish (Mandera *et al.* 2014) and French, the latter being accessible on *Lexique3*[20] (New 2006).

For 10 years, data relating to naming and lexical decision times for tens of thousands of words and non-words in different languages have been collected and made available to researchers. This data is accessible via the Lexicon Project in different languages; for example, in English, they can be found in *The English Lexicon Project*[21] (Balota *et al.* 2007) or in the *British Lexicon Project*[22] (Keuleers *et al.* 2012), in French in the *French Lexicon Project* (Ferrand *et al.* 2010), in Dutch (Brysbaert *et al.* 2016, Keuleers *et al.* 2010), in Chinese (Sze *et al.* 2016) or in Spanish (Aguasvivas *et al.* 2018).

Databases specifically related to children's lexicon are also available in different languages. In addition to the Subtlex databases which also include information on TV programs for children, statistics based on text corpora

[16] http://celex.mpi.nl/.

[17] https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm.

[18] http://www.lexique.org/?page_id=241.

[19] http://crr.ugent.be/archives/1423.

[20] www.lexique.org.

[21] https://elexicon.wustl.edu/index.html.

[22] http://crr.ugent.be/programs-data/lexicon-projects.

for children such as *the American Heritage Word-Frequency Book* (Carroll *et al.* 1971), or corpora of interactions with children like the CHILDES database (MacWhinney 2000) are available. TheChildFreq tool[23] (Bååth 2010) also makes it possible to search the CHILDES database in order to retrieve information relating to the interactions of American and British children.

However, it is possible that the research question involves examining a variable for which there is no published standard. It is then up to the researchers to select and validate the chosen words by implementing a pretest. For example, such a pretest was necessary for a study aimed at determining the influence of the context on the activation of color in mental representation, conducted by Connell and Lynott (2009). In this experiment, the participants read sentences describing objects whose typical color or alternative color was implied by the context. For example, one sentence described either a bear in the forest, which activated the brown color, or a bear at the North Pole, which activated the white color. Another story described either a ripe banana, which activated the yellow color, or an unripe banana, which activated the green color. In order to build this experimental material, it was first necessary to identify the objects, animals or plants which could take up different colors. A pretest was then conducted to determine the typical color of these objects and their alternative color, which were activated in most participants.

6.4.2. *Filler items*

Once the experimental items have been prepared, the filler items can be created so that the experiment is complete. The experiment includes this type of item for two reasons. First, they are essential for all tasks requiring a YES/NO response from the participants. In these tasks, the independent variable is generally manipulated for the items associated with the YES response. For participants to have the opportunity to answer NO to some of the elements presented to them, filler items associated with the NO response need to be added. In a lexical decision task, for example, filler items are non-words. In a verification task, they correspond to the elements which have not been presented in the text.

[23] http://childfreq.sumsar.net/.

Secondly, filler items make it possible to conceal the real purpose of the task from participants. We have already mentioned the fact that it is necessary to test naive people, for the results to not be biased. By concealing the goals of the task, we aim to reduce the risk of participants suspecting which independent variables are being manipulated. For example, in the study by Connell and Lynott (2009) presented above, filler items were sentences describing objects which had no typical color associated with them. In the filler items, it is also possible to manipulate a completely different variable other than the one actually being investigated, in order to divert the participants' attention. For example, in the study by Zufferey *et al.* (2015b) on the understanding of connectives by learners, some of the filler items included obvious grammatical errors, such as an incorrect subject–verb agreement. The number of filler items should generally be equal to or greater than the number of experimental items (Havik *et al*. 2009; Jegerski 2014).

6.4.3. *Other aspects of the material*

Experiments on language comprehension, implementing reading paradigms such as self-paced reading or eye-tracking, require participants to be presented with comprehension questions after reading some items, usually fillers. These comprehension questions aim to ensure that participants carry out the task properly by following the instructions, reading and trying to understand the sentences or short texts presented to them. Reading time or eye movement measurements obtained from participants who did not really process the text, but simply pressed the response keys only to move forward within the experiment, would be purposeless and may even risk preventing the demonstration of the effect. Comprehension questions are generally simple questions, expecting a YES/NO type of response. It is important to note that the degree of simplicity of these questions may influence the natural reading process, as participants sometimes develop strategies for answering these questions (Havik *et al*. 2009; Jegerski 2014).

6.4.4. *The concept of lists*

As discussed above, we encourage the use of repeated-measures designs, since they make it possible to assign the same participants to different experimental conditions. In the same way as participants, the items selected for the experiment represent only a sample of all possible items. These items also have their own characteristics which can influence the results, depending on the words or the types of sentences chosen. In order to control this influence, within-item designs should be implemented whenever possible. In other words, the same item should appear in all conditions. However, in most experiments, it is preferable to avoid having a participant see the same item in more than one condition, so as to prevent the familiarity effect. Indeed, the answer given during the second run risks being influenced by the fact that the item has already been seen and processed. For every item to be presented in each condition and for every participant to see every item only in one condition, it is necessary to organize items as lists. Every participant should see no more than one list during the experiment.

Let us go back to our fictitious example of an experiment on how the presentation modality might impact sentence comprehension. For this experiment, we imagine that we have created 40 items (40 sentences having the same structure). In order to set up a within-subject and within-item design, all of the items should appear once in their written form and once in their spoken form. Every participant should also be exposed to items in their spoken form and others in their written form. It would therefore be necessary to set up two lists of items. The first would contain items 1–20 in the written form and items 21–40 in the spoken form, whereas the second list would contain items 1–20 in the spoken form and 21–40 in the written form.

If we add the variable related to sentence complexity to this experiment, and we want to follow a within-subject and within-item design, every item and every participant should be confronted with the four possible conditions (all variables combined). In this case, we would have to create four lists according to the following model:

	List 1	**List 2**	**List 3**	**List 4**
Spoken-complex	Items 1–10	Items 31–40	Items 21–30	Items 11–20
Spoken-less complex	Items 11–20	Items 1–10	Items 31–40	Items 21–30
Written-complex	Items 21–30	Items 11–20	Items 1–10	Items 31–40
Written-less complex	Items 31–40	Items 21–30	Items 11–20	Items 1–10

Table 6.3. *List possibilities for an experiment*

6.4.5. *Number of items to be included in an experiment*

An important question when creating experimental items concerns the number of items that must be included in the experiment. It is impossible to answer this question in a definite manner, as it depends on the effect size, the task implemented and the characteristics of the material. The experiment by Connell and Lynott (2009), for example, contained only 10 items, due to the very rare specificities of the words used, namely the fact of representing an object with a clearly defined typical/atypical color. Conversely, some studies implementing lexical decision tasks include more than a hundred, and sometimes even several hundred items (e.g. Carreiras *et al.* 2005; Perea *et al.* 2015). For a long time, choosing the number of items was decided on the basis of what was regularly done in a specific field. Recently, it has been suggested to target a minimum of 1,600 observations per condition when measuring reaction time in repeated-measures designs (Brysbaert and Stevens 2018). This number of observations is the product of the number of participants and the number of experimental items per condition, representing 40 participants seeing 40 items, or 20 participants seeing 80 items, for example. The criteria to be considered when choosing these numbers depends on the task: for simple tasks, processing 80–100 experimental items and 80–100 filler items poses no problem. For more complex tasks, the accumulation of items may induce fatigue effects which should preferably be avoided. In such cases, it might be better to test fewer items on more participants.

6.5. Building the experiment

The need to randomize the order in which items are presented makes it difficult to collect data without using experiment presentation software or a dedicated web interface. These tools also make data collection easier, since the responses are recorded as a ready-to-use database. Some of these software or interfaces require a license, which can be expensive for institutions or individuals; this is the case of *EPrime* (Psychology Software Tools, Pittsburgh, PA) and *Qualtrics* (Qualtrics, Provo, UT), just to mention a few examples. The software *PsychoPy* (Peirce *et al.* 2019) and the online interface *PsyToolkit* (Stoet 2010, 2017) offer free and rather easy to use alternatives. We will not develop the characteristics of each of these interfaces in detail, but we encourage those interested to directly consult their documentation, which is available online. Many examples of experiments are also available.

However, in order to be able to program an experiment, it is necessary to get a good representation of the stages involved. We will describe these steps later in this chapter.

6.5.1. *Instructions*

Every experiment begins by clearly explaining to the participants how the task will unfold and what is expected from them. As we will see below, a task is made up of different trials, which are repeated a certain number of times and for which the participants have to perform the same action. For example, in a lexical decision task, a trial corresponds to the categorization of a string of letters into words and non-words. The instructions have to make it clear to participants how to give their answers. It is also essential to ask participants to keep their fingers on the answer keys throughout the experiment, in order to be able to react as quickly as possible. For a lexical decision task, the instructions could be as follows:

> In this experiment, you will perform a lexical decision task. This means that we will present you with strings of letters and that you will have to decide, for each of them, whether they form an existing word in English or not.
>
> The experiment will proceed as follows. First, you will see the message "Are you ready?" on the screen. When you are ready, you can press the YES key. This will bring up a fixation point in the center of the screen, for half a second. Please fix this point. The fixation point will then be replaced by a string of letters. At that moment, you will have to decide as rapidly and as accurately as possible, whether this string of letters is a word that exists in English or not. If it is a word, press the YES key. If it is not a word, press the NO key. The experiment will consist of 10 training trials, then 60 trials. It should last about 15 minutes. At the end of the experiment, you will see a message indicating that the experiment is over.
>
> Please place your forefingers on the YES and NO keys and keep them on these keys throughout the experiment. From the moment you start a trial, be sure to give your answer rapidly

> and accurately. You can take a break at any time when you see the message “Are you ready?”.
>
> If you have any questions, you can ask the person in charge of the experiment now.

6.5.2. *Experimental trials*

After presenting the instructions, the task begins. As we saw above, a task is divided into trials, each corresponding to an item. Depending on the task, the trials vary, but some characteristics remain the same, namely how items are presented and how responses are recorded. For online tasks, it is customary to precede every trial with a message such as “Ready to continue?”, to which participants reply YES in order to start the trial. This enables participants to prepare for the task and allows them to take a break when needed during the experiment. As soon as the trial begins, the elements are presented and responses to the different elements are recorded. When the experiment requires the recording of reaction times or reading times of single words, the presentation of the item (word or phrase) is generally preceded by a fixation point at the location where the item will appear, in order to attract the attention of the participants and reduce variations in the data collected. When recording the reading time of sentences or text segments, the use of a fixation point is not compulsory. On the other hand, in eye-tracking experiments, the accuracy of the measurement is very important and every trial begins with a fixation point. Figure 6.2 illustrates the different types of trials for some of the tasks presented in previous chapters.

The construction of experimental trials also requires defining how long the items will be presented. Sometimes, the participants set their own pace, as in self-paced reading tasks, in which pressing a key determines the progression in the experiment. In other cases, for example, in experiments based on a priming effect, it is necessary to precisely define the duration allocated to the presentation of the prime, and the time lapsed until the presentation of the target.

In order to prevent possible spill-over effects, the presentation of the items has to be randomized. In other words, for every participant, the order for presenting the items is established randomly. It is then highly unlikely that two participants will see the items in the same order. Randomization also makes it possible to avoid some items being systematically presented at the beginning or at the end of the task, which could lead to learning or fatigue effects, or cause the processing of a certain item to be regularly influenced by the one preceding it.

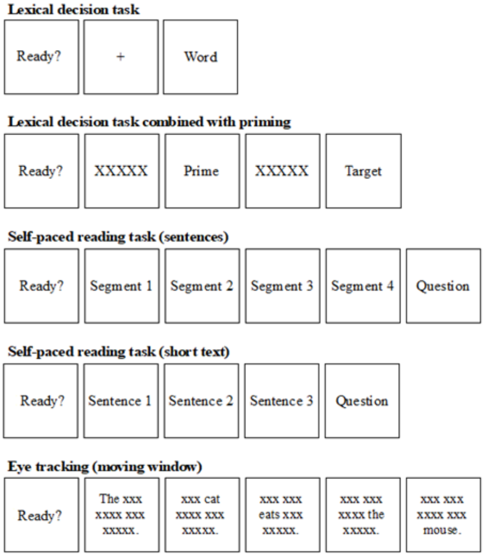


Figure 6.2. *Illustrations of experimental trials in different tasks in experimental linguistics*

Depending on the number of items and the type of experimental design, the number of trials may be very high, which could induce loss of attention or fatigue in the participants, and thus jeopardize the quality of the data collected. One solution to this problem may be to divide the trials into blocks, so that the experiment can be segmented into shorter portions. The use of blocks can also be convenient for presenting all the items in one condition, before moving on to another condition. For example, to study oral and written comprehension, it is preferable to present the sentences in only one modality before moving on to the other, since having to constantly switch from one modality to the other could pose additional difficulties to participants. When using blocks, it is advisable to try to counterbalance not only the order of presentation of the items within the blocks, but also the blocks themselves.

For participants to get used to the task, the experiment begins with the presentation of training items, during which the experimenter can verify the correct understanding of the instructions and re-explain them if necessary. The experiment ends when all the items have been presented. At this point, the experimenter usually thanks the participants and answers any questions they may have. It may also be useful to ask the participants for feedback on their perception of the experiment and to survey their intuitions on the nature of the question being investigated, in order to determine whether their behavior may have been influenced in a way that could affect the experiment's quality.

6.6. Data collection

For data collection to take place, it is necessary to recruit participants. Very often, participants are university students who voluntarily take part in studies proposed in their field, or who take part in them in exchange for credits or a sum of money. Recruitment is simply done by posting ads which briefly present the research project and provide contact details for enrollment.

When the research question requires participants with a specific profile, such as a certain L2 proficiency level or certain cognitive skills, two solutions may be contemplated. The first is simply to test voluntary participants and then build groups based on their individual differences, once the data has been collected. This method, which is simple to implement, nonetheless entails various risks. First of all, it is likely that the groups obtained do not have a similar size, which is something that may cause problems when analyzing the data. It is also possible that the participants' individual differences do not make it possible to establish clearly different groups, or may produce groups with high scores, or on the contrary, very low scores regarding the characteristic of interest. For example, Gillioz *et al.* (2012) grouped their participants depending on their level of empathy, in order to verify the potential influence of that variable on their comprehension of emotions while reading. Since the participants were for the most part students in psychology, the groups, while differing from each other, still presented rather high scores compared to the standard in the general population. As the results of the task on emotional inferences did not show any differences between groups, it was difficult to determine whether empathy had an influence on emotion comprehension (or not), or whether the groups tested did not make it possible to shed light on such influence. In order to avoid these problems, it may be useful to set up a preliminary selection of participants, by testing for the desired variable and then only including the people which correspond to the desired criteria of the study. For example, this can be useful for recruiting participants who share a certain linguistic profile.

For some research questions, it may also be necessary to recruit people from specific populations, such as people with autism spectrum disorder (ASD) or people with aphasia. It is therefore necessary to contact the competent institutions and associations in order to be able to gain access to these people. In general, access to such populations is rather restricted and requires important administrative work. Even more than with neurotypical participants, the ethical principles developed in the next section must be guaranteed in the experiments involving these populations.

The rest of data collection may take different forms depending on whether the participants are tested by meeting them in person, or remotely, by means of a questionnaire, or a task which can be completed on the Internet.

In the first case, data collection begins with the welcoming of the participants, where they are presented with the study and the task to be completed. This step has two main purposes. The first is to make participants feel at ease for the rest of the procedure. The second is to obtain their consent (see the next section on research ethics), which is a prerequisite for any empirical research with humans. Once the participants have provided their written consent, their demographic data is usually collected. In general, information such as gender, age, mother tongue and educational level is relevant for linguistic studies. The information collected at this stage will make it possible to describe the participants when the results are communicated afterwards. Depending on the research question or method, other types of information may be required. This is the case for laterality (left-handedness/right-handedness) for experiments measuring response times, because the participants must provide the YES responses with their dominant hand. During eye-tracking experiments, it is useful to check the participants' visual acuity and to take note in case glasses or lenses are worn.

The standardization of the procedure is extremely important in order to grant the quality of the results. Care must be taken to present every participant with the same test conditions. In case data collection is always carried out by the same person, it is very important to maintain the same attitude with all the participants and to ensure that the instructions are identical for everyone, being specially trained in advance and making sure to use the same formulations. When several people are in charge of the data collection, every experimenter should test participants in all conditions, in order to not induce bias associated with the experimenter, or, at least, to keep such bias under control.

Collecting data in the laboratory has the advantage of being able to observe the participants, to monitor their behavior during the experiment and to interact with them in order to determine their impressions. This allows us to enrich the data collected with observations made during the completion of the task. It is also important to keep track of the participants' involvement, by logging relevant information for every subject (such as time, the experimenter, any problems encountered during the task or any other observation that may be useful later). For example, it is useful to record the cases where participants show fatigue, or find the task difficult to perform, or even find out the variables that are being manipulated in the experiment. However, laboratory data collection has the disadvantage of being costly in terms of time, material resources and staff.

Whenever possible, one solution is to turn to the Internet, specifically to the various online data collection platforms. In this case, recruiting participants can be done in a much broader way, through social media, for example. It is also possible to ask each participant to forward the link of the experiment to one or more acquaintances, which quickly generates a snowball effect that will help when acquiring new data. Finally, there are also websites linking researchers with participants. For example, this is the case for *Amazon Mechanical Turk*[24] or *Prolific*[25], the latter being specifically intended for research. In order to recruit participants, it indicates the number and characteristics of the desired people. It is thus easier to access French-speaking women between 30 and 45 years old, for example. People recruited through these platforms are paid to participate in the experiment, which means there has to be a budget available.

[24] https://www.mturk.com.

[25] https://www.prolific.co.

Online studies also have the advantage of being able to test a higher number of participants as well as a wider variety of people. Consequently, they are more generalizable and have a higher ecological validity than laboratory studies (Reips 2000). Several studies have recently shown that studies carried out on the Internet offer results quite similar to studies carried out in the laboratory (Reimers and Stewart 2007; Schubert *et al.* 2013; Kim *et al.* 2019), opening new avenues for this form of experimentation.

6.7. Ethical principles

Before finishing this chapter devoted to the practical aspects of experimentation, we will present the different facets of ethics involved in experimental research. Ethical questions arise at different stages of the research process, not only during the conceptualization phase, but also when recruiting participants and when publishing results. We will develop these various aspects below, without dwelling on the principles of integrity relating to the general principles of research. Readers who are interested can turn to the codes of conduct drawn up by various research institutions, such as, the European Federation of Academies of Sciences and Humanities (ALLEA).

In experimental linguistics, it is generally essential to recruit participants in order to obtain the data which will allow us to answer research questions. Most participants take part in experiments which do not involve significant risks or benefits to their health, which could be the case in other disciplines, such as medicine. It is, however, necessary to respect certain ethical principles in order to ensure the respectful treatment of participants. Most importantly, this requirement implies their right to confidentiality. Data protection is a legal obligation, and researchers have to determine in advance how the data will be anonymized and then stored, who will have access to it and the way in which it will be used in publications or public presentations. Data confidentiality is essential to the trust between all those involved in the study and must be ensured throughout the research process.

Another important element related to ethics concerns the well-being of the participants during and after the study. We must therefore ensure that the participants do not leave in a degraded condition, compared to their initial condition. In the majority of linguistic studies, the only risk that participants face is getting bored during the experiment. However, in some cases, the research question may relate to a characteristic of the participants, such as intelligence, memory, specific skills or an impairment, such as ASD or dyslexia, for example. The evaluation of these characteristics should then be done in a neutral manner, to avoid judging the participants, or categorizing them openly, for having a higher or lower level concerning these characteristics. When it is necessary to perform experiments with particular populations, such as children, illiterate people or people suffering from ASD, every precaution must be taken to avoid unpleasant moments for them.

The well-being of participants can also be endangered when research is manipulating the conditions in which language is produced or understood. Going back to an example discussed in Chapter 1, it might be interesting to examine the influence of stress on articulation rate, which might require placing participants in stressful and less stressful conditions. This study would therefore need to find a way to stress some of the participants, without this stress having an excessively negative impact on them. At the end of the experiment, it would then be essential to eliminate the stress induced by the manipulation, either by debriefing the participants or by offering them a moment of relaxation before they go home. Another example of research, which could affect the well-being of participants, is that of Eilola *et al.* (2007), presented in Chapter 5, in which participants were presented with emotionally loaded words, including words with negative connotations and taboo words. The presence of such words can offend some sensitive people, and it is therefore necessary to warn them that the experiment includes such material. This enables participants to make an informed decision whether or not they consent to getting involved in the experiment.

Finally, it is important to ensure the equality of participants between conditions, when these can influence their reality in one way or another. For example, in order to study the effectiveness of a language learning method that it is very likely to offer better results than other methods, it is necessary to inform the participants about this difference. Once the study is finished, it would be desirable to offer a catch-up to participants in the more "disadvantaged" group.

In summary, ethical questions arise at different levels of a research project. Ethical principles are first taken into account by the researchers at the conceptualization stage of research and are then generally submitted to an Ethics Committee, which decides on the respect of the ethical principles for the suggested research project. If these are considered adequate, the Committee gives the green light and the study can be carried out.

Any scientific research complying with ethical principles compulsorily has to collect the *free and informed consent* of participants. This implies that participants can freely decide to take part in a study, in an informed manner. In other words, participants must receive complete and honest information about the study, the task to be completed and the potential positive or negative consequences of this task. Moreover, the participants must be able to decide to participate freely, without any external constraints linked to an advantage or loss of an advantage. It is also essential to inform participants about the possibility of ending their participation at any time during the study or even after, by requesting the withdrawal of their data. In order to attest to the participants' consent, researchers must collect their signature on a written document. This document should generally:

– adequately present the content of the research project to participants;

– present the task that the participants will have to complete;

– present the risks, side effects and possible benefits associated with participation;

– mention the total freedom to participate in the study, the possibility of withdrawing at any time and the procedure to follow in the event of withdrawal;

– provide contact information for further details on the study.

6.8. Conclusion

In this chapter, we have reviewed various practical aspects which are useful for creating an experiment. First, we introduced the sources we can consult for formulating a research question, as well as the means of accessing these. We have seen that the research question needs be operationalized by defining the levels of the independent variables, as well as indicating how the dependent variable will be measured. External variables that could influence the results also have to be determined at this stage in order to choose the appropriate experimental design. This experimental design may include independent (between-subject) or repeated (within-subject) measurements. We have described the advantages of repeated-measures designs, which enable a better control of external variables, as well as their limitations, which must be taken into account when building the experiment. These may refer to spill-over effects, which can be controlled by counterbalancing the conditions and/or by randomizing the items' order of presentation. Item lists are essential in repeated-measures designs and we have shown how to build them. We then described factorial designs, involving several independent variables, as well as different effects (main and interaction) which can be observed in this type of design. In the second part of the chapter, we discussed the important elements that need to be respected when building experimental material and we presented resources for selecting this material. We have seen that the material is made up of experimental items and filler items, which allow the task to be carried out while concealing its objectives. We then discussed the various stages of the experiment itself, how to recruit participants and how to collect data. We concluded the chapter by describing the ethical principles inherent in research on human beings and the main elements to be observed in this context.

6.9. Revision questions and answer key

6.9.1. *Questions*

1) Transform the following hypothesis into an operational hypothesis, then schematize it by including the external variables that you consider the most important, in terms of items and participants: "A person's accent influences the credibility of what he or she says."

2) Choose how to counterbalance the conditions in the following situations:

a) An experiment studying the influence of sentence complexity on the comprehension of anaphora in children.

b) An experiment studying the influence of a concurrent task in working memory (participants must remember strings of letters in parallel with the reading task) on the construction of predictive inferences while reading.

3) An experiment aims to study the influence of grammatical gender on the representation of common nouns. Based on Borodistky *et al.* (2003; see section 2.5), you ask French-speaking participants to choose associations of

common nouns and first names, which can either be of the same gender or a different one.

a) How do you choose common nouns and first names? What are the variables to be controlled?

b) Choose the words to create a dozen pairs in French.

c) Create lists to implement a repeated-measures design. Every item should be presented in the different conditions without the participants seeing the same item several times.

4) Write the instructions for a self-paced reading experiment comprising of 40 items. Every item corresponds to a 5-sentence short story describing a situation in everyday life, and whose fourth sentence is the target sentence. Stories are sometimes followed by questions.

5) Write the free and informed consent form for that same experiment.

6.9.2. *Answer key*

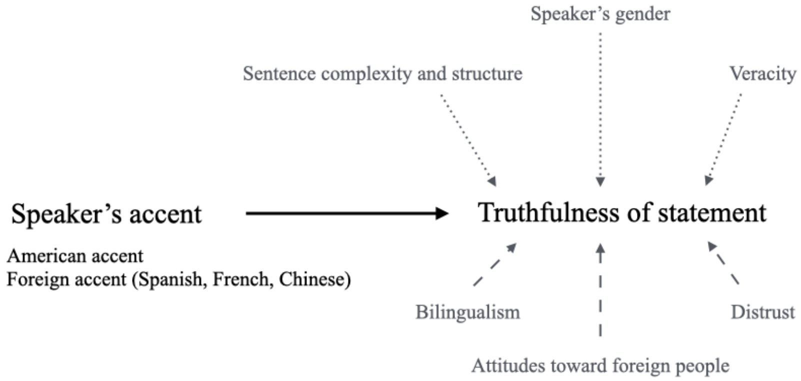
1) There are different ways of approaching this question. Here, we will follow the assumption that people give less credibility to statements made by a person speaking with a foreign accent than by a person without an accent. To study this question, we suggest recording statements made by speakers with or without a foreign accent, to present to participants and ask them to assess, on a scale from 1 to 10, the truthfulness of such statements. By comparing the scores obtained in the different conditions, it would be possible to determine a connection, if existent, between foreign accent and credibility. Let us imagine that we will test native American-English speakers.

For this study, we can identify different variables to control. First of all, at the item level, the veracity of the information to be evaluated should be checked. It would be appropriate to present true and false statements, so that participants can respond to a full range of credibility. Secondly, the structure and complexity of the items should be kept constant, so that these parameters do not influence the truthfulness judgment. Thirdly, it might be necessary to present as many statements as possible, that are equally uttered by male and female speakers, in order to control a potential influence of the speaker's gender on credibility. It is possible that the participants consider men more credible than women, due to the existence of certain stereotypes in society. The last very important aspect to check would be the speaker's accent. It would be appropriate to vary the accents in order to generalize the results.

As far as participants are concerned, the main variable that could interfere with the variables being investigated in the study, relates to their general attitude towards people with a strong foreign accent. This is undoubtedly influenced by their general attitude towards foreigners, and it might be useful to measure this variable in order to take it into account when analyzing the results. The attitude towards speakers with foreign accents may also be influenced by a subject speaking one or more languages, as bilingual people probably tend to be more tolerant of a foreign accent than monolingual people. Finally, the level of mistrust in relation to a statement certainly varies from one person to another, and this would also be a variable to be kept under control.

The best solution for this experiment would be to build a within-subject design, that is, a design where every participant sees all the conditions of the experiment, and a within-item design, in other words, a design where every statement is presented either with or without a foreign accent.

The operational hypothesis, as well as the variables to be checked, can be schematized as follows:


Speaker's gender
Sentence complexity and structure
Veracity
Speaker's accent
Truthfulness of statement
American accent
Foreign accent (Spanish, French, Chinese)
Bilingualism
Distrust
Attitudes toward foreign people

2) a) The conditions of this study are related to the complexity of the sentences presented. This means that children see more complex sentences and less complex sentences. In order to counterbalance the conditions, the

best solution would be to present the sentences in the two conditions randomly, without separating the conditions themselves. Every child could see one or more complex sentences, before seeing one or more less complex ones, and then see complex sentences again, and so on.

b) In this experiment, the conditions relate to the manipulation of the working memory. This implies that, for half of the sentences in the experiment, participants only have to read them, whereas for the other half of the sentences, a working memory task has to be performed in parallel with the reading task. In this case, the most appropriate technique would be to separate the experiment into two blocks, depending on the working memory condition. The order of presentation of the blocks would then have to be alternated among participants. Furthermore, the different sentences should be presented in one block or the other, and the presentation order of the sentences should be random.

3) a) In order to answer this question, it is necessary to identify which variables can influence the memorization of a pair, consisting of a common noun and a first name. One possible variable is whether the common noun represents a living being, or not. The pair COW-AGATHE seems intuitively easier to remember than the pair SPOON-AGATHE. It would therefore be appropriate to decide to only test pairs with inanimate objects. A second variable that can influence the memorization of pairs is the frequency or length of the different words included in the task. For French common nouns, we can consult *Lexique* to find out their frequency and their length. Concerning first names, there are statistics provided by the national statistical institutes, on the classification of first names during the last decades. These may help to control to what extent participants are familiar with such names (depending on their age group).

b) The pairs should be made up of male and female common nouns and male and female first names. For the example, we searched on *Lexique* for common nouns with a length oscillating between three and four syllables, and a frequency between 10 and 100 appearances per million (in books and movies). We chose the following nouns: *ambulance, batterie, caméra, cigarette, pharmacie, télévision, ascenseur, canapé, escalier, hélicoptère, magazine* and *pantalon.* In order to choose the appropriate first names, we consulted the classification of the most widely spread first names in France for the period 1995–2000, on the National Institute of Statistics and

Economic Studies[26] website. We retained the first names Manon, Camille, Pauline, Marie, Chloé, Sarah, Thomas, Clément, Maxime, Lucas, Quentin and Julien.

c) In order for each item to be presented in each condition, it should appear once associated with a male name, and once with a female name. It is therefore necessary to create two lists, which should each present half of the male and female common nouns associated with a male first name, and the other half with a female first name. One possibility would correspond to the following lists:

List 1:

ambulance-Thomas, batterie-Maxime, caméra-Julien, cigarette-Manon, pharmacie-Camille, télévision-Chloé, ascenseur-Clément, canapé-Lucas, escalier-Quentin, hélicoptère-Pauline, magazine-Sarah, pantalon-Marie.

List 2:

ambulance-Manon, batterie-Camille, caméra-Chloé, cigarette-Thomas, pharmacie-Maxime, télévision-Julien, ascenseur-Pauline, canapé-Sarah, escalier-Marie, hélicoptère-Clément, magazine-Lucas, pantalon-Quentin.

4) In this experiment, you will read short stories that are five sentences long, describing situations in everyday life. The goal is to read these stories in a natural way, as you would have if you were at home. Before each story, you will see the message "Ready to continue?". When you are ready, press the YES button to bring up the first sentence in the story. Read the sentence, then press YES to go to the next sentence and so on until the end of the story. Some stories will be followed by simple questions about the story. If such a question appears, you must answer the question as quickly and as accurately as possible, by pressing either YES or NO. It is very important to read each story without stopping. If you want to take a break during the experiment, it is possible to do so when you see the message "Ready to continue?". Please keep your fingers on the YES and NO keys during the whole experiment, so that you can easily progress through the stories and answer the questions. The experiment will start with some training stories. If

[26] https://www.insee.fr/fr/statistiques/3532172.

you have questions, you can ask the experimenter. The experiment will last between 20 and 30 minutes.

5) In this study, we are interested in the process of reading comprehension. Please read the explanation of the experiment you are going to take part in, as well as the risks and benefits it may present, before deciding to participate.

If you agree to take part in this study, you will complete a reading task presented on a computer screen. It will take between 20 and 30 minutes.

You will not get any direct benefit from this experiment, but it will allow us to improve our knowledge about the comprehension processes at work while reading texts. As compensation, you will receive 10 Euros.

There is no direct risk associated with your participation in this experiment, except that of feeling bored. Participation involves an investment of 20–30 minutes of your time.

You are free to accept or refuse to take part in the study. You can now choose not to take part in it. If you choose to participate, you can still withdraw from the study at any time, without any need for justification. If you take part in the study and decide to withdraw from it following your participation, you can ask for your data to be deleted. In all cases, the 10 Euros compensation will be given to you.

All the data obtained during the experiment will be treated in strict confidence. You will only be identified by a randomly assigned number, and neither your name nor any means of identification will appear anywhere. No data identifying you will be used in the publications or presentations which result from this study.

At any time, you may ask questions or request further details from Ms. X, Address, Phone No.

6.10. Further reading

Gonzalez-Marquez *et al.* (2007a) present the structure of a scientific article, explain how to read such sources and detail the stages involved in the literature review. Abbuhl *et al.* (2013) develop the advantages and limitations of different experimental designs, as well as the particularities of research carried out on children. These two sources also review other general principles to take into account when developing an experimental study. Jegerski (2014) presents the construction of the materials used in self-paced reading experiments in detail, such as experimental items, filler items and comprehension questions. Kim *et al*. (2019) present the advantages of studies conducted in the laboratory or on the Internet, as well as the results of their study comparing these two methods, with a task involving choice reaction time. For the ethical principles associated with scientific research, there are various documents published by the national research societies. It is relevant to refer to the specific recommendations of the country where the study is conducted.

7

Introduction to Quantitative Data Processing and Analysis

This chapter presents the basic principles for the analysis of quantitative data. We start by describing how raw data is generally organized after the data collection. We then show that data follows a certain distribution, and we present one distribution in particular: the normal distribution. We also discuss the different ways to visualize and describe data. The second part of the chapter deals with data modeling for statistical tests, before describing the logic underlying such tests. Then, we briefly discuss some tests that have traditionally been applied to linguistic experimental data, and point out the inherent limitations in these. We see that nowadays there are more reliable models for analyzing this particular type of data, through the use of mixed linear models, for example. We present these models, as well as the results obtained with such analyses. We end the chapter by discussing some questions that may arise while analyzing data.

7.1. Preliminary observations

In Chapter 2, we saw that there are different types of variables which can be measured along different scales. The dependent variables used in linguistic experiments are generally measured along continuous scales, such as reaction or reading times, the number of items retained after reading a text, or an item's acceptability on a scale from 1 to 10. Independent variables are often categorical in order to compare the data collected in the conditions

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

of experiments. In this chapter, we will focus on the principles and analyses which are specifically related to this type of variable.

Data analysis requires a good understanding of different mathematical and statistical principles, the complexity of which can vary widely. In this introductory chapter, we will not be able to go into the detail of mathematical and statistical models. Hence, we encourage those interested in the concepts explained in this chapter to deepen their knowledge of them using the suggested reading at the end of the chapter.

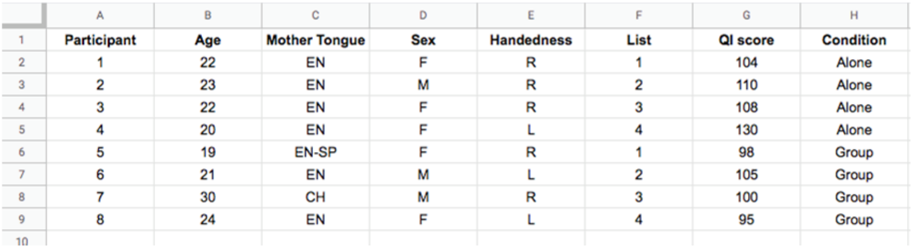
Most of the analyses presented in this chapter also require the use of statistical software. Today, there is a very powerful and free access tool, *R*[1] (R Development Core Team 2016), which has become the standard in language science. This software certainly requires a period of familiarization in order to understand and to learn how to apply the necessary codes for different functions. However, this training time quickly pays for itself, since the possibilities offered by *R* are prolific. Besides, *R* relies on a community of researchers who create and share *packages*, or in other words, reproducible code units, and who also provide documentation, software updating and technical support. When communicating the results of research, it is also increasingly expected to make available the data as well as the code on which the results are based, in order to favor the reproducibility and the openness of science. The use of *R* meets these expectations, enabling the proper management and recording of all the stages involved: processing, visualization and data analysis. Not only due to all the advantages mentioned, but also because this software is widely used within the scientific community, we encourage beginners to turn to *R*. An excellent basis to start learning statistics and to discover their application with *R* can be found in Winter's book (2019), which is specially intended for researchers in language sciences.

7.2. Raw data organization

At the end of the data collection, a lot of data are integrated into one or more databases, depending on the technique used in the experiment. In order

[1] https://www.r-project.org.

to illustrate a first possible type of database, let us consider the one devoted to participants, which should contain relevant demographic information identified by researchers. This database could take the form shown in Table 7.1 for participants 1 to 8 in a fictitious experiment.

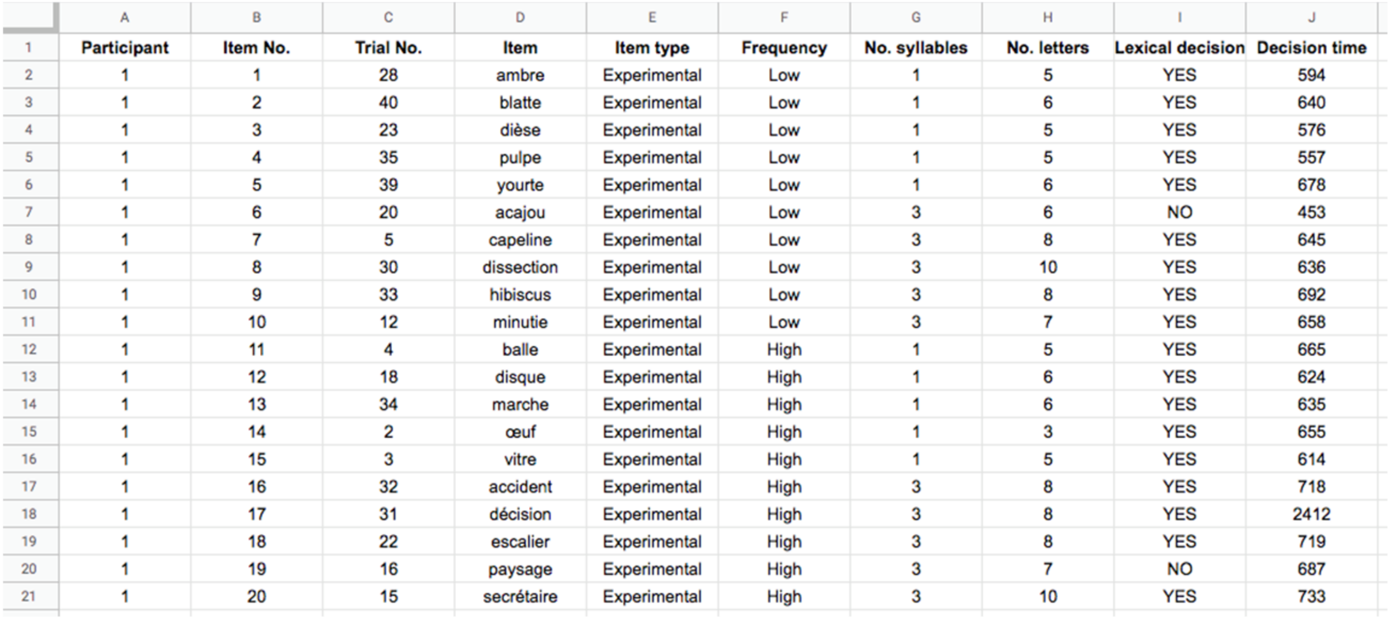


	A	B	C	D	E	F	G	H
1	Participant	Age	Mother Tongue	Sex	Handedness	List	QI score	Condition
2	1	22	EN	F	R	1	104	Alone
3	2	23	EN	M	R	2	110	Alone
4	3	22	EN	F	R	3	108	Alone
5	4	20	EN	F	L	4	130	Alone
6	5	19	EN-SP	F	R	1	98	Group
7	6	21	EN	M	L	2	105	Group
8	7	30	CH	M	R	3	100	Group
9	8	24	EN	F	L	4	95	Group
10								

Table 7.1. *Example of a database describing the characteristics of participants*

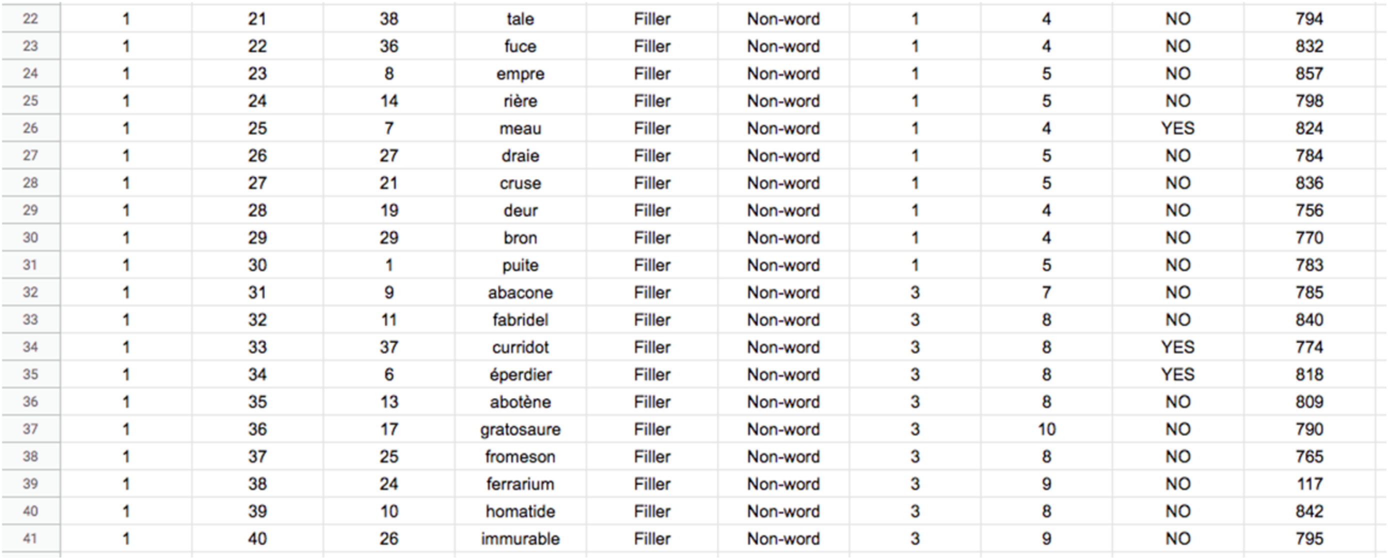
In this type of database, every row corresponds to one participant and every column is related to a demographic variable (age, mother tongue or laterality, for example) or to a variable included in the experiment (such as the list or condition assigned to each person, for example). In Table 7.1, we can see that lists were assigned to the participants in a sequential manner, as well as the conditions in which the experiment was carried out (in a solitary manner or in a group). In this example, we can also see a number representing each person. This is to respect the ethical principle of anonymity inherent in research. In fact, at no time should the identity of a participant be related to his/her performance in the task.

The data collected during the tasks themselves are presented in a similar way, except for the fact that every row now corresponds to an item in the experiment, rather than to a participant. In an experiment comprising 20 experimental items, 20 filler items and 20 participants, the database would thus include 800 rows, 40 for every participant and 20 for every item. This type of coding is called long format data. Table 7.2 illustrates, for one participant, the fictitious data obtained during a lexical decision task, in which word frequency and word length (in syllables) were manipulated. For every item included in the experiment, we can see the order in which it appeared, its type (experimental or filler item) and the different frequency and length conditions. The answers given by the participant and the time associated with every answer are also shown.



	A	B	C	D	E	F	G	H	I	J
1	**Participant**	**Item No.**	**Trial No.**	**Item**	**Item type**	**Frequency**	**No. syllables**	**No. letters**	**Lexical decision**	**Decision time**
2	1	1	28	ambre	Experimental	Low	1	5	YES	594
3	1	2	40	blatte	Experimental	Low	1	6	YES	640
4	1	3	23	dièse	Experimental	Low	1	5	YES	576
5	1	4	35	pulpe	Experimental	Low	1	5	YES	557
6	1	5	39	yourte	Experimental	Low	1	6	YES	678
7	1	6	20	acajou	Experimental	Low	3	6	NO	453
8	1	7	5	capeline	Experimental	Low	3	8	YES	645
9	1	8	30	dissection	Experimental	Low	3	10	YES	636
10	1	9	33	hibiscus	Experimental	Low	3	8	YES	692
11	1	10	12	minutie	Experimental	Low	3	7	YES	658
12	1	11	4	balle	Experimental	High	1	5	YES	665
13	1	12	18	disque	Experimental	High	1	6	YES	624
14	1	13	34	marche	Experimental	High	1	6	YES	635
15	1	14	2	œuf	Experimental	High	1	3	YES	655
16	1	15	3	vitre	Experimental	High	1	5	YES	614
17	1	16	32	accident	Experimental	High	3	8	YES	718
18	1	17	31	décision	Experimental	High	3	8	YES	2412
19	1	18	22	escalier	Experimental	High	3	8	YES	719
20	1	19	16	paysage	Experimental	High	3	7	NO	687
21	1	20	15	secrétaire	Experimental	High	3	10	YES	733

Table 7.2a. *Example of a long format database (where one row corresponds to one item)*



22	1	21	38	tale	Filler	Non-word	1	4	NO	794
23	1	22	36	fuce	Filler	Non-word	1	4	NO	832
24	1	23	8	empre	Filler	Non-word	1	5	NO	857
25	1	24	14	rière	Filler	Non-word	1	5	NO	798
26	1	25	7	meau	Filler	Non-word	1	4	YES	824
27	1	26	27	draie	Filler	Non-word	1	5	NO	784
28	1	27	21	cruse	Filler	Non-word	1	5	NO	836
29	1	28	19	deur	Filler	Non-word	1	4	NO	756
30	1	29	29	bron	Filler	Non-word	1	4	NO	770
31	1	30	1	puite	Filler	Non-word	1	5	NO	783
32	1	31	9	abacone	Filler	Non-word	3	7	NO	785
33	1	32	11	fabridel	Filler	Non-word	3	8	NO	840
34	1	33	37	curridot	Filler	Non-word	3	8	YES	774
35	1	34	6	éperdier	Filler	Non-word	3	8	YES	818
36	1	35	13	abotène	Filler	Non-word	3	8	NO	809
37	1	36	17	gratosaure	Filler	Non-word	3	10	NO	790
38	1	37	25	fromeson	Filler	Non-word	3	8	NO	765
39	1	38	24	ferrarium	Filler	Non-word	3	9	NO	117
40	1	39	10	homatide	Filler	Non-word	3	8	NO	842
41	1	40	26	immurable	Filler	Non-word	3	9	NO	795

Table 7.2b. *Example of a long format database (where one row corresponds to one item) (continued)*

7.3. Raw data processing

Now that we have described different formats of raw databases, we will move on to the inspection of these data. Let us go back to the responses by participant #1 (Table 7.2) and take a look at them. We can see that the participant answered correctly in the majority of cases. In fact, there are only two NO replies related to words and three YES replies related to pseudowords. The participant presents a correct response rate of 87.5% (35/40), suggesting that she has understood the task. This inspection of the responses given by the participants is an important step that should be carried out before analyzing the data, in order to exclude those who have not understood the task or who have not followed the instructions. The exclusion criteria should be clarified before observing the data, on the basis of theoretical principles, or the criteria used in prior research projects.

Now, let us take a look at this participant's response times. These are generally around 700 milliseconds (ms), which further confirms that she followed the instructions and gave her answer quickly. Nevertheless, two response times are very far from the others. This is the case for item #17, associated with a response time of 2,412 ms, and item #24, with a time of 100 ms. Due to their distance from other response times, these particular measurements may not reflect the processes involved in a lexical decision task, but other processes. The time corresponding to 117 ms is very probably related to a participant's error, who may have pressed the NO key prematurely. It seems highly unlikely that a person would be able to read the word and then categorize it in such a short time. The 2,412 ms time is more difficult to interpret. It could arise from a difficulty in processing, and then categorizing the word *decision*, thus suitably reflecting the processes investigated in the experiment. However, this time could also result from other processes, such as a decrease in concentration, for example. In the same way as the decision to exclude a participant from the experiment, the decision to exclude this type of response time must comply with the criteria chosen before data collection. We will return to this point later.

A final step before analyzing the data is to determine and to select which data will be analyzed. In this example on the lexical decision time, the data analysis should let us take a glimpse of the relationship between word recognition time and the variables *word frequency* and *word length*. In order to do this, only the word-related decision times should be taken into account, excluding the times related to non-words, as these were not related to the research question but were only used to make the lexical decision task possible. Among the response times for words, it is also necessary to consider only the times related to correct answers (or YES replies), showing that the words have actually been recognized as such.

7.4. The concept of distribution

The data acquired in an experiment can be summarized using different indicators. When communicating the results, it would be inappropriate to present all the individual data obtained in the experiment, as this would in no way be informative. Rather, a relevant summary of these data should be provided, thus enabling those interested to quickly understand the results. But before summarizing the data, it is essential to observe their distribution, that is, the frequency of the different values collected. This can be done through the use of a histogram, such as the one presented in Figure 7.1, which summarizes the fictitious data (YES correct responses) acquired for all the participants in the study described above, except for the extreme value.

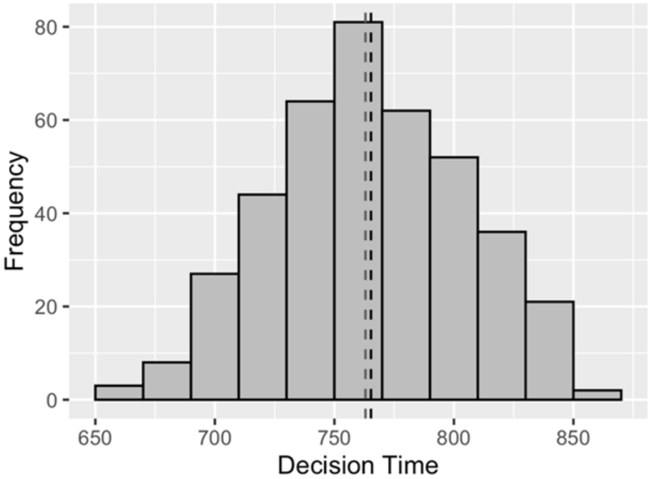


Figure 7.1. *Histogram representing the distribution of data acquired in an experiment. The black dashed line designates the mean of the distribution, whereas the gray dashed line is the median*

On a histogram, different values are grouped into classes, the width of which can be adapted (here, the classes represent 20-ms intervals), and the height of which corresponds to the number of values contained in each class. Different types of information may be deduced from this histogram. First, the distribution has a single peak, between 750 and 770 ms, meaning that it is unimodal. Second, this peak is located at the center of the distribution and there is no positive or negative asymmetry. Third, we can observe that there are no extreme values located far from the other values. This distribution approximately corresponds to a normal distribution, since it is centered symmetrically around the central class, and presents a gradual decrease in the frequency of classes as we move away from the center. The normal distribution corresponds to a theoretical distribution for modeling the data observed empirically. It is generally represented by a probability density function, which indicates the probability of observing certain values. Two parameters, the mean and the standard deviation, define the distribution. The mean corresponds to the center of the distribution and indicates its location on the *x* axis. The standard deviation can be considered as the mean deviation around the mean. For a theoretical normal distribution, 68% of the data are located in the area between –1 and +1 standard deviation from the mean, 95% in the area between -2 and +2 standard deviations and 99.7 % in the area between –3 and +3 standard deviations (see Figure 7.2). This means that if a value were to be chosen randomly from such distribution, in 68% of the cases, it would correspond to a value placed within one standard deviation of the mean, and in 95% of the cases, it would correspond to a value placed within two standard deviations of the mean.

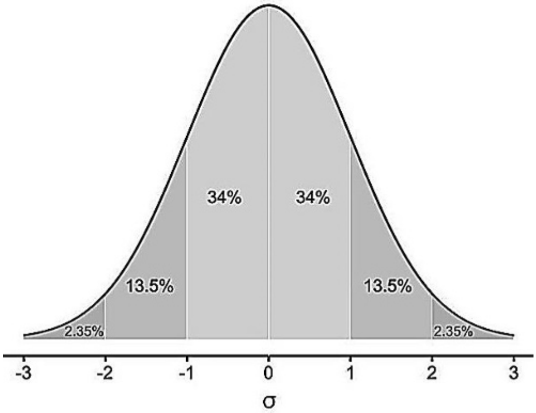


Figure 7.2. *Normal distribution, with a mean of 0 and a standard deviation of 1*

Many statistical tests are based on the normal distribution, as it corresponds to the form that likelihood takes. By analogy, we can estimate that many variables are distributed in a normal way in the population.

7.5. Descriptive statistics

To summarize the data, we generally report the center of the distribution, also known as the central tendency. The central tendency can be measured in three ways. A first way – which we have already discussed – is the mean, which can be simply obtained by adding the values and then dividing the total obtained by the number of values. For a normal distribution, the mean indicates the location of the center of the distribution. In our example, it would correspond to 765 ms (the black dashed line in Figure 7.1). However, the mean alone cannot summarize data in an informative manner because a distribution is also defined by its dispersion, which can be evaluated by the standard deviation, for example.

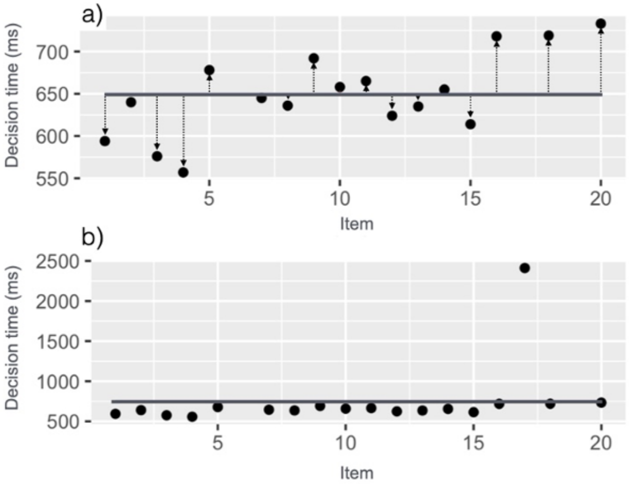


Figure 7.3. *Graphic illustrations of decision times by participant #1 and the mean for correctly recognized experimental items (gray line). Panel (a) does not contain the extreme value of 2,412 ms, whereas panel (b) contains it*

To illustrate this, let us consider the correct YES decision times by participant #1 and plot them on a graph. Figure 7.3, panel (a), shows the mean (gray line) for all the data concerning participant #1, with the exception of the time of 2,412 ms. We can see that every value deviates from the mean by a certain distance (arrow). In order to quantify the total distance, we might imagine adding the individual distances. However, this is not a good solution, since the distances of the values below the mean are compensated by those located above the mean, and their sum would therefore be equal to 0. In order to remedy this problem, it is necessary to transform negative values into positive ones, which can be achieved by squaring them. The variance of the sample is thus calculated on the basis of the sum of the squares of the distances from the mean, divided by the number of observations minus one. By calculating the square root of the variance, we obtain the sample's standard deviation, which can be considered as an indicator of the average distance from the mean, which can be seen as the average error around the mean. In the example we are interested in, the mean is 649 and the standard deviation is 49. In other words, the central value of the distribution is 649 ms and the data move away from it by 49 ms on average. Let us now observe the effect on the mean and the standard deviation when we add the 2,412 ms value (Figure 7.3, panel (b)). Introducing this value into the calculation would increase the mean to 747 and the standard deviation to 418. This illustrates that a single value can strongly influence the mean and the standard deviation of a distribution. This type of value is called an extreme value.

Managing extreme values is a complicated issue for researchers. While there is general consensus that extreme values corresponding to impossible values, for example, resulting from a coding error, should be eliminated from the data, there is no clear procedure concerning extreme values such as the one we have here. In the literature, we find two ways for dealing with them, which can be combined in some cases. A first solution would be to exclude data outside an acceptable range that should be defined by researchers before the data collection, on the basis of theoretical criteria. For example, for a lexical decision task, several articles have reported the exclusion of times shorter than 200 ms and longer than 2,000 ms (e.g. Ferrand *et al*. (2010)). In this case, the criteria depend on the task or the type of process investigated and should be set for every new study.

A second solution, which can be implemented following the first, would be to eliminate or to replace extreme values on the basis of the distribution of the results. We can often read in the literature that for every participant and/or every item, data further than 2, 2.5 or 3 standard deviations from the mean have been eliminated or replaced by their threshold value. In the example we are discussing, the value of 2,412 deviates by more than 3 standard deviations from the mean. We could decide to eliminate this value or replace it with a less extreme value, like 2,002, which corresponds to the mean plus 3 standard deviations. Although we will not go into further detail in this book, it is important to note that the scientific community is not unanimous regarding the validity of this approach. For further developments, we refer those interested to McClelland (2014) who discusses these options, and to Leys *et al*. (2013) who present an alternative solution. In all cases, whatever the criteria chosen for the treatment of extreme values, these must be reported together with the results. In addition, it is customary to indicate the percentage of data that have been replaced or deleted.

The examples described above show that the mean is a good central tendency indicator when the data are distributed symmetrically. In other cases, it may be appropriate to summarize the data using another measure, such as the median. This simply corresponds to the value that separates the distribution into two equal parts. Half of the values are below the median, and the other half are above (the gray dashed line in Figure 7.1). Unlike the mean, the median is only slightly influenced by extreme values. If, for example, a third of the values in the distribution had values higher than 2,000 ms, the median would always remain the same. For this reason, this measure of central tendency is sometimes preferred to the mean when the distribution contains extreme data or when it follows an asymmetrical curve.

So far, we have considered all the correct YES decision times gathered in the experiment, regardless of the condition in which they were obtained. The fictitious study that we are examining, however, focuses on the influence of two variables, word length and word frequency, on decision times. It would therefore be more appropriate to observe the distribution of times in the different conditions resulting from the combination of these variables, as illustrated in Figure 7.4.

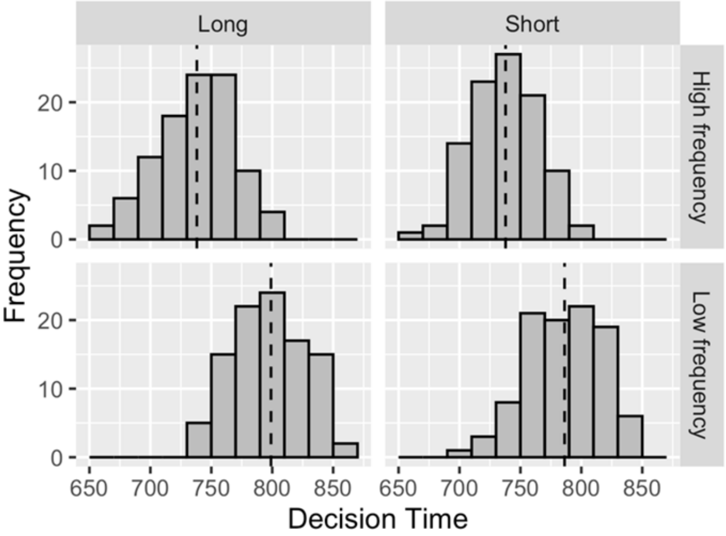


Figure 7.4. *Histograms representing the decision times obtained in the different conditions of the experiment, as well as the mean for each condition (dashed line)*

On these representations, we can see approximately normal distributions in the different conditions. We can also observe that, in general, the decision times for low frequency words (793 ms, bottom-line) are higher than those related to high frequency words (738 ms, top-line). The difference between short words (762 ms, right column) and long words (769 ms, left column) is very narrow. These results suggest that in our sample, it is only word frequency – not word length – that influences the lexical decision time.

7.6. Linear models

In order to represent data and to analyze the influence of independent variables on dependent variables, it is necessary to rely on models. A model makes it possible to predict the values of the dependent variable based on the values of the independent variable (also called *predictor*) . In other words, a model aims to mathematically represent the relationship between two or more variables.

In this chapter, we will present statistics that are based on linear models. This simply means that the predicted data are summarized by a straight line.

In order to define this line, it is necessary to determine its intercept, that is, the point where it crosses the *y* axis, as well as its slope, or the evolution that the line follows as the value of the predictor increases. For a model with only one predictor, the equation corresponds to:

$$y = (b_0 + b_1 * x) + \varepsilon \qquad [7.1]$$

In this equation, *y* represents the predicted value and *x* the value of the predictor; b_0 is the intercept and b_1 the slope of the straight line. The intercept corresponds to the value of *y* when *x* = 0, and the slope corresponds to the change in the value of *y* when the value of *x* changes by one.

A very simple model that we have already mentioned is the mean. In this model, only the intercept (the mean) is defined. If we predicted the decision times only on the basis of the mean, this prediction would be imprecise, because as we have already seen, the observed values never correspond exactly to the mean and deviate a certain distance. This difference between predicted and observed values is called error and is symbolized by ε in the equation.

In order to build a model that describes the data more accurately than the mean, we can add a regression coefficient b_1 associated with a relevant predictor. Adding this predictor aims at reducing the model's error. Figure 7.5 illustrates the participants' mean decision time in low and high frequency conditions (where every point represents a participant). On panel (a), the line represents a model based on the mean. We can quickly notice that this model does not help in the prediction of observed times; it reflects data in general, without taking into account the specifics related to the conditions. On panel (b), a linear model was calculated, in order to include the role of frequency on decision time. This model is as follows:

$$y = (738 + 54 * \text{Frequency}) + \varepsilon \qquad [7.2]$$

In a model including categorical predictors, it is necessary to assign a value to their different modalities. The choice made on panel (b) corresponds to what is called dummy coding, in which a value of 0 is assigned to the reference modality and a value of 1 is assigned to the second modality (the order of the modalities is determined alphabetically in *R*). For the reference modality (high frequency), the equation simply predicts the mean time of this condition (738+54*0). Thus, we can see that the intercept is equivalent to the mean of the reference category. If we apply the equation to the second modality of the variable (738+54*1), we then obtain the mean of the low frequency condition. When dummy coding is used, the regression coefficient represents the difference between the means in the two conditions.

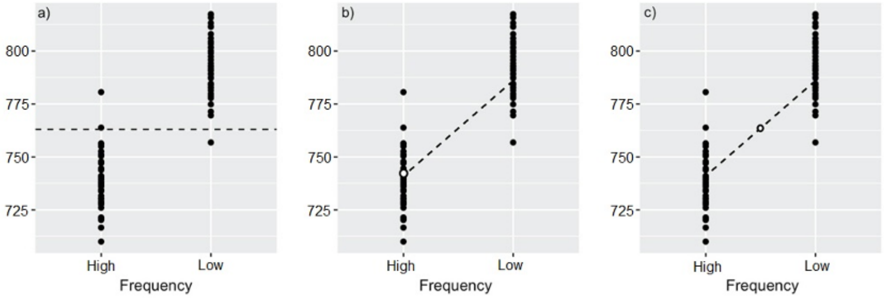


Figure 7.5. *Mean decision times of participants in high and low frequency conditions. On graph (a), the line represents a model based only on the mean. On graphs (b) and (c), the model takes into account the frequency predictor, using two different codings for the predictor ((b) = binary coding, (c) = sum coding). The intercept corresponds to the white circle*

Another possibility for coding categorical predictors is to use sum coding, which assigns opposite values to the modalities of the predictor, so that their sum is equivalent to 0.

In this type of coding, the value 1 corresponds to the reference category. In this specific case, value 1 is assigned to the high frequency condition and value -1 to the low frequency condition. This means that the intercept calculated in such a model no longer corresponds to the mean of the reference condition, but to the general mean, taking the times of the two conditions into account, as in equation [7.3], which corresponds to panel (c) in Figure 7.5.

$$y = 765 - 27 * \text{Frequency} + \varepsilon \qquad [7.3]$$

When sum coding is used, the regression coefficient indicates the difference between the general mean and the mean in each condition. In this precise case, it would be necessary to subtract 27 ms (-27*1) to the intercept to obtain the mean of the high frequency condition, and to add 27 ms (-27*(-1)) in order to obtain the mean of the low frequency condition.

Linear models can also be built on the basis of several predictors. To do this, we can simply add a regression coefficient associated with one or more additional predictors to the equation, as in the following example:

$$y = b_0 + b_1 * x + b_2 * x + \cdots + b_\mathrm{n} * x + \varepsilon \qquad [7.4]$$

If we included word length in the model, using sum coding, we would obtain:

$$y = 765 - 27 * \text{Frequency} + 3 * \text{Length} + \ \varepsilon \qquad [7.5]$$

When several predictors are included, it is possible to specify the terms of interactions between predictors. However, as we mentioned in the previous chapter, it is not recommended to include too many predictors in a model, especially when they involve interactions, because these models quickly become difficult to interpret.

The influence of the predictors included in the model is then tested using inferential statistics, whose logic we will describe in section 7.7.

7.7. Basic principles of inferential statistics

The primary goal of an experiment is to answer general questions and to reach conclusions that can be applied to a group of people or to a set of linguistic items. In the fictitious experiment we are discussing, the aim is to analyze the influence of word frequency and word length on the lexical decision time for people and words in general. However, in reality, it is impossible to test all the people or all the words. For this reason, data are collected from a sample of people, based on a sample of items. The parameters estimated at the sample level (means and standard deviations, for example) are then used to make deductions at the population level. There are different approaches for inferring conclusions for a whole population on the basis of a sample (Dienes 2008). The most common and widely used so far corresponds to the null hypothesis significance testing (NHST) framework.

7.7.1. *The null hypothesis significance testing*

In the NHST framework, the research hypothesis related to every independent variable of the study is expressed in the form of two opposite statistical hypotheses. The first one, the *null hypothesis* (H_0), states that the independent variable does not influence the dependent variable. In other words, the means observed in the different conditions should be the same. The second one, the *alternative hypothesis* (H_1), says that the independent variable influences the dependent variable. In other words, the means observed in the different conditions should differ. In order to illustrate these different hypotheses, let us go back to our example by imagining that we have examined the influence of a single variable, word frequency.

The statistical hypotheses in this situation would be as follows:

> H_0: the mean decision time for frequent words is equal to the mean decision time for less frequent words.
>
> H_1: the mean decision time for frequent words is different from the mean decision time for less frequent words.

It is important to note that these statistical hypotheses are not related to the sample from where the observed data were retrieved, but to the population for which one wants to draw conclusions.

The various statistical tests carried out within the NHST framework aim to check the compatibility of the data observed in an experiment with the data that would be predicted for the population by the null hypothesis. In our example, this logic would amount to saying "if word frequency did not influence decision time, what would be the probability of observing a difference of 54 ms between the conditions in our sample?". In order to answer this question, the statistical tests are based on the models we mentioned in the previous section. They are generally carried out on the basis of the difference in means between conditions, as well as on the basis of variations in data and sample size. Actually, a difference between means, as in our example, may be related to two possible sources. The first source corresponds to the systematic variation in the data that can be attributed to the manipulation, or in other words, to the fact that the word was frequent or not. The second source of variation corresponds to the unsystematic variation in the data, which can be attributed to the participants, such as their being faster/slower to respond, or to the items, such as being more or less familiar, for example.

In a simplified way, statistical tests in the NHST framework calculate the ratio between the systematic variation and the unsystematic variation explained by the model. If this ratio is greater than 1, this means that the systematic variation is greater than the unsystematic variation and that the independent variable has an effect. In order to know whether this effect is significant, the value returned by the statistical test is compared with a sampling distribution specifically related to the test performed. This makes it possible to determine the probability of observing an equal result or a more extreme result than the one observed in our sample if H_0 were true, which is called the *p*-value. If the *p*-value is small enough, then it is possible to reject the null hypothesis. In a relatively arbitrary manner, the majority of the scientific community has set a threshold below which the *p*-value would be acceptable, conventionally placed at 0.05 in Human Sciences. When p is smaller than 0.05, a result is said to be statistically significant.

The *p*-value is a conditional probability since it represents the probability of obtaining the observed data (or more extreme data) if H_0 were true. It is very important to remember this, so as not to draw the wrong conclusions from the *p*-value. In sum, the *p*-value gives no indication of the probability of H_0 being true. The null hypothesis is only a theoretical distribution and it is not possible to take a stand as to its veracity. Likewise, the *p*-value does not let us express an opinion as to the veracity of H_1. Indeed, the *p*-value only provides an indication as to the compatibility of the data observed with H_0, nothing else. Finally, the *p*-value does not directly reflect the size of an effect. For this reason, and in order to better interpret the statistical effects discovered, it is essential to couple the *p-value* with other measurements we will describe below.

7.7.2. *Effect sizes and confidence intervals (CIs)*

Imagine that, in our example, the difference observed between the frequency conditions is statistically significant, that is, that the *p-value* associated with the 54 ms difference is smaller than 0.05. The importance of this difference can be estimated on the basis of several points, as described by Winter (2019). First, it can correspond to the magnitude of the difference itself, here 54 ms. The greater this magnitude, the more the effect can be considered as important. Second, the importance of the difference depends on the variability observed in data: when the variability is small (reflected by a small standard deviation), we can be more confident about the existence of the difference than when the variability is large. These two points can be combined in order to calculate *effect size statistics*, which helps us to estimate how important a difference between conditions is. Examples of such measurements are Cohen's *d*, Pearson's *r*, the eta-square (η^2) or the omega-square (ω^2). The *p-value* and effect size are indicators which help us decide on complementary aspects of a result, the former evaluating the confidence we can place on the result, and the latter the importance or magnitude of such result.

Interestingly, the relationship between the *p-value* and effect size is not fixed. In fact, it is possible to demonstrate very small effects if the sample size is large enough. In this case, the effect could be statistically significant while being very small. Similarly, it is possible to demonstrate very important effects on the basis of a small sample.

The sample size also plays a role in the confidence we can place on an effect: large samples make it possible to reach more reliable conclusions than small samples. As explained above, the measurements made on the sample aim to infer or to estimate the population parameters. In order to assess the accuracy of this estimate, we can calculate the *standard error* on the basis of the variability in the data and sample size. Based on the standard error, it is possible to calculate confidence intervals at 95% around the estimated value, in which the *real* value is expected to fall 95% of the time. For more information on this topic, see Cumming (2014).

7.7.3. *Potential errors and statistical power*

Now, let us go back to the logic behind the null hypothesis testing. As described above, in this type of approach, the result of the statistical test allows us to evaluate the probability for the data obtained (or more extreme data) to correspond to the one predicted by H_0, if the latter were true. H_0 can be rejected when this probability is smaller than a certain threshold. However, the fact that the *p*-value is below the threshold does not necessarily mean that it would be impossible to be wrong by rejecting H_0. Likewise, the fact that the *p*-value is greater than the threshold does not necessarily mean that the effect does not exist. There are thus two types of errors that may come up in the NHST framework. The Type I error corresponds to the probability of rejecting the null hypothesis – and concluding the effect is present – when the null hypothesis is true and the effect does not exist at the population level. The Type II error corresponds to not rejecting the null hypothesis – thus not concluding that an effect exists – while it actually exists in the population.

The statistical power corresponds to the probability of obtaining a significant effect when it is actually present in the population. This is represented by the Type II error formula. In general, the recommended power is 0.8, in order to minimize the risk of Type II error, which would always be 0.2 in this case (20%, indicating that one study in five could potentially find no effect, while this is nonetheless real). As the result of a test, statistical power is dependent on the effect size, the sample size, as well as the variability in the latter. This means that when studying large effects, it is possible to carry out research by testing limited samples, and/or samples whose variability is not kept to a minimum. On the other hand, when we aim to investigate a phenomenon with a small effect size, it is necessary to collect a lot of data and/or to try to minimize the variability within the sample.

Statistical power is thus an essential element to take into account both when developing an experiment and when interpreting the results. Even if a power of 0.8 is generally recommended, it is rare to observe such a power in the studies carried out until now in the field of linguistics, or in related fields such as psycholinguistics (for a review of the problems raised by an inadequate statistical power, see Ioannidis (2005)).

In order to carry out research with adequate statistical power, it is necessary to estimate it in advance. Power calculation can be done based on an estimate of the expected effect size to determine the adequate sample size. To do this, we can turn to the results of previous studies or to expected effect size estimates. For simple models, there is a tool freely accessible online, *G*Power* (Faul *et al.* 2007)[2], which makes it possible to estimate different parameters related to statistical power, on the basis of known parameters. For more sophisticated models, such as the mixed linear models which we

[2] See http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html. See also https://stats.idre.ucla.edu/other/gpower/ for examples of applications.

will present later, power calculation is more complicated. For more information, we recommend Brysbaert and Stevens (2018), a reference already mentioned in the previous chapter when we addressed the question of the number of items and participants to be included in an experiment[3].

7.8. Types of statistical effects

Different effects can be assessed during statistical modeling. When only one independent variable is investigated, the variable is introduced into the model and the result of the test makes it possible to decide on its effect. For example, by introducing the frequency variable into the model, the result would allow us to know whether the frequency effect observed in our data is significant. When several variables are investigated at the same time, as is the case in our example, two types of effects can appear.

The first type of effect, the main effect, corresponds to the effect of a variable independent of the others. In an experiment with two variables, two main effects can appear, one for each of them. Compared to our example experiment, the first main effect, the frequency effect, describes the situation in which decision times are higher in one frequency modality than in the other. Based on the literature, such an effect should be predicted. In fact, less frequent words should be recognized more slowly than more frequent words. The second main effect, related to word length, would correspond to the fact that decision times are higher in one modality of the length variable than in the other. Intuitively, we could assume that longer words take longer to recognize than shorter words.

The second type of effect, the interaction effect, indicates that the effect of one variable depends on the modality of another variable. In practice, an interaction effect may result from different configurations, as illustrated in Figure 7.6. First, it is possible for the effect of the first variable to be present

[3] An alternative approach would be to turn to the Bayes factor. Unlike the null hypothesis significance testing, in which the data are evaluated in relation to their adequacy with the null hypothesis, the Bayes factor makes it possible to evaluate the adequacy of data with the alternative hypothesis. It is thus possible to determine which hypothesis, either the null hypothesis or the alternative hypothesis, is supported by the results obtained in an experiment. For more information, see Dienes and Mclatchie (2018).

in a single modality of the second variable. In our example, this would be equivalent to observing a frequency effect for shorter words but not for longer words, for example (panel (a)). Second, the effect of the first variable may be greater in one modality of the second variable than in the other. It could appear that frequency influences decision time in general, but that this influence is more important for shorter words than for longer words (panel (b)). Finally, the effect of the first variable could be opposite depending on the modality of the other variable. In this case, decision times would be faster for high frequency words than for low frequency words when these are short, whereas when they are long, decision times would be faster for low frequency words than for high frequency words (panel (c)). This last possibility is unlikely for our example, but helps us make our point.

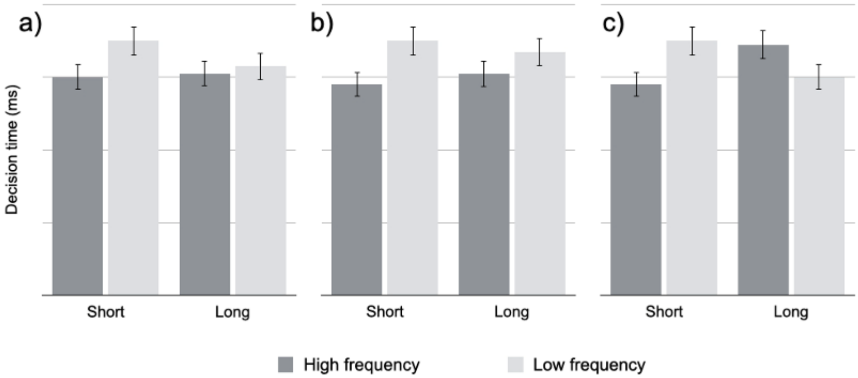


Figure 7.6. *Illustrations of interaction effects for a design with two independent variables*

When more variables are introduced into the model, the number of main effects, and especially the number of interaction effects, increases. Higher-order interactions also appear, which require more interpretation stages. Thus, in an experimental design comprising three independent variables, seven effects are possible. These effects correspond to the main effects, of which there are three in total as with the number of independent variables. Apart from the main effects, we have the interaction effects between two variables, of which there are also three (VI1*VI2, VI1*VI3 and VI2*VI3), as well as the interaction effect between the three variables (VI1*VI2*VI3). Decomposing such higher-order effects requires a systematic approach which we do not develop here, but which is presented in detail by Field *et al.* (2012).

7.9. Conventional procedures for testing the effects of independent variables

For a long time, the effects of the independent variables were evaluated using various tests based on the above-mentioned linear models. Choosing the test to be performed is based on different criteria, such as the number of independent variables involved, the number of modalities of these independent variables, the type of independent and dependent variables, and the type of experimental design (independent groups or repeated measures). Classic procedures to test the effect of one or more categorical variables on a quantitative dependent variable are called *Student test* and *ANOVA* (for analysis of variance). These tests are based on linear models and require compliance with the following assumptions of parametric data[4]:

– homogeneity of variance, meaning that the data are distributed similarly around the means in the different groups;

– normally distributed data;

– independence of observations, meaning that every person only contributed once to the data and that data from different participants are independent.

If you think about the example we are developing in this chapter, it becomes clear that the last assumption is violated, since participants have seen all the words in the experiment. The data collected for the same person or for the same item are therefore correlated and interdependent. For example, if a participant responds slowly in general, her responses will tend to be slower in all conditions. Therefore, it might be possible to predict her response times on the basis of previous response times. Likewise, if an item is more complex than the others, it is generally likely to be processed more slowly by participants. One response time for this item could also be partially predicted based on the other times associated with it.

In order to take into account the interdependence of the data in repeated-measures designs, for each test, there is a version adapted to this type of

[4] These assumptions must be checked systematically before applying a parametric test. When these assumptions are violated, it is necessary to turn to other types of tests, such as nonparametric tests.

design, in which some parameters are modified. In the repeated-measures tests, it is possible for every participant to contribute to all conditions. It is nonetheless necessary to reduce the initial database in order to enter only one data point per participant per condition. This is done through an aggregation process, whereby the data obtained by a person are summarized, in the majority of cases using the mean.

Data aggregation, however, poses different problems in terms of data modeling. First, using only one indicator for summarizing a data set entails loss of information. If we reconsider how data are distributed around the mean, we will reckon that this indicator does not let us take the variability of measurements into account.

Secondly, aggregation only lets one source of variation be considered. In fact, when we summarize data per participant, the effect of the variables is tested against the participants' means and the variation existing between the items is lost.

In order to fully understand what this loss of information represents, we will use an example similar to the one provided by Brysbaert (2007), focusing only on the experiment's variable "frequency". Imagine that the participants' mean times per condition, calculated on the basis of five items per condition, are as presented in Table 7.3.

In order to examine the influence of frequency on decision times, we should carry out a *Student test* with repeated measurements. This would return the following result: $t(9) = 5.43$, $p<0.001$, which we will break down before going any further. The *t* indicates that a *Student test* has been used. The number between brackets after the *t*, 9, corresponds to the analysis number of degrees of freedom[5]. In a simplified manner, the reported value for *t* represents the relationship between the systematic variation and the unsystematic variation of the model (see section 7.7.1).

[5] Simply put, the degrees of freedom can be considered as the number of parameters which may vary while keeping the mean even. If we had two numbers, 18 and 24, and we wanted to change these numbers but get the same mean, only one number could vary freely. The second one would necessarily depend on the first, since it would be conditioned by the mean we intend to obtain. If we changed 18 to 10, 24 would necessarily have to be changed to 32 in order to obtain a mean of 21. The number of degrees of freedom for a Student test is equivalent to the number of participants minus 1.

Participant	High frequency	Low frequency
1	634	680
2	657	701
3	711	712
4	623	658
5	655	745
6	599	669
7	632	673
8	678	704
9	659	697
10	661	688
Mean	**650.9**	**692.2**

Table 7.3. *Examples of participants' mean decision times in the two frequency conditions*

On the basis of this value and the number of degrees of freedom, it is possible to calculate the *p-value*, which corresponds to the probability that such a ratio could be obtained if the null hypothesis were true, for example, if people in general did not show different reaction times for high and low frequency words[6].

On the basis of the analysis carried out on the data aggregated per participant, the conclusion would be that the frequency effect is statistically significant. However, as we mentioned previously, the means obtained by the participants do not let us observe the variation between the items. In order to do this, we would also have to aggregate response times per item, as in Table 7.4.

[6] The result of a repeated-measure ANOVA on the same data would appear as follows: $F(1, 9) = 29.47$, $p<0.001$, where the value of F represents the relationship between the systematic and unsystematic variations of the model in a simplified manner. Contrary to t, F is defined by two degrees of freedom. The first one corresponds to the variables' degrees of freedom (number of conditions -1) and the second one corresponds to the error's degrees of freedom, (number of conditions -1)*(number of participants -1).

The means of the two conditions are similar to those obtained during the aggregation per participant. In this case, however, the effect would not be significant, $t(4.22)$[7] $= 1.23$, $p = 0.28$. The frequency effect observed for the sample of words used in the experiment could therefore not be generalized to all of the high and low frequency words. This is due to the fact that one value (824 ms for item 8) in the low condition is much higher than the others, and alone explains the difference in mean between the frequency conditions.

Item	Frequency	Time	Mean
1	High	657	650.8
2	High	644	
3	High	651	
4	High	635	
5	High	667	
6	Low	660	692.2
7	Low	658	
8	Low	824	
9	Low	649	
10	Low	670	

Table 7.4. *Examples of decision time means for items, depending on their frequency*

This example shows that when we analyze data from experiments testing not only a sample of participants, but also a sample of items, it is necessary to take into account both the variations between participants and between items. For this reason, in classical analyses, it is customary to perform two analyses, one based on the aggregation per participant, and the other based on the aggregation per item.

[7] In this case, Welch correction was applied in order to take into account the fact that the variances are not equal in the two conditions. Of course, performing a statistical test on such a small sample is not optimal but helps us illustrate the points presented.

In order to conclude that the independent variable had an effect, both types of analyses have to return significant results[8]. This process aims to generalize the results both to the populations of participants and of items ((Clark 1973); for an explanation of the procedure to be followed, see Brysbaert (2007)). However, this type of analysis has raised some concerns, particularly regarding the increased risk of a Type I error in models that do not fully account for data dependence (Judd *et al*. 2012; Barr *et al*. 2013).

Despite their limitations, it is this type of analyses that you will generally come across in the literature published before 2010, since the technical means available before that time made it difficult to calculate more complex models, taking into account data interdependence. After Baayen *et al*. (2008) presented a solution, the linguistic research community has increasingly turned to these new analyses, which we introduce in the following section.

7.10. Mixed linear models

In order to build an accurate model, it is necessary to include as much information as possible, namely concerning the variation between the measurements for the same item or the same person (which are lost when using models based on aggregate information, such as the mean). Mixed linear models make this possible.

7.10.1. *Fixed and random effects*

In a mixed model, two types of effects are taken into account and evaluated: fixed effects and random effects. These notions are at the center of these models; that is why it is important to understand what they mean. This will allow you to define the different effects in a relevant way so as to model the data you want to analyze.

[8] It is interesting to note that experts recommended calculating an additional indicator, minF’, based on the results of each analysis, in order to simultaneously generalize results to the populations of participants and of items.

The first type of effect, fixed effects, simply correspond to the effects of the variables manipulated in the experiment. More precisely, they represent the general effect of a variable, independently of the unsystematic variability present in data (Singmann and Kellen 2020). Following this definition, fixed effects are assumed to be constant from one experiment to another. In our example, frequency and length are the fixed effects that we investigate[9].

The second type of effect, random effects, are those related to the unsystematic variations in data, which cannot be explained by fixed effects, and which may come from different sources. The random effects typically considered in an experiment are participants or items. Indeed, as we have already mentioned many times, the data obtained in a linguistic experiment come from a sample of participants, whose own characteristics may influence responses. For example, we might easily imagine that some participants react more quickly than others in general, which could lead to measurements partly depending on this general speed. Likewise, data are collected on the basis of a sample of items, whose specific characteristics may influence responses. For example, some may be processed more quickly than others, regardless of their frequency or their length but in relation to their phonological properties, and this could eventually influence the results. It can be useful to imagine the random effects as corresponding to the groups beyond which we want to generalize the data (Singmann and Kellen 2020).

In order to better understand what random effects correspond to in statistical models, let us observe Figure 7.7 displaying the decision time distributions for participants (panel (a)) and for items (panel (b)) included in our fictional experiment.

[9] It may be appropriate to introduce additional fixed effects, which are not directly related to the independent variables investigated, but which are relevant for the prediction of the dependent variable. For example, it is known that reading or decision times decrease as the experiment progresses. The trial number is a fixed effect which it might be useful to include in the model in order to acknowledge this fact.

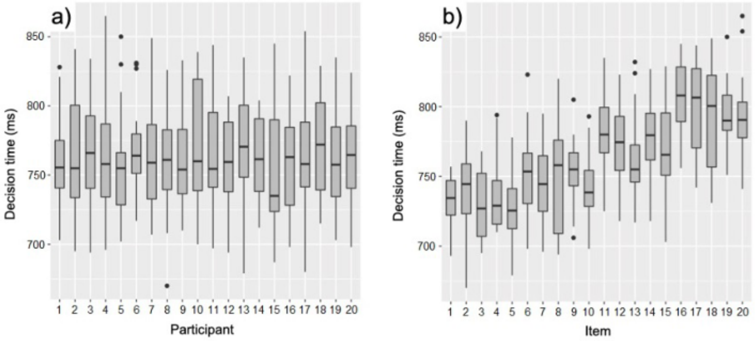


Figure 7.7. *Reaction times distribution for participants (panel (a)) and for items (panel (b)). Items 1–10 are high frequency and items 11–20 are low frequency items*

In general, for participants, we can see that the medians (indicated by the black horizontal lines) are placed at different levels, and that the data are distributed around the medians in a different way between participants. The same type of observation can be made for items. This illustrates the variability that participants and items provide to data. In order to better understand how the variability between participants or items may influence a model, let us now turn to Figure 7.8, illustrating the data for three participants in the experiment.

The graph shown in this figure should remind you of the one we already saw when we discussed the linear model. Here, we can see what the best model for representing data for every participant might look like. What we can first notice is that each model has a specific intercept. In a simplified manner, this represents the fact that every participant presents a particular mean decision time in the reference condition, partly related to their personal characteristics. If we turn to the lines connecting the two conditions, we can see that their slopes differ for every participant. This means that the participants' decision times are not influenced in the same way by the independent variable, word frequency. In order to model these data accurately, it is necessary to resort to a model enabling a variation of intercepts and slopes among participants. According to the representation of items in Figure 7.7, it would also be appropriate to include the random effects related to the items in the model.

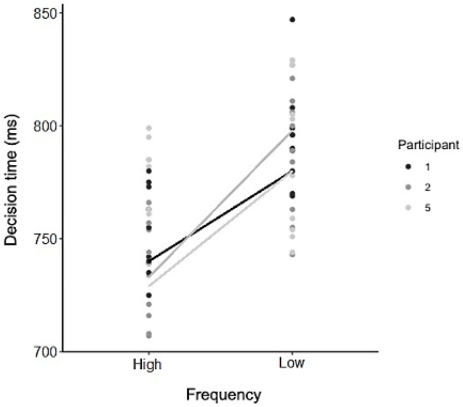


Figure 7.8. *Decision times from participants 1, 2 and 5. Each point corresponds to a decision time (10 per condition). The lines represent the best model for each participant*

7.10.2. *Building mixed models*

As we have seen so far, a mixed model enables data modeling on the basis of fixed effects and random effects. In this section, we will present the different options for such effects. In the case we are interested in, first we will try to observe a frequency-related fixed effect, where highly frequent words are recognized more quickly than less frequent words.

Now, let us consider the random effects that may influence this model. The experimental design used in this experiment is based on repeated measures for the participants who react to all the conditions of the experiment, that is, low frequency words and high frequency words. We have already discussed the fact that it is likely that participants execute the task more or less quickly, which implies that their intercepts should be able to vary in the model. Similarly, it is probable that word frequency does not affect participants in the same way. In order to reflect this, the slope of participants should also be able to vary. Now, let us turn to the items. Unlike participants, every item can only appear in one condition of the experiment because every word can only correspond to a high or to a low frequency. This means that there is no slope associated with each item, since items are not repeated between conditions. Nonetheless, it is likely that the specific characteristics of each item will influence the mean decision time associated with it. This means that in this experiment, only the item intercept should be able to vary.

7.10.3. *Testing a mixed model using R*

In this section, we will describe the general steps for testing a mixed model using *R*, as well as the results returned by the test. We will present the different *packages* used for this approach, as well as the possibilities or limitations they have.

There are different functions for testing mixed models in *R*. The best known and most widely used is the `lmer ()` function, available in the `lme4` package (Bates *et al*. 2015)[10]. As with most functions in *R*, they require long format data (see section 7.2), containing the information needed for building the model. These may include the identity of the participants, the identity of the items, decision time, or frequency and length conditions.

In order to build the mixed model we are interested in with `lme4`, we have to write the following formula:

```
m_freq<- lmer(Time ~ Frequency +
(1+Frequency|Participant) + (1|Item), data=D, REML = FALSE[11])
```

Let us break down the formula following the term `lmer` in order to understand what it contains. The first terms `(Time ~ Frequency)` simply mean that we want to predict time based on the frequency category. These are the same terms we would enter into a classic linear model. The following terms represent the variations in the data we wish to explain using random effects. The term `(1|Item)` corresponds to the variations in the items' intercepts (a different intercept is estimated for every item). The term

[10] In this chapter, we will only discuss the modeling of continuous quantitative data. When analyzing category-specific data (yes/no answers, for example), it is possible to turn to the glmer () function, which is also available in lme4.

[11] The term REML = FALSE indicates that the model is based on maximum likelihood. This is important so that you can later determine the influence of the fixed effects included in the models.

`(1+Frequency|Participant)`[12] means that not might only participants' intercepts vary (*1*), but also their slope (the term + *frequency*).

By default, the contrast used in *R*, and therefore in `lme4`, follows a dummy coding, in which a condition is defined as a reference category. When models include interactions, it may become difficult to interpret results based on this type of contrast. In order to remedy this problem, it is possible to set up a sum coding, as shown in the following example, which also presents an extract of the results obtained on the basis of a mixed model:

```
contrasts(D$Frequency)<-contr.sum(2)
m_freq<-lmer(Time~Frequency+(Frequency|Participant)+
(1|Item), data=D, REML=FALSE)
summary(m_freq)

## Random effects:
##  Groups      Name        Variance Std.Dev. Corr
##  Participant (Intercept)  28,44    5,333
##              Frequency1   10,73    3,275    0,85
##  Item        (Intercept)  18,43    4,293
##  Residual                859,38   29,315
## Number of obs: 400, groups: Participant, 20; Item, 20
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  765,295      2,119   361,1
## Frequency1   -27,305      1,899    14,4
```

Based on the formula, the model intercept and the regression coefficient for the frequency predictor are estimated. In addition, an intercept for every item, as well as an intercept and a regression coefficient for every participant are calculated. First, let us look at the fixed effects, in the lower part of the results. The intercept equals 765. This corresponds to the decision time general mean. The fixed frequency effect, evaluated at -27, corresponds to the difference between the general mean and the reference category of the frequency predictor, in other words, the high frequency condition. This part of the results is similar to the one we would obtain on the basis of a linear model.

[12] This term is equivalent to (Frequency|Participant), where the intercept is automatically included.

Let us now turn to the upper part of the results, which concerns the random effects introduced in the model. You can see the *standard deviation* (*Std.Dev*) calculated for every random parameter in the model. For example, the standard deviation concerning the participants' random intercept is approximately 5 ms and represents the participants' variation around the general mean. Correlation (*Corr*) between the intercept and participants' slope random effects is also shown. This represents the relationship between the intercept and the slope, which in this case is large and negative. This means that the more the intercept increases, the more the slope decreases. In our case, this would mean that the slower the participants respond, the smaller the difference between conditions.

To determine whether the effect of a predictor is statistically significant, it is necessary to compare the model containing the predictor we want to test with a completely similar model to the previous one, apart from the absence of the predictor to be tested. In order to test the model above, it would be necessary to build a model based on the same random structure but not containing the predictor, that is, a model exclusively based on the intercept (1), as here:

```
m_int <- lmer(Time ~ 1 + (Frequency|Participant) +
(1|Item), data = D, REML = FALSE)
```

Models can be tested using the `anova` function, which is also available in `lme4`. If the difference between the models is significant, that is, if the model containing the predictor explains the data sufficiently better than the reduced model, then we can consider that the predictor in question plays a significant role in the model:

```
anova(model_frequency,modele_intercept)
```

```
## Data: D
## Models:
## m_int: Time ~ 1 + (Frequency | Participant) + (1 | Item)
## m_freq: Time ~ Frequency + (Frequency | Participant) + (1 | Item)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m_int   6 3921.1 3945.1 -1954.5   3909.1
## m_freq  7 3872.0 3900.0 -1929.0   3858.0 51.065      1  8.935e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova function performs a likelihood ratio test between the two models. In other words, it compares the goodness of fit of the models. This comparison is made on one of the likelihood indicators of the model (*logLik*, log-likelihood), and tests the null hypothesis that log-likelihood is not different between the models. The test used is a chi-square test (*Chisq*) whose degree of freedom (*df*) corresponds to the difference in the number of parameters of the models, equal to 1 in this case. We can see that the *p-value* associated with this test is significant ($p<0.001$), which means that it is extremely unlikely that both models have similar goodness of fit. On this basis, we can conclude that the frequency predictor has an effect, and we can report it as follows: "A likelihood ratio test on the models, including the frequency predictor (or not), showed a significant difference between the models, $\chi^2(1) = 52.03$, $p<0.001$."

Models can be compared for fixed effects as well as for random effects, provided that certain specifications of the model are respected (for more detail concerning these specifications, see Winter (2019)). Different methods can be used for making these comparisons, which present advantages as well as limitations, in particular in terms of risks related to the Type I error when there are not enough random factor levels (Singmann and Kellen 2020). For the comparison of fixed effects, the Kenward–Roger approximation (Kenward and Roger 1997; Halekoh and Højsgaard 2014) and the Satterthwaite approximation (Satterthwaite 1946; Kuznetsova *et al.* 2016) are generally recommended.

We can easily imagine that when the models include more than one predictor and potential interactions between predictors, the test of the influence of each predictor can quickly lead to the construction and to the subsequent comparison of a lot of models. In order to simplify this procedure, it is possible to turn to the afex *package* (Singmann *et al.* 2017). The mixed function, contained in this *package*, is based on the lmer function and enables circumventing certain difficulties inherent in the latter. In particular, the mixed function returns the p-values associated with each fixed effect in the results by default. If interactions are present in the model, it is possible to inspect them using the functions contained in the emmeans package (Lenth 2017).

7.10.4. *Which random structure to choose?*

The main difficulty of mixed models is the definition of the model's random structure, especially when the models involve several predictors. Since mixed models appeared, different practices have been followed. These practices may be classified according to a continuum of complexity, ranging from structures consisting only of random intercepts for participants and items to maximal random effect structures, encompassing all random effects justified by the experimental design, as recommended by Barr *et al.* (2013). This maximal model contains the intercepts and the random slopes relative to all the fixed effects which may vary depending on the levels of a grouping factor. For example, if we wanted to include frequency and length predictors in the model and also assumed the existence of an interaction between frequency and length, the maximum model fitted using the `mixed` function would correspond to this:

```
model_maximal <- mixed(Time ~ Frequency*Length +
(Frequency*Length|Participant) + (1|Item), data = D)
```

It has been shown that using a less complete random structure than needed, that is, ignoring a random effect which should be taken into account, increases the probability of Type I error (Judd *et al.* 2012; Barr *et al.* 2013). According to this, it would be more appropriate to build a maximal model in order to limit this risk. However, fitting a maximal model is often problematic when sample sizes are limited, as is often the case in a linguistics experiment. Given the number of parameters to be estimated in this type of model, as well as the complexity of the underlying structure, it can happen that it is impossible to calculate it due to an insufficient amount of data.

In these cases, different solutions may be followed, both in terms of model specifications and in terms of the means used for fitting it. We will not go into further detail concerning this topic here, but the reader may refer to Winter (2019) and Singmann and Kellen (2020) for a review of the existing possibilities. In addition, it has been shown that systematically using the maximal model could lead to a loss of statistical power, and therefore an increase in Type II error, that is, the risk of not demonstrating an effect when it exists (Bates *et al.* 2015; Matuschek *et al.* 2017).

There is thus no definitive procedure to follow regarding the specification of the random structure of a mixed model. This choice depends on the predictors of the model, the groups that can give rise to random effects, the potential influences that can be expected on the basis of the theory, as well as on the data on which the model is built.

In order to be able to make the decisions needed for the development of mixed models, it is therefore essential to understand how these models are built and their implications. All in all, the elements reviewed in this chapter only represent a part of what it is required to understand before embarking on such analyses. It is therefore up to you now to take the time to deepen and to assimilate these concepts, turning to the specialized literature recommended at the end of the chapter.

7.11. Best-practices for collecting and modeling data

Before concluding this introductory chapter to the analysis of quantitative data, it is important to discuss an essential aspect of this approach: subjectivity. When talking about statistics, we can often hear that it is possible to make figures say everything and their opposite. Although this negative conception of statistics indicates a lack of knowledge of the basic principles underlying their use, it is sometimes relevant. It is indeed quite possible to make the figures speak so as to favor certain results or to conceal others. This can be the result of deliberate choices, for example, by selecting and analyzing only certain data. We would then be in the presence of a flagrant case of bad practice, which is unacceptable in scientific research. However, there is a whole range of practices, often unconscious, which can lead to a misinterpretation of the data. These practices stem from different processes related to the cognitive biases of people involved in research and a limited comprehension of the statistics implemented.

As we have seen throughout this chapter, researchers must make many decisions during the design of the experiment and during the analysis phase, decisions which may have an impact on the quality of statistical inference. Let us first look at the choice of the sample size to be studied. We have seen that statistical power is related to different parameters that must be taken into account when determining the number of people to test and the number of items to use. This step is very important in order to implement an experiment with enough power to enable researchers to make a decision concerning the effect. It is also essential in order to avoid entering into problematic practices, such as testing a certain number of people, testing the desired effect, finding that it is not present and continuing to collect data until the effect appears. By doing this, the probability of having a Type I error would rapidly increase, as every test performed would depend on the result of the previous one (Simmons *et al.* 2011). It has also been shown that increasing the number of tests increases the risk of Type I error, the latter even rising to 10% for two tests, and reaching more than 60% for 20 tests (Winter 2019).

However, it is sometimes difficult to precisely assess the size of the desired statistical effects needed for the calculation of power *a priori*. In this case, it is possible to carry out sequential tests, at different stages of the data collection, while taking into account the increased risk of Type I error. For more information on this procedure, the interested reader may refer to Lakens (2014).

When analyzing data, different decisions must be made such as whether or not to include extreme values in the model, eliminating participants or choosing the specifications of the statistical models. These decisions may have a significant impact on the results obtained (Simmons *et al.* 2011; Gelman and Loken 2014). Let us repeat it once more, it would be incorrect to test different models or different ways of processing data only choosing those that will best meet the expectations of researchers. By doing this, the risk of seeing a significant result appear when the null hypothesis is true is high, as the number of tests performed increases.

In order to avoid introducing biases in data analysis, it is necessary to consider the statistical analyses from the beginning of the construction of the experimental design, and to clearly specify beforehand:

1) the number of participants and items to be included in the experiment;

2) the stages involved in data processing;

3) the statistical analyses for each hypothesis as accurately as possible.

A document containing this information can then be used to pre-register the study on a platform such as *The Open Science Framework*[13] (for a brief description, see Foster and Deardorff (2017)).

[13] http://osf.io.

7.12. Conclusion

In this chapter, we introduced the necessary theoretical bases for analyzing quantitative data drawn from continuous variables. We first described the formats of databases in linguistic experiments, and then presented some necessary steps for processing raw data before proceeding with analyses.

We then saw that the data follow a distribution and can be summarized by different parameters, such as the mean or the standard deviation. Then, we approached data modeling using the general linear model. We have seen that it was possible to mathematically describe the relationship between a dependent variable and one or more independent variables (also called predictors), among other things, to test the influence of the independent variables on the dependent variable. This relationship can be tested through inferential statistics techniques based on null hypothesis testing.

In this approach, two hypotheses, the null hypothesis and the alternative hypothesis, must be specified. The data collected in the experiment are evaluated in relation to their probability of occurrence in case the null hypothesis were true. When this probability is smaller than 5%, it is possible to reject the null hypothesis, and to consider that the results obtained are not caused by chance.

We have also seen that the classic tests (Student test and ANOVA) require compliance with certain assumptions which can be problematic in experiments where large amounts of data are collected for every participant and every item.

Mixed linear models make it possible to model this type of data, by allowing us to specify not only the fixed effects, but also the random effects, related as much to participants as to the items in the experiment. We provided some examples of these models and the general procedure for testing them in *R*. Finally, we discussed the best practices to be followed when collecting and analyzing data.

7.13. Revision questions and answer key

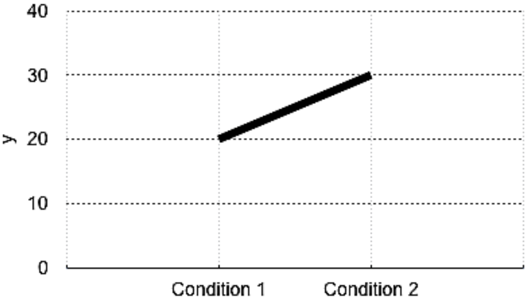
7.13.1. *Questions*

1) In the following table, which box corresponds to each of the following concepts: Type I error, Type II error, power, correct decision?

	Reality in the population	
Sample-related decision	**H_0 is true**	**H_0 is false**
Do not reject H_0 ($p>0.05$)		
Reject H_0 ($p<0.05$)		

2) Imagine that the distribution of results (response times) obtained in an experiment follows a normal law, with a mean of 632 ms and a standard deviation of 133. Between which values would 68% of the data fall? And 95% of the data?

3) Determine the linear model that characterizes the following relationship, first using dummy coding and then sum coding.



4) Meier and Robinson (2004) examined the association between the position of a word on the vertical axis and its affective evaluation. Their study was based on the existence of a conceptual metaphor *up is good*, which may influence concept representation. According to this metaphor, the objects placed in a higher position are generally positive, whereas those placed at the bottom are generally negative. For example, we can think of paradise and hell, the position of the results in a ranking or the fact of placing one's thumb up or down. In order to assess the link between affective assessment and spatial position, the authors chose words with a

positive (such as *hero*) or a negative valence (such as *liar*), and presented them either at the top or at the bottom of a computer screen. Participants had to assess whether the words were positive or negative by pressing a key for *positive* and another key for *negative*. Decision times were recorded.

a) What are the two independent variables of this experiment?

b) What are the possible effects based on these two variables?

c) Which of these effects would reflect a relationship between vertical position and affect?

5) Let us go back to the example of the fictitious experiment in Chapter 6, which aimed to study the influence of presentation modality (written vs. spoken) of a sentence on its comprehension. Imagine an experimental design with repeated measures, meaning that participants saw the sentences in all the conditions and that the items were also presented in all the conditions. In order to counterbalance the conditions, half of the participants started with the spoken modality and the other half with the written modality. In order to analyze the results, a mixed linear model must be fitted.

a) What are the fixed effects to introduce in the model?

b) What are the random effects?

c) Which maximal model should be built?

d) Which should be the reduced model for comparing the maximal model, in order to decide on the influence of the sentence's presentation modality?

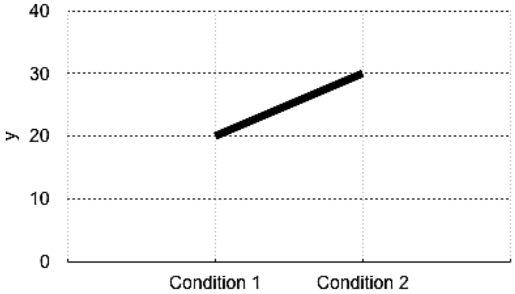
7.13.2. *Answer key*

1)

	Reality in the population	
Sample-related decision	***H_0* is true**	***H_0* is false**
Do not reject H_0 **(p>0.05)**	Correct decision	Type II error
Reject H_0 **(p<0.05)**	Type I error	Correct decision and power

2) To answer this question, we should focus on the properties of a normal distribution, for which 68% of the data are located at one standard deviation from the mean, and 95% of the data at two standard deviations from the mean. The interval located at one standard deviation from the mean corresponds to the values between (632 - 1*133) and (632 + 1*133), that is, between 499 ms and 765 ms. The interval located at two standard deviations from the mean corresponds to the values between (632 - 2*133) and (632 + 2*133), that is, between 366 ms and 898 ms.

3)



When using dummy coding, the intercept corresponds to the mean of the reference condition (here, condition 1) and the slope corresponds to the difference between the reference condition and the second condition. The equation will therefore be:

$$y = 20 + 10 * x + \varepsilon$$

When using a sum coding, the intercept corresponds to the general mean (the mean of the two conditions), and the slope to the difference between the intercept and the reference condition (here condition 1, coded 1). The equation would therefore be:

$$y = 25 - 5 * x + \varepsilon$$

4) a) The first independent variable corresponds to the valence of the word presented (positive vs. negative). The second independent variable corresponds to the location of the word on the screen (on top vs. at the bottom).

b) In an experiment with two independent variables, three effects may appear, namely a main effect for each variable and an interaction effect between the variables. The main effect of the word's valence might correspond to the fact that the words in one valence condition are generally evaluated more quickly than the words in the other condition. For example, we might assume that positive words are evaluated faster than negative words. The main effect concerning the location in the screen would correspond to the fact that the words appearing in one location are generally evaluated more quickly than the words in the other location. Finally, the interaction effect between the valence and location might correspond to the fact that the effect of the "valence" variable might depend on the modality of the "location" variable.

c) The effect underlying the relationship between emotional valence and location is the interaction effect. As a matter of fact, if there is an actual relationship between these two variables, then we might expect the positive words to be evaluated more quickly than the negative ones when shown at the top of the screen, whereas the negative words should be evaluated more quickly than positive ones when presented below.

5) a) The fixed effects correspond to the effects of the variables we want to study and which are manipulated in the experiment. In this case, the fixed effect corresponding to the variable is the presentation modality of the sentence (written vs. spoken modality). We could also introduce as a fixed effect the order in which the participants saw the conditions and the trial number.

b) Random effects correspond to those effects related to unsystematic variations in data, which cannot be explained by the fixed effects. In this case, as every participant probably has their own characteristics influencing comprehension, and as it is likely that the manipulation does not affect all the participants in a similar way, it would be appropriate to introduce in the model a random intercept and a random slope for participants, in the form (1 + Presentation modality | Participant). This is probably also the case for items, since these have their own characteristics and are tested in the two presentation conditions. It would therefore also be necessary to introduce a random intercept and a random slope for the items in the form (1 + Presentation mode | Item).

c) Based on the fixed and random effects identified above, the maximal model would be as follows:

```
Comprehension ~ ConditionOrder + TrialNumber + PresentationModality +
(1+ PresentationModality|Participant) +
(1+ PresentationModality|Item)
```

d) In order to build the reduced model to assess the influence of the sentence’s presentation modality, it would suffice to remove this term from the model, while preserving the other fixed effects and the same random structure. The model would then be as follows:

```
Comprehension ~ ConditionOrder + TrialNumber + (1 +
PresentationModality|Participant) + (1 + PresentationModality|Item)
```

7.14. Further reading

Field *et al*. (2012) is an excellent introductory manual to statistics, and the use of *R*. Winter (2019) provides a thorough introduction to understanding and modeling data, the statistical methods applied to data collected in linguistics experiments and the procedure for carrying them out using the *R* software. Vasishth and Nicemboim (2016) present the fundamental principles of inferential statistics based on the frequentist approach, as well as the practices to avoid when using them. The article by Clark (1973) is the reference concerning the presence of random effects related not only to participants but also to items. Brysbaert (2007) illustrates this question in a simple way and presents the interest of analyses per participant and per item within the framework of classical analyses, such as the Student test or ANOVA, and also using mixed linear models. For more information on mixed linear models, we recommend Baayen *et al*. (2008), as well as Barr *et al*. (2013), Bates *et al*. (2015), Luke (2017) and Matuschek *et al*. (2017). For a more accessible presentation of the use of mixed linear models, Winter (2013, 2019) and Singmann and Kellen (2020) are excellent resources. A step-by-step description of data processing, descriptive statistics, the construction of mixed models and their interpretation using *R* is provided by Singmann (2019). Finally, Dienes (2008) presents the scientific philosophy and the reasoning on which the different statistical models are based. It is a very interesting resource for understanding the principles of statistical inference, as well as the different possibilities offered for data analysis.

References

Abbuhl, R., Gass, S., and Mackey, A. (2013). Experimental research designs. In *Research Methods in Linguistics*, Podesva, R.J. and Sharma, D. (eds). Cambridge University Press, Cambridge, 116–134.

Aguasvivas, J.A., Carreiras, M., Brysbaert, M., Mandera, P., Keuleers, E., and Duñabeitia, J.A. (2018). SPALEX: A Spanish lexical decision database from a massive online data collection. *Frontiers in Psychology*, 9, 2156.

Aitchison, J. (2012). *Words in the Mind: An Introduction to the Mental Lexicon*, 4th edition. John Wiley & Sons, New York.

Algom, D., Chajut, E., and Lev, S. (2004). A rational look at the emotional Stroop phenomenon: A generic slowdown, not a Stroop effect. *Journal of Experimental Psychology: General*, 133(3), 323–338.

Athanasopoulos, P., Damjanovic, L., Krajciova, A., and Sasaki, M. (2011). Representation of colour concepts in bilingual cognition: The case of Japanese blues. *Bilingualism: Language and Cognition*, 14, 9–17.

Augustinova, M., Almeida, E., Clarys, D., Ferrand, L., Izaute, M., Jalenques, I., Juneau, C., Normand, A., and Silvert, L. (2016). Que mesure l'interférence Stroop ? Quand et comment ? Arguments méthodologiques et théoriques en faveur d'un changement de pratiques dans sa mesure. *L'Année psychologique*, 116(1), 45–66.

Avanzi, M. (2019). "Yaourt" ou "yoghourt" ? *Français de nos régions* [Online]. Available at: https://francaisdenosregions.com/2019/02/01/yahourt-ou-yoghourt/ [Accessed 5 April 2019].

Avanzi, M., Barbet, C., Glikman, J., and Peuvergne, J. (2016). Présentation d'une enquête pour l'étude des régionalismes du français. *SHS Web of Conferences*, 27, 03001.

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

Bååth, R. (2010). ChildFreq: An online tool to explore word frequencies in child language. *LUCS Minor*, 16, 1–6

Baayen, R.H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database [CD-ROM]. *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia.

Baayen, R.H., Feldman, L.B., and Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55(2), 290–313.

Baayen, R.H., Davidson, D.J., and Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., and Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459.

Bard, E.G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1), 32–68.

Baron-Cohen, S. and Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2), 163–175.

Barr, D.J., Gann, T.M., and Pierce, R.S. (2011). Anticipatory baseline effects and information integration in visual world studies. *Acta Psychologica*, 137(2), 201–207.

Barr, D.J., Levy, R., Scheepers, C., and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Beaver, J. and Carter, M. (2019). *Developmental Reading Assessment*, 3rd edition. Pearson Assessments, Minneapolis.

Berko, J. (1958). The child’s learning of English morphology. *Word*, 14(2–3), 150–177.

Berman, R.A. and Slobin, D.I. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Lawrence Erlbaum Associates, Mahwah.

Bernicot, J., Laval, V., and Chaminaud, S. (2007). Nonliteral language forms in children: In what order are they acquired in pragmatics and metapragmatics? *Journal of Pragmatics*, 39(12), 2115–2132.

Bernstein Ratner, N. (2000). Elicited imitation and other methods for the analysis of trade-offs between speech and language skills in children. In *Methods for Studying Language Production*, Menn, L. and Bernstein Ratner, N. (eds). Lawrence Erlbaum Associates, Mahwah, 291–312.

Besner, D., Stolz, J.A., and Boutilier, C. (1997). The Stroop effect and the myth of automaticity. *Psychonomic Bulletin & Review*, 4(2), 221–225.

Blanc, N., Kendeou, P., van den Broek, P., and Brouillet, D. (2008). Updating situation models during reading of news reports: Evidence from empirical data and simulations. *Discourse Processes*, 45(2), 103–121.

Bloom, P. (2002). *How Children Learn the Meanings of Words?* MIT Press, Cambridge.

Bloom, L., Hood, L., and Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, 6(3), 380–420.

Bonin, P. (2013). *Psychologie du langage : la fabrique des mots*, 2nd edition. De Boeck Supérieur, Brussels.

Borghi, A., Caramelli, N., and Setti, A. (2016). How abstract is risk for workers? Expertise, context and introspection in abstract concepts. *Reti, saperi, linguaggi, Italian Journal of Cognitive Sciences*, 95–118.

Boroditsky, L., Schmidt, L.A., and Phillips, W. (2003). Sex, syntax, and semantics. In *Language in Mind: Advances in the Study of Language and Thought*, Gentner, D. and Goldin-Meadow, S. (eds). MIT Press, Cambridge, 61–79.

Boroditsky, L., Fuhrman, O. and McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition*, 118, 123–129.

Branigan, H.P. and Pickering, M.J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, 40, E282.

Branigan, H.P., Pickering, M.J., Liversedge, S.P., Stewart, A.J., and Urbach, T.P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, 24(6), 489–506.

Bransford, J.D. and Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 717–726.

Bransford, J.D., Barclay, J.R., and Franks, J.J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, 3(2), 193–209.

Brysbaert, M. (2007). “The language-as-fixed-effect fallacy”: Some simple SPSS solutions to a complex problem. Report, version 2, Royal Holloway, University of London.

Brysbaert, M. (2013). Lextale_FR a fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, 53(1), 23–37.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.

Brysbaert, M. and Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), 1–20.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., and Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5), 412–424.

Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, 42, 441–458.

Cai, Q. and Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS ONE*, 5(6), e10729

Carlson-Radvansky, L.A. and Radvansky, G.A. (1996). The influence of functional relations on spatial term selection. *Psychological Science*, 7(1), 56–60.

Carreiras, M., Carriedo, N., Alonso, M.A., and Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, 25, 438–446.

Carreiras, M., Ferrand, L., Grainger, J., and Perea, M. (2005). Sequential effects of phonological priming in visual word recognition. *Psychological Science*, 16(8), 585–589.

Carroll, J.B., Davies, P., and Richman, B. (1971). *The American Heritage Word Frequency Book*. Houghton Mifflin, New York.

Chan, A., Meints, K., Lieven, E., and Tomasello, M. (2010). Young children’s comprehension of English SVO word order revisited: Testing the same children in act-out and intermodal preferential looking tasks. *Cognitive Development*, 25(1), 30–45.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359.

Clifton, C., Staub, A., and Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements. A Window on Mind and Brain*, van Gompel, R.P.G., Fischer, M.H., Murray, W.S., and Hill, R.L. (eds). Elsevier, Oxford, 341–371.

Coady, J.A. and Evans, J.L. (2008). Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language & Communication Disorders*, 43(1), 1–40.

Colonna, S., Schimke, S., and Hemforth, B. (2012). Information structure effects on anaphora resolution in German and French: A crosslinguistic study of pronoun resolution. *Linguistics*, 50(5), 991–1013.

Colston, H.L. and Gibbs, R.W. (2002). Are irony and metaphor understood differently? *Metaphor and Symbol*, 17(1), 57–80.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4), 497–505.

Connell, L. (2007). Representing object colour in language comprehension. *Cognition*, 102(3), 476–485.

Connell, L. and Lynott, D. (2009). Is a bear white in the woods? Parallel representation of implied object color during language comprehension. *Psychonomic Bulletin & Review*, 16(3), 573–577.

Coventry, K.R., Prat-Sala, M., and Richards, L. (2001). The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3), 376–398.

Crepaldi, D., Amenta, S., Pawel, M., Keuleers, E., and Brysbaert, M. (2015). SUBTLEX-IT. Subtitle-based word frequency estimates for Italian. *Proceedings of the Annual Meeting of the Italian Association For Experimental Psychology*. September 10–12, Rovereto, Italy.

Cuetos, F., Glez-Nosti, M., Barbón, A., and Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32(2), 133–144.

Culicover, P.W. and Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences*, 14(6), 234–235.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.

Dąbrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1), 1–23.

Davis, M.H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10(4), 85.

DeKeyser, R. and Larson-Hall, J. (2005). What does the critical period really mean. In *Handbook of Bilingualism: Psycholinguistic Approaches*, Kroll, J.F. and de Groot, A.M.B. (eds). Oxford University Press, Oxford, 88–108.

Dell'Acqua, R. and Grainger, J. (1999). Unconscious semantic priming from pictures. *Cognition*, 73(1), B1–B15.

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference.* Macmillan International, Basingstoke.

Dienes, Z. and McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 25(1), 207–218.

Dimitropoulou, M., Duñabeitia, J.A., Avilés, A., Corral, J., and Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in Psychology*, 1, 218.

Drieghe, D. and Brysbaert, M. (2002). Strategic effects in associative priming with words, homophones, and pseudohomophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(5), 951–961.

Dunn, L.M. and Dunn, D.M. (2007). *Peabody Picture Vocabulary Test*, 4th edition. Pearson Assessments, Minneapolis.

Durrleman, S., Hippolyte, L., Zufferey, S., Iglesias, K., and Hadjikhani, N. (2015). Complex syntax in autism spectrum disorders: A study of relative clauses. *International Journal of Language & Communication Disorders*, 50(2), 260–267.

Eilola, T.M., Havelka, J., and Sharma, D. (2007). Emotional activation in the first and second language. *Cognition & Emotion*, 21(5), 1064–1076.

Eisenbeiss, S. (2010). Production methods in language acquisition research. In *Experimental Methods in Language Acquisition Research*, Blom, E. and Unsworth, S. (eds). John Benjamins, Amsterdam, 11–34.

Engelen, J.A., Bouwmeester, S., de Bruin, A.B.H., and Zwaan, R.A. (2014). Eye movements reveal differences in children's referential processing during narrative comprehension. *Journal of Experimental Child Psychology*, 118, 57–77.

Erickson, T.D. and Mattson, M.E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5), 540–551.

Evers-Vermeul, J. and Sanders, T. (2011). Discovering domains – On the acquisition of causal connectives. *Journal of Pragmatics*, 43(6), 1645–1662.

Faul, F., Erdfelder, E., Lang, A.G., and Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.

Ferrand, L. (2001). *Cognition et lecture : processus de base de la reconnaissance des mots écrits chez l'adulte*. De Boeck, Brussels.

Ferrand, L. and Grainger, J. (1992). Phonology and orthography in visual word recognition: Evidence from masked non-word priming. *The Quarterly Journal of Experimental Psychology Section A*, 45(3), 353–372.

Ferrand, L. and Grainger, J. (1993). The time course of orthographic and phonological code activation in the early phases of visual word recognition. *Bulletin of the Psychonomic Society*, 31(2), 119–122.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2), 488–496.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203.

Ferreira, F. and Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes*, 56(7), 485–495.

Ferreira, F., Bailey, K.G.D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 11–15.

Field, A. and Hole, G. (2003). *How to Design and Report Experiments*. Sage, London.

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. Sage, London.

Foster, E.D. and Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA*, 105(2), 203–206.

Gass, S.M. (2015). Experimental research. In *Research Methods in Applied Linguistics*, Paltridge, B. and Phakiti, A. (eds). Bloomsbury Publishing PLC, London, 180–209.

Gass, S.M. and Mackey, A. (2007). *Data Elicitation for Second and Foreign Language Research*. Routledge, Abingdon-on-Thames.

Gelman, A. and Loken, E. (2014). Ethics and statistics: The AAA tranche of subprime science. *Chance*, 27(1), 51–56.

Gibson, E. and Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6), 233–234.

Gibson, E. and Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1–2), 88–124.

Gillioz, C., Gygax, P., and Tapiero, I. (2012). Individual differences and emotion inferences during reading comprehension. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66, 239–250.

Gonzalez-Marquez, M., Becker, R.B., and Cutting, J.E. (2007a). An introduction to experimental methods for language researchers. In *Methods in Cognitive Linguistics*, Gonzalez-Marquez, M., Mittelberg, I., Coulson, S., and Spivey, M.J. (eds). John Benjamins Publishing, Amsterdam, 53–86.

Gonzalez-Marquez, M., Mittelberg, I., Coulson, S., and Spivey, M.J. (eds) (2007b). *Methods in Cognitive Linguistics*. John Benjamins Publishing, Amsterdam.

Gordon, P.C. and Hendrick, R. (1997). Intuitive knowledge of linguistic co-reference. *Cognition*, 62(3), 325–370.

Gordon, P.C., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1411–1423.

Gordon, P.C., Hendrick, R., Johnson, M., and Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(6), 1304–1321.

Halekoh, U. and Højsgaard, S. (2014). A Kenward–Roger approximation and parametric bootstrap methods for tests in linear mixed models – The R package pbkrtest. *Journal of Statistical Software*, 59(9), 1–30.

Havik, E., Roberts, L., Van Hout, R., Schreuder, R., and Haverkort, M. (2009). Processing subject–object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning*, 59(1), 73–112.

van Heuven, W.J.B., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6), 1176–1190.

Hoeben Mannaert, L.N., Dijkstra, K., and Zwaan, R.A. (2017). Is color an integral part of a rich mental simulation? *Memory & Cognition*, 45(6), 974–982.

Huettig, F., Rommers, J., and Meyer, A.S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Jegerski, J. (2014). Self-paced reading. In *Research Methods in Second Language Psycholinguistics*, Jegerski, J. and VanPatten, B. (eds). Routledge, Abingdon-on-Thames, 20–49.

Judd, C.M., Westfall, J., and Kenny, D.A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69.

Just, M.A. and Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.

Kaiser, E. (2013). Experimental paradigms in psycholinguistics. In *Research Methods in Linguistics*, Podesva, R.J. and Sharma, D. (eds). Cambridge University Press, Cambridge, 135–168.

Karmiloff, K. and Karmiloff-Smith, A. (2003). *Comment les enfants entrent dans le langage ?* Retz, Paris.

Kehler, A. and Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1–2), 1–37.

Kelter, S., Kaup, B., and Claus, B. (2004). Representing a described sequence of events: A dynamic view of narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 451–464.

Kenward, M.G. and Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997.

Keuleers, E., Brysbaert, M., and New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44(1), 287–304.

Kidd, E., Donnelly, S., and Christiansen, M.H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 22(2), 154–169.

Kim, Y. and McDonough, K. (2008). Learners' production of passives during syntactic priming activities. *Applied Linguistics*, 29(1), 149–154.

Kim, J., Gabriel, U., and Gygax, P. (2019). Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PLoS ONE*, 14(9), e0221802.

Kissine, M., Cano-Chervel, J., Carlier, S., Brabanter, P.D., Ducenne, L., Pairon, M.-C., Deconinck, N., Delvenne, V., and Leybaert, J. (2015). Children with autism understand indirect speech acts: Evidence from a semi-structured act-out task. *PLoS ONE*, 10(11), e0142191.

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22, 519–534.

Kuznetsova, A., Brockhoff, P.B., and Christensen, R.H.B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710.

Lassonde, K.A. and O'Brien, E.J. (2009). Contextual specificity in the activation of predictive inferences. *Discourse Processes*, 46, 426–438.

Lemhöfer, K. and Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44(2), 325–343.

Lenth, R.V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1), 1–33.

Levinson, S.C. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In *Language and Space*, Bloom, P., Peterson, M.A., Nadel, L., and Garrett, M.F. (eds). MIT Press, Cambridge, 109–169.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766.

Litosseliti, L. (ed.) (2018). *Research Methods in Linguistics*, 2nd edition. Bloomsbury Publishing PLC, London.

Logan, G.D. and Sadler, D.D. (1996). A computational analysis of the apprehension of spatial relations. In *Language and Space*, Bloom, P., Peterson, M.A., Nadel, L., and Garrett, M.F. (eds). MIT Press, Cambridge, 493–529.

Luke, S.G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49, 1494–1502.

MacKay, D.G. and Ahmetzanov, M.V. (2005). Emotion, memory, and attention in the taboo Stroop paradigm: An experimental analogue of flashbulb memories. *Psychological Science*, 16(1), 25–32.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah.

Madden, C.J. and Zwaan, R.A. (2003). How does verb aspect constrain event representations? *Memory & Cognition*, 31, 663–672.

Mandera, P., Keuleers, E., Wodniecka, Z., and Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47(2), 471–483.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.

Mayer, M. (1969). *Frog, Where Are You?* Dial Book for Young Readers, New York.

McClelland, G.H., (2014). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In *Handbook of Research Methods in Social and Personality Psychology*, Reis, H.T. and Judd, C.M. (eds), 2nd edition. Cambridge University Press, Cambridge, 608–626.

McConkie, G.W. and Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6), 578–586.

McDaniel, D., Cairns, H.S., and McKee, C. (1998). *Methods for Assessing Children's Syntax*. MIT Press, Cambridge.

McKenna, F.P. and Sharma, D. (1995). Intrusive cognitions: An investigation of the emotional Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6), 1595–1607.

Meier, B.P. and Robinson, M.D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological Science*, 15(4), 243–247.

Menn, L. and Bernstein Ratner, N. (eds) (2000). *Methods for Studying Language Production*. Lawrence Erlbaum Associates, Mahwah.

New, B. (2006). Lexique 3 : une nouvelle base de données lexicales. In *Actes de la Conférence Traitement Automatique des Langues Naturelles, TALN 2006*. Louvain, Belgium.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4), 661–677.

Ortega, L. (2014). *Understanding Second Language Acquisition*. Routledge, Abingdon-on-Thames.

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J.K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203.

Perea, M., Vergara-Martínez, M., and Gomez, P. (2015). Resolving the locus of cAsE aLtErNaTiOn effects in visual word recognition: Evidence from masked priming. *Cognition*, 142, 39–43.

Phakiti, A. (2015). Quantitative research and analysis. In *Research Methods in Applied Linguistics*. Paltridge, B. and A. Phakiti (eds). Bloomsbury Publishing PLC, London, 180–209.

Pinker, S. (1994). *The Language Instinct*. Harper Perennial Modern Classics, New York.

Potter, M.C. and Levy, E.I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1), 10–15.

Pouscoulous, N., Noveck, I.A., Politzer, G., and Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14(4), 347–375.

Price, P.C., Jhangiani, R., and Chiang, I.-C.A. (2013). *Research Methods in Psychology*, 2nd edition. BCcampus, Vancouver.

Psychology Software Tools, Inc. (2016). E-Prime: Documentation article [Online]. Available at: https://www.pstnet.com.

Rasinger, S.M. (2010). Quantitative methods: Concepts, frameworks and issues. In *Research Methods in Linguistics*, Litosseliti, L. (ed.). Continuum, New York, 49–67.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457–1506.

Rayner, K. and Bertera, J.H. (1979). Reading without a fovea. *Science*, 206(4417), 468–469.

R Development Core Team (2016). R: A language and environment for statistical computing *R Foundation for Statistical Computing*, Vienna [Online]. Available at: http://www.R-project.org/.

Reimers, S. and Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, 39(3), 365–370.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, 49, 243–256.

Rossi, J.-P. (2008). *Psychologie de la compréhension du langage*. De Boeck, Brussels.

Rowland, C. (2013). *Understanding Child Language Acquisition*. Routledge, Abingdon-on-Thames.

Royle, P. and Reising, L. (2019). Elicited and spontaneous determiner phrase production in French-speaking children with developmental language disorder. *Canadian Journal of Speech-Language Pathology and Audiology*, 43(3), 167–187.

Sanford, A.J. and Graesser, A.C. (2006). Shallow processing and underspecification. *Discourse Processes*, 42, 99–108.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110–114.

Schilling, N. (2013). Surveys and interviews. In *Research Methods in Linguistics*, Podesva, R.J. and Sharma, D. (eds). Cambridge University Press, Cambridge, 96–115.

Schmalhofer, F. and Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language*, 25, 279–294.

Schubert, T.W., Murteira, C., Collins, E.C., and Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS ONE*, 8(6), e67769.

Schumann, J. and Zufferey, S. (2020). Connectives and straw men. Experimental approach on French and English. In *Proceedings of the 12th OSSA Conference: Evidence, Persuasion & Diversity*, OSSA Conference Archives, University of Windsor, Canada.

Schumann, J., Zufferey, S., and Oswald, S. (2019). What makes a straw man acceptable? Three experiments assessing linguistic factors. *Journal of Pragmatics*, 141, 1–15.

Schütze, C.T. (2016). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Language Science Press, Berlin.

Schütze, C.T. and Sprouse, J. (2013). Judgment data. In *Research Methods in Linguistics*, Podesva, R.J. and Sharma, D. (eds). Cambridge University Press, Cambridge, 27–50.

Schwab, S. and Avanzi, M. (2015). Regional variation and articulation rate in French. *Journal of Phonetics*, 48, 96–105.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366.

Singmann, H. (2019). Mixed model reanalysis of RT data [Online]. Available at: https://cran.r-roject.org/web/packages/afex/vignettes/afex_mixed_example.html [Accessed 24 June 2019].

Singmann, H. and Kellen, D. (2020). An introduction to mixed models for experimental psychology. In *New Methods in Cognitive Psychology*, Spieler, D. and Schumacher, E. (eds). Routledge, Abingdon-on-Thames, 4–32.

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (2017). Afex: Analysis of factorial experiments. R package version 0.17-8 [Online]. Available at: http://cran.r-project.org/package=afex.

Skordos, D. and Papafragou, A. (2012). Lexical alternatives improve 5-year-olds' ability to compute scalar implicatures. In *BUCLD 36 Online Proceedings Supplement* [Online]. Available at: http://www.bu.edu/bucld/files/2012/07/Skordos-36.pdf.

Smith, N. (2004). *Chomsky: Ideas and Ideals*. Cambridge University Press, Cambridge.

Smith, N.J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

Soares, A.P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J.J., Comesaña, M., and Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology*, 68(4), 680–696.

Spivey, M.J. and Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science*, 10(3), 281–284.

Staub, A. and Rayner, K. (2007). Eye movements and on-line comprehension processes. In *The Oxford Handbook of Psycholinguistics*, Gaskell, M.G. (ed.). Oxford University Press, Oxford, 327–342.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643–662.

von Stutterheim, C. and Nüse, R. (2003). Processes of conceptualization in language production: Language-specific perspectives and event construal. *Linguistics*, 41(5), 851–882.

Sze, W.P., Rickard Liow, S.J., and Yap, M.J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, 46(1), 263–273.

Vasishth, S. and Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational Ideas – Part I. *Language and Linguistics Compass*, 10(8), 349–369.

de Vega, M., Diaz, J.M., and Leon, I. (1997). To know or not to know: Comprehending protagonists' beliefs and their emotional consequences. *Discourse Processes*, 23, 169–92.

Vigliocco, G., Vinson, D.P., Paganelli, F., and Dworzynski, K. (2005). Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General*, 134(4), 501–520.

Wagner, E. (2015). Survey research. In *Research Methods in Applied Linguistics*, Paltridge, B. and Phakiti, A. (eds). Bloomsbury Publishing PLC, London, 166–197.

Weskott, T. and Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 87(2), 249–273.

Williams, J.M.G., Mathews, A., and MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, 120(1), 3–24.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications [Online]. Available at: http://arxiv.org/abs/1308.5499 [Accessed 22 September 2018].

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R*. Routledge, Abingdon-on-Thames.

Ziegler, J.C., Ferrand, L., Jacobs, A.M., Rey, A., and Grainger, J. (2000). Visual and phonological codes in letter and word recognition: Evidence from incremental priming. *The Quarterly Journal of Experimental Psychology Section A*, 53(3), 671–692.

Zufferey, S. (2020). *Introduction to Corpus Linguistics*. ISTE Ltd, London and John Wiley & Sons, New York.

Zufferey, S. and Gygax, P. (2020). "Roger broke his tooth. However, he went to the dentist": Why some readers struggle to evaluate wrong (and right) uses of connectives. *Discourse Processes*, 57, 184-200.

Zufferey, S., Mak, W.M., and Sanders, T.J.M. (2015a). A cross-linguistic perspective on the acquisition of causal connectives and relations. *International Review of Pragmatics*, 7(1), 22–39.

Zufferey, S., Mak, W., Degand, L., and Sanders, T. (2015b). Advanced learners' comprehension of discourse connectives: The role of L1 transfer across on-line and off-line tasks. *Second Language Research*, 31(3), 389–411.

Index

Introduction to Experimental Linguistics,
First Edition. Christelle Gillioz and Sandrine Zufferey.

Other titles from

ISTE

in

Cognitive Science and Knowledge Management

2020

NGUYEN-XUAN Anh
Cognitive Mechanisms of Learning (Learning, Development and Cognitive Technologies Set)

OSIURAK François
The Tool Instinct

ZUFFEREY Sandrine
Introduction to Corpus Linguistics

2019

CLAVEL Chloé
Opinion Analysis in Interactions: From Data Mining to Human-Agent Interaction

KAROUI Jihen, BENAMARA Farah, MORICEAU Véronique
Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media

MARTINOT Claire, BOŠNJAK BOTICA Tomislava, GEROLIMICH Sonia, PAPROCKA-PIOTROWSKA Urszula
Reformulation and Acquisition of Linguistic Complexity: Crosslinguistic Perspective
(Interaction of Syntax and Semantics in Discourse Set – Volume 2)

2018

BONFANTE Guillaume, GUILLAUME Bruno, PERRIER Guy
Application of Graph Rewriting to Natural Language Processing
(Logic, Linguistics and Computer Science Set – Volume 1)

PENNEC Blandine
Discourse Readjustment(s) in Contemporary English
(Interaction of Syntax and Semantics in Discourse Set – Volume 1)

2017

KURDI Mohamed Zakaria
Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications

MAESSCHALCK Marc
Reflexive Governance for Research and Innovative Knowledge
(Responsible Research and Innovation Set – Volume 6)

PELLÉ Sophie
Business, Innovation and Responsibility
(Responsible Research and Innovation Set - Volume 7)

2016

BOUVARD Patricia, SUZANNE Hervé
Collective Intelligence Development in Business

CLERC Maureen, BOUGRAIN Laurent, LOTTE Fabien
Brain–Computer Interfaces 1: Foundations and Methods
Brain–Computer Interfaces 2: Technology and Applications

FORT Karën
Collaborative Annotation for Reliable Natural Language Processing

GIANNI Robert
Responsibility and Freedom
(Responsible Research and Innovation Set – Volume 2)

GRUNWALD Armin
The Hermeneutic Side of Responsible Research and Innovation
(Responsible Research and Innovation Set – Volume 5)

KURDI Mohamed Zakaria
Natural Language Processing and Computational Linguistics 1: Speech, Morphology and Syntax

LENOIR Virgil Cristian
Ethical Efficiency: Responsibility and Contingency
(Responsible Research and Innovation Set – Volume 1)

MATTA Nada, ATIFI Hassan, DUCELLIER Guillaume
Daily Knowledge Valuation in Organizations

NOUVEL Damien, EHRMANN Maud, ROSSET Sophie
Named Entities for Computational Linguistics

PELLÉ Sophie, REBER Bernard
From Ethical Review to Responsible Research and Innovation
(Responsible Research and Innovation Set - Volume 3)

REBER Bernard
Precautionary Principle, Pluralism and Deliberation
(Responsible Research and Innovation Set – Volume 4)

SILBERZTEIN Max
Formalizing Natural Languages: The NooJ Approach

2015

LAFOURCADE Mathieu, JOUBERT Alain, LE BRUN Nathalie
Games with a Purpose (GWAPs)

SAAD Inès, ROSENTHAL-SABROUX Camille, GARGOURI Faïez
Information Systems for Knowledge Management

2014

DELPECH Estelle Maryline
Comparable Corpora and Computer-assisted Translation

FARINAS DEL CERRO Luis, INOUE Katsumi
Logical Modeling of Biological Systems

MACHADO Carolina, DAVIM J. Paulo
Transfer and Management of Knowledge

TORRES-MORENO Juan-Manuel
Automatic Text Summarization

2013

TURENNE Nicolas
Knowledge Needs and Information Extraction: Towards an Artificial Consciousness

ZARATÉ Pascale
Tools for Collaborative Decision-Making

2011

DAVID Amos
Competitive Intelligence and Decision Problems

LÉVY Pierre
The Semantic Sphere: Computation, Cognition and Information Economy

LIGOZAT Gérard
Qualitative Spatial and Temporal Reasoning

PELACHAUD Catherine
Emotion-oriented Systems

QUONIAM Luc
Competitive Intelligence 2.0: Organization, Innovation and Territory

2010

ALBALATE Amparo, MINKER Wolfgang
Semi-Supervised and Unsupervised Machine Learning: Novel Strategies

BROSSAUD Claire, REBER Bernard
Digital Cognitive Technologies

2009

BOUYSSOU Denis, DUBOIS Didier, PIRLOT Marc, PRADE Henri
Decision-making Process

MARCHAL Alain
From Speech Physiology to Linguistic Phonetics

PRALET Cédric, SCHIEX Thomas, VERFAILLIE Gérard
Sequential Decision-Making Problems / Representation and Solution

SZÜCS Andras, TAIT Alan, VIDAL Martine, BERNATH Ulrich
Distance and E-learning in Transition

2008

MARIANI Joseph
Spoken Language Processing