Assessing Grammaticality

Fiona Scallan
B.A. (Mod.) CSLL
Final Year Project, May 2012
Supervisor: Dr. Carl Vogel
Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university.

__________________________ April 4, 2012
Fiona Scallan
Permission to Lend

I agree that the Library and other agents of the College may lend or copy this thesis upon request.

__________________________  April 4, 2012
Fiona Scallan
Acknowledgements

First and foremost I would like to thank Carl Vogel for all of his help throughout this year, and indeed over all four years. Carl always had a calming word to say on days when I felt stressed and under pressure, he always said that there was time and not to worry, I guess he was right in the end.

My parents are the next obvious people to thank as if they had not provided me with the opportunity to attend college and see me through the four years, this dissertation would not even exist right now.

Finally, I would like to thank John, my brothers, and my friends for listening to my rants and complaints when things were not running as smoothly as I had perhaps hoped, and naturally for completing some surveys for me.

Thank you everybody.
You can be a little ungrammatical if you come from the right part of the country. (Robert Frost)
# Contents

1 Introduction .......................................................... 1

2 Grammaticality ......................................................... 3
  2.1 Introduction ...................................................... 3
  2.1.1 What is Grammaticality? ..................................... 3
  2.1.2 Degrees of Ungrammaticality ................................. 7
  2.1.3 Stimulus Factors .............................................. 8
  2.1.4 Definitions of Grammaticality Judgements ................. 11
  2.1.5 Sentence Processing ........................................... 12
  2.1.6 Conclusion .................................................... 15

3 Acceptability .......................................................... 16
  3.1 Introduction ...................................................... 16
  3.1.1 What is Acceptability? ....................................... 16
  3.1.2 Judging Acceptability ........................................ 20
  3.1.3 Acceptability and Unacceptability in Dialects .......... 21
  3.1.4 Differences between Acceptability and Grammaticality 22
  3.1.5 Why is a sentence judged to be ungrammatical or un-acceptable? .................................................. 23
  3.1.6 Conclusion .................................................... 25

4 Experiment Background ............................................... 27
  4.1 Introduction ...................................................... 27
  4.1.1 Types of Scales ................................................. 27
  4.1.2 The Experiment ................................................ 32
  4.1.3 Conclusion .................................................... 43
# CONTENTS

5 Results and Discussion 44
  5.1 Introduction ........................................ 44
  5.1.1 Statistical Background ............................ 45
  5.1.2 Working with the Data ............................... 46
  5.1.3 Analysis ........................................... 47
  5.1.4 Grammatical Controls ............................... 55
  5.1.5 Ungrammatical controls ............................. 58
  5.1.6 Test Items ......................................... 59
  5.2 Conclusion ............................................ 62

6 Conclusion 64

Bibliography 66

A Instructions To Candidates 72

B Tables 81

C Html Code 88
List of Figures

5.1 Age corresponding to Rating on Likert Scale . . . . . . . . . . . . . . . . . . . . . . . . . . 56
B.1 Information sheet for Participants . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
B.2 Example of Question One . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86
B.3 Example of Question Two . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
B.4 Conent Form . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
Abstract

The current dissertation investigated the claim that Magnitude Estimation is the best method for assessing grammaticality. In order to conduct such research, eight grammaticality judgement tests were distributed online through the use of an online survey tool known as Limesurvey. There were four main methods used; two binary scales differing in the way the question was worded, the Likert scale and of course Magnitude Estimation. As such a tool was not readily available within Limesurvey, in order to account for the randomisation of the question order the four experiments mentioned had their question order reversed, creating four further experiments and this is how the final total of eight emerged. The items analysed were taken from a paper by (Rimmer, 2006) and the results provided a clear cut answer to the claim listed above, in vague terms, no. Overall the results showed that there exist no distinct differences between the scales of Magnitude Estimation, the Likert scale and the binary scale in relation to grammaticality judgement tests. The current dissertation provides an account of grammaticality and acceptability as a whole, as well as the results obtained in the experiment, implications for future work and the conclusions drawn from the study.
Chapter 1

Introduction

Have you ever encountered a sentence or an utterance which you initially felt was grammatical, only to find that after some thought you believed it to be ungrammatical? Have you ever questioned the reasons as to why humans are capable of making such distinctions or why an utterance may be acceptable upon one reading but not upon another? Have you ever wondered what the best method for assessing grammaticality is or whether it should be analysed as a gradient concept or by a simple binary distinction of grammatical or ungrammatical? These are the central ideas under investigation in this dissertation. What are the factors which govern such an abstract notion such as grammaticality and do the methods available for assessing grammaticality differ in their analysis of grammaticality judgements? Grammaticality and acceptability judgement tests are a relatively new age methodology, which is a growing area of research. The method of Magnitude Estimation has its origins in psychophysics for analysing physical stimuli, however has made its way into grammaticality judgement experiments and at present stands as the most informative test in the area, according to many linguists. Though the test may allow for as many levels of ungrammaticality as desired, the act of carrying out the test of Magnitude Estimation is not as user friendly as previous methods in the area. Therein lies the motivation for the present study, is it really necessary to burden participants with all of the detail required for Magnitude Estimation? The current study found no significant difference between each scale, with a significant positive correlation also existing between them.

The present dissertation begins with an overview of the relevant literature in the areas of grammaticality and acceptability, as well as providing
some implications as to why one judges a sentence to be grammatical or un-
grammatical and the differences between grammaticality and acceptability
judgements. The report then continues to explain the methodology behind
each individual scale and gives an overview of how they differ from each other
and what arguments exist for and against certain factors within them. The
report then continues to explain how the experimental items were arranged,
what factors were taken into account when conducting each survey and how
the data was organised in order to be able to work with it within the statisti-
cal platform known as R. The areas of the project which did not go according
to plan will also be discussed in this section.

The final chapters then give an account of the results obtained from the
experiment, as well as an in depth discussion of the analysis and further
factors which could be taken into account if the present experiment were to
be recreated in the future.
Chapter 2
Grammaticality

2.1 Introduction

As the present study entails grammaticality, an obvious starting point is to define what exactly the notion of grammaticality means and what factors govern such a concept. The following chapter contributes an overview of the relevant literature in the area, whilst simultaneously introducing the notion of grammaticality, how grammaticality judgements are formed and why they are important. The basis and factors upon which a sentence is assigned the status of being ungrammatical and the best methods for analysing such judgements are the main ideas under investigation in this dissertation.

2.1.1 What is Grammaticality?

Whether a sentence is grammatical or not is a question of competence (Chomsky, 1965). This definition has been adopted by many linguists, such as (Fetzer, 2004), who claims that the theoretical notion of grammaticality presupposes a native speaker with linguistic competence, which provides them with the ability to differentiate between grammatical and ungrammatical sentences, and also provides them with the ability to produce and interpret grammatical sentences. Grammaticality itself is therefore based on premises that language is a rule-governed system which is defined by rules and constraints, and sentences or constructions which are produced in accordance with these rules and constraints are grammatical, and the opposite are ungrammatical. Fetzer also claims that it is the native speaker who assigns a sentence the status of being grammatical or ungrammatical. Speakers are
only seen as native if they can speak properly by avoiding grammatical mistakes in sentences. However, native speakers make mistakes with grammar every day, in every day speech. So to say that a non-native is only native like when they no longer make mistakes with grammar in speech, is not really a very fair description is it? I will give examples of such bad grammar used by native speakers later on in this dissertation. As a quick example in support of this argument, one of the items analysed in this dissertation was, 'either you or I are wrong'. Although this sentence technically violates a grammatical rule, this sentence was generally seen as acceptable to the vast majority of subjects, however there were a few discrepancies. Some believed that 'Either you or I am wrong' would be better. Other responses indicated a certain amount of doubt in their response, such as 'not sure if it’s 100% ‘grammatically’ correct - but perfectly common and acceptable’, or, 'I think this is correct...’. Nonetheless the majority of people claimed the sentence made perfect sense. This is in keeping with the findings of (Rimmer, 2006), where the items were originally found, in which 60.7% of his participants judged the sentence acceptable, compared to an average of 74% in the present case. Both results show a relatively high acceptability rate for such an utterance, despite its nature as a technically ungrammatical sentence.

A grammatical sentence is generated by the speakers grammar. In order to answer the question of what grammaticality is, one must first draw their attention to the underlying internal grammar that each individual possesses. In other words, what do grammars actually stand for? Linguistic intuitions of native speakers? A certain set of rules? Systematic language use? This is one of the crucial problems of current grammatical theories and opinion differs somewhat on the answer. It has been stressed that linguistic abilities and that cognitive strategies play an important role in our verbal performances (Van Dijk, 1977, p.39). Carson T. Schütze’s view is that grammar is what linguistics traditionally defines as the mental representation of the generative system for the language we have actually acquired (Schütze, 2005, p.462).

However, two of the most widespread criticisms of generative grammar are that it involves constructing theories of intuition rather than of language use and that it is highly subjective and biased by the views of the linguist (Schütze, 1996, p.1). It is therefore important to formulate linguistic theories based on the average native speaker, rather than relying on the notions of the linguist themselves and this strategy was used in the present experiment. A further problem with defining what grammars stand for is that it is based on tacit knowledge, and so the formulation of a concrete answer is very difficult.
CHAPTER 2. GRAMMATICALITY

to construct. Nonetheless the following paragraph attempts to develop a further insight into such a complex area, by taking a more in depth look at what the internal grammar of a person actually consists of.

The internal grammar of a person is a grammar, which is free of mistakes to them. Chomsky (1965) clarifies this point in stating that although the grammar of a person expresses his or her knowledge of a language, the person is not aware of the rules, and the set of rules he is adhering to may not necessarily be correct. A person attempting to judge the grammaticality of an example in his/her native language typically asks, introspectively, 'Would/could I say this? How does this sound/feel to me?' Here the accent is on 'I' and 'me', that is, the concern is not with whether an example conforms to the norms of some external linguistic community, but with whether it conforms to the individual’s internal grammar (Bley-Vroman & Masterson, 1989, p.212). When judging the grammaticality of an English sentence, one is deciding whether a sentence is English or is not English to them, but this does not necessarily mean it is good English. What may be true of one person is not necessarily true of anyone else. Even though two people may be native speakers of a language X, their grammars at descriptive levels will not be identical, i.e. both speakers are appealing to a different set of internal rules. How different these rules are can depend on a number of different factors (Davies & Kaplan, 1998, p.42). For example, the grammar of a person who speaks the English of England and the grammar of a person who speaks the English of Scotland will be quite different, as compared to the difference between the grammars of a father and daughter, in which the differences will still exist, however to a lesser degree. Therefore, native speakers of a given language who have been exposed to roughly the same sorts of input during the development process all eventually come to possess essentially similar grammars. This theory could be investigated in a further experiment, in which the nature of a subject’s language acquisition is taken into account. In other words grammaticality judgement tests could be used to analyse families from one part of the world with families from another part who are natives of the same language. The results of such an experiment would be very interesting.

As mentioned, for a sentence to be grammatical, it must follow a certain set of rules that are defined in the language. For example, the grammatical constituents have to be ordered in accordance with a linear order specified in the syntax of the grammar. There is obviously a distinct word order for every language. Studies have shown that if an object and verb are not ad-
adjacent in English, this can lead to an ungrammatical sentence. Generally speaking, the set grammar of a language rarely changes. 'Gling' is a possible word of English which may be introduced at some stage into the English language; however 'gnimp' could never be an English word, unless the grammar of English was to change drastically. The reason for this is that no English syllable has a sequence of a consonant and nasal, where the nasal is not syllabic (Fetzer, 2004, p.21). Therefore grammaticality does not only refer to constructions and sentences, but it also refers to properties of linguistic expression and is thus grounded in the structure of the expression. It is obvious that the factors used in judging grammaticality will differ from sentence to sentence. It may be the case in one sentence that a noun or verb is missing, in another that the word order is incorrect, and so forth.

However, although such factors are needed for a grammatical sentence, this does not necessarily mean the sentence will have to make sense. Although the following famous Chomskyan sentence does not make any sense, it is grammatical; 'Colourless green ideas sleep furiously'. On the other hand 'Furiously sleep ideas green colourless' is not. This is because the first sentence contains an adjectival phrase followed by a verb phrase and if one really tried to create some understanding of it, they could in fact conclude that there is something named green ideas, which is colourless and sleeping in an angry state, whereas the second construction is an impossible English sentence. Constructions which are grammatical but not acceptable are frequently discussed in the literature. Another sentence given by (Chomsky, 1965) is 'The man who the boy who the students recognized pointed out is a friend of mine'. Such a sentence is generally conceived to be sticky and not really possible English. I used such a sentence as one of my items in this study; 'The plane that the pilot that the police questioned flew crashed' and the results generally agreed with what has been said here. The centre-embedding of this sentence clearly caused problems for participants, with the vast majority of binary results judging it to be ungrammatical. A number of people mentioned that the sentence actually annoyed them as they found it to be a bizarre and awful construction, or that it was 'confusing' and 'clumsy', even though they could work out what was being conveyed in the sentence. One person wrote that it would be o.k. if commas were included and another said that it was grammatical but would never use such a construction. This was the common consensus amongst participants, as this construction is generally not used for more than two sentences in the English language. Even though the relative clauses technically match, the sentence
lacks linguistic grace and is thus generally not used in common conversation. All in all, an overwhelming majority of the subjects gave this sentence a very low rating, with only around 7% of participants judging it to be grammatical. This is less than half the amount of that in Rimmers study, in which 15.1% of participants claimed the sentence to be grammatical, however this is still a very low rating and clearly demonstrates the dislike for such a construction.

2.1.2 Degrees of Ungrammaticality

On the surface, grammaticality may appear to entail the dichotomy between grammatical and ungrammatical in that there either exist good sentences and good constructions, which do not contain mistakes, or bad sentences and bad constructions, which contain minimally one mistake. Fetzer (2004) claims that a good sentence is parsed and interpreted completely and bad sentences are only parsed partially and therefore are not assigned a logical form. Many linguists have wanted to maintain the principle that grammaticality is a dichotomous notion, however many others believe that such a concept cannot be defined in such a simple way. Schütze (1996, p.62,63) summarises the views of (Lakoff, 1973) on grammaticality occurring on a continuum, stating that rules of grammar do not simply apply or fail to apply, they apply to a degree, that grammatical elements are not simply members or non-members of grammatical categories, rather they apply to a degree and grammatical constructions are not simply islands or non-islands or environments or non-environments, rather they are each to a degree. According to (Fetzer, 2004, p.48), the majority of grammaticality judgements no longer categorise sentences as grammatical or as ungrammatical, rather, they are categorised as expressing a higher or lower degree of ungrammaticality. Levelt and Neijers (1977, p.89) also agree that there may be different degrees of ungrammaticality. They claim that replacing ‘absolute grammaticality’ with this ‘preservation-of-order’ is advantageous as they conclude that the reliability of absolute grammaticality judgments turns out to be very low. This dissertation investigated the claims made here by having different speakers analyse the same set of items via binary judgements, a 5 point Likert scale and by Magnitude Estimation, which allows for infinite degrees of ungrammaticality and the results can be found under the section entitled Results and Discussion. Note, different degrees of grammaticality do not exist, if a sentence is grammatical, it is grammatical. There may be sentences which are perhaps written in more idiomatic and colourful language, but this does
not make a grammatical sentence more grammatical than a simple one. If a sentence is ungrammatical, it may be slightly ungrammatical, such as containing the wrong verb conjugation, or drastically ungrammatical in which the sentence does not make any sense.

It is common convention to mark ungrammaticality with an asterisk (*) and is the most widely used symbol to indicate the ungrammaticality of a given utterance. This convention is also used to mark unacceptability. However, the question mark (?) marks ungrammaticality to a lesser degree. Although some rule has been violated, it is not bad enough to be marked with an asterisk (Bauer, 2007, p.96). Ross (2004) differentiates six degrees of ungrammaticality symbols, ranging from completely grammatical to completely ungrammatical; O.K, ?,??,?,?,*,**.

2.1.3 Stimulus Factors

Grammaticality itself is a very complex notion and does not have one solitary definition, rather, it can be characterised in a variety of ways. Independent of any theoretical framework, as mentioned, it can be viewed as a measure of whether a sentence is permitted by the rules, patterns and conventions ('grammar') of a language (Murphy & Vogel, 2001, p.5). As well as this however, a number of stimulus factors may influence grammaticality and acceptability. Examples of such factors include meaning, parsability, frequency, morphology and spelling, repetition, context, cloze tests and true/false formats (Schütze, 1996). The cloze test procedure is also widely used to assess proficiency levels of second language learners. In such a test there exist blanks which the reader must fill in (Schütze, 2005, p.474). Fetzer (2004, p.45) refers to (Allen & Seidenberg, 1999, p.125) who also state that the grammaticality of an utterance cannot be defined with respect to the form of that sentence alone, rather it must make reference to the meaning which gave rise to it. Schütze (1996, p.176) also agrees that "on the assumption that one of the first things one does when processing a sentence for judgement is to simply try to understand it, the usual parsing strategies will be involved, and therefore by assumption so will the linguistic competence that they may draw on and the general cognitive resources they may use, with their incumbent limitations." Schütze (1996, p.70) does also agree however that although fully grammatical sentences can be judged without much reference to their meaningfulness, interpretability becomes an important factor in some ungrammatical sentences. In other words, the closer a person comes to
figuring out what an ungrammatical sentence is supposed to mean, the more likely they are to judge it to be acceptable. One of the experimental items in this experiment was ‘the woman sitting next to door’s shoes are like mine’. Although less than 35% of people judged this sentence to be grammatical, people did write that they understood what was being said and that they may indeed utter such a phrase in quick conversation. There are also cases where inserting a word or joining two sentences together changes the meaning of the phrase and may again make the phrase unacceptable. For instance, ‘Yesterday I had a funny dream. I was president and..’ in comparison to *‘Yesterday I had a funny dream and I was president..’, or ‘Can you tell me the time? I have no watch.’ compared with *‘Can you tell me the time and I have no watch.’ (Van Djik, 1977, p.46). The second sentence in both cases either means something different to the first or is awkward and ungrammatical. Interestingly, it would be perfectly acceptable to replace the ‘and’ with ‘because’ in the second example, ‘Can you tell me the time because I have no watch.’ It is these minute differences in language which make it so difficult to analyse. Once more the ungrammaticalness of these sentences is essentially for pragmatic reasons (Van Djik, 1977, p.48).

Other factors which may contribute to the acceptability of a sentence include repetition, morphology and spelling. Repetition of an utterance may deem it more acceptable as the more people hear a form, the more they may like it and morphology and spelling may cause a person to alter their opinion on the acceptability of the word as, for example, spelling ‘night’ as ‘nite’ may consequently interfere with an utterance’s acceptability (Schütze, 1996, p.167).

As mentioned, context is a stimulus factor in judging grammaticality and there are numerous ways that context can influence grammaticality. There have been many incidents where initially one would think a sentence to be ungrammatical, only to have someone suggest a real-world situation where it is plausible. Schütze (1996, p.185) claims that ratings of sentences in context cannot be compared with those made in isolation. If the context of the previous sentence is needed in order to judge a sentence acceptable, naturally this needs to be included in the study. Fetzer (2004) states that native speakers are generally asked to evaluate sentences with regard to minimally three categories; acceptable, ok, not ok, or not sure and context is generally what seems to determine the grammaticality judgements of a sentence. For example the use of personal pronouns at the beginning of a sentence is correct so long as the ‘he or she’ has been defined in a previous sentence. As a further
example, take the following sentences of English:

Q: With what has the postman been murdered? A: John thinks with a knife.

In the course of conversation, answer 'A' is perfectly acceptable, however linking the phrase 'with a knife' to the predicate 'think' would not usually occur alone and may not be deemed acceptable without the question. Therefore, the preceding utterance is needed for the interpretation (Van Djik, 1977, p.44).

Frequently, borderline cases exist in judging the grammar of a sentence. Fetzer (2004) gives the example of 'More doors are higher than windows are wide.' When given this sentence some native speakers will judge it as grammatical and well-formed, whereas others will judge it as ungrammatical and ill-formed. The judgment of a sentence such as this will again be based on context. In this case the sentence will be grammatical if there is a specific set of doors which are being compared to a specific set of windows. Levelt and Neijers (1977) also argue that a subject’s acceptability judgement depends widely on his or her ability to imagine a context for the sentence or phrase they are judging, as does (Heringer, 1970), who compared reactions to sentences which were given with and without context. For example 'John left until 6pm', was not accepted by any of the subjects in the study however when presented with the context, [John left earlier and is going to come back at 6], 15 out of 39 accepted it (Schütze, 1996, p.152).

Schütze (1996, p.155) also refers to the affect neighbouring stimulus sentences may have on the acceptability of a sentence in context. It is thought that borderline grammatical sentences may appear more acceptable when preceded by much worse examples, and less acceptable when preceded by much better ones. This is naturally a problem that is difficult to overcome, and unless it is possible to analyse sentences one by one in isolation, it is a problem which will continue to reside in most grammaticality judgement tests. There are certain measures which could be taken into account in order to try to limit the effects one sentence has on the next. In the present experiment, the items were given to different groups in one given order and to another group in the reverse of that order, for each scale. The results can be seen in the chapter entitled 'Results and Discussion'. Also in an attempt to eliminate the effects of one sentence on the other each question was shown on a separate webpage. In other words, after the participant had answered on one question, they had to click a button to proceed to the next question and they were not allowed to return to the previous question. Still, it would be
almost impossible to eliminate this affect completely from any given study.

2.1.4 Definitions of Grammaticality Judgements

"What sort of process underlies the formation of a grammaticality judgment? The only way to approach this question is to ignore all a priori linguistic restrictions and to regard it as a problem in human information processing" (Levelt & Neijers, 1977).

Grammaticality judgements operate on the performance-competence interface and test whether constituents are grammatical or ungrammatical (Fetzer, 2004, p.33)). Fetzer states that grammaticality judgements are informed by two domains, they are either based on the premise of linguistic competence, i.e. speakers have knowledge of a language and know how to identify grammatical and ungrammatical sentences, or they are confronted with language material in an idealised manner.

Grammaticality judgments provide an alternative path to the grammar. While more factors are involved in grammaticality judgments, they might be less mysterious than those connected to language use. Such judgements have been used to study relationships between memory and language and to examine the processing of sentences by children, bilingual speakers and even aphasics (Smith, 2010, p.2). On this note, many kinds of second language acquisition research have been interested in grammaticality. For example, researchers studying how learners master the article system of English, and what uses of the articles are more difficult than others, will investigate grammaticality, or researchers interested in global proficiency may want to determine the degree to which the learners grammar corresponds to the target language grammar. This relates back to what has been said earlier about individuals acquiring different grammars and the different studies which could be conducted. Even a task such as reading may correlate widely with grammaticality judgements. For example, (Hakes, 1980) found that synonymy judgements were based on superficial form in the youngest children, but on meaning and form together at a later stage. Acceptability for the youngest children was determined by whether or not they understood the sentence. Hence, literacy and/or schooling seem to be a prerequisite for the ability to explain grammaticality judgements, but not for the ability to make them (Schütze, 1996, p.94, 96). Perhaps this should be taken into account when choosing subjects for an experiment in which subjects are invited to explain their answers.
An interesting experimental approach to judging grammaticality was carried out by (?). They examined the effects of distraction on grammaticality judgements by having participants read sentences with various types of auditory distraction. Participants made acceptability judgments on sentences varied by the presence of homophones or orthographically similar words in simple and more complex sentences. In a similar way, Smith (2010) also conducted an experiment in which she investigated the effects of auditory distraction on judging grammaticality. The participants made more errors in more complex sentences but she concluded that the judgement accuracy was not affected (Smith, 2010, p.3).

A line of work that could be considered a model of certain aspects of the grammaticality judgement process is that of Catt. Catt created a computer program for computer-assisted language instruction that was designed to perform automatic error diagnosis and correction of ungrammaticalities produced by second-language learners. The system was above all a model of a foreign language instructor. It could classify errors in a sentence as being due to phrase structure, transformations, morphology, verb subcategorisation, or even certain direct translations from the learners native language (Schütze, 1996, p.172)

2.1.5 Sentence Processing

Yet another aspect which relates to grammaticality is well-formedness. A sentence is ‘well-formed’ if it conforms to the rules of grammar of a given language. Grammaticality and well-formedness together examine linguistic form but they differ when the sentence processing is taken into account. In order to be a well-formed sentence, it must be grammatical and must be processed with a reasonable amount of time and with a reasonable amount of cognitive effort, however a grammatical sentence does not necessarily have to be well-formed (Fetzer, 2004, p.16). Blackwell A. and Fisher. (1996) also state that judging grammaticality is a sentence processing task involving both linguistic and cognitive processing. The participant reads (or listens to) a sentence and judges its acceptability, rendering a decision that relies, in part, upon metalinguistic skills (Smith, 2010). Sentence processing takes place on several levels: immediate, online interpretive processes, and offline, post interpretive processes (Caplan & Waters, 1999). Schütze (1996) also talks about ‘the judgement process’ itself, in which he examines what people might actually be doing when judging the grammaticality of a sentence. He
claims, however, that there exists only one piece of concrete information, the rest is mere speculation. "The only thing we do know is that we do not know what they are doing" (Schütze, 1996, p.81).

Language itself is a complex task. The fact that we can process grammar and other areas of language so quickly is quite remarkable. Dabrowska (2004) uses the example of hearing the sentence 'Tess blew her nose on a paper tissue'. If one hears such a sentence, they quickly and seemingly effortlessly distinguish the meaning, being that a woman, Tess, picks up a tissue, covers her nose, and expels air out through it. In order to arrive at this interpretation, a number of processes had to be performed. For example the words were identified and accessed, grammatical relationships were determined and lexical and syntactic information was processed to lead to the understanding of the sentence. We take this process for granted. The speed at which language is processed is extremely fast. In informal conversation, people produce on average 150 words per minute (Dabrowska, 2004, p.13). In a conversation, when one person has finished speaking, there is about a half a second before the listener begins to speak. Within this tiny interval of 500 milliseconds, the listener has decoded the preceding utterance, decided on their response, constructed their reply and then begins to speak. It really is quite remarkable. Syntactic analysis and the integration of grammatical, semantic and contextual information occurs so fast, that it has been difficult to determine whether they each have processing stages or whether the processor integrates all the information available all at once (Dabrowska, 2004, p.15). The idea that processing occurs quickly can be proven by event-related brain potentials.

**ERP: A method for scientifically analysing grammaticality**

Event-related brain potential (ERP) is a technique used to study the relationship between electrical activity in the brain and language comprehension and has been widely used in assessing grammaticality. However, ERP experiments are costly and relatively difficult to conduct. In order to conduct such an experiment, electrodes are applied to the scalp and this non-invasive method can then provide a direct index of brain activity (Gonzalez-Marquez & Spivey, 2007, p.401).

One of the main components of the ERP is referred to as the N400 component. The N400 components are generated whenever stimulus events involve meaningful processing to the stimuli, and its size is sensitive to fairly subtle
differences in the processing difficulty of the words that elicit it. For this reason the N400 component of the brain waves has been used in many psycholinguistic experiments (Gonzalez-Marquez & Spivey, 2007, p.405). The N400 is a negative-going wave that peaks approximately 400ms after the onset of the stimulus (Gonzalez-Marquez & Spivey, 2007, p.405). The P300 is a positively-going component, which peaks 300ms after the onset of a stimulus; however the N400 is more specific to language, though the P300 has been used in experiments. The N400 is elicited by words in all modalities, whether written, spoken or signal and is also sensitive to contextual factors related to meaning (Gonzalez-Marquez & Spivey, 2007, p.407). N400 amplitudes are large for unexpected items, smaller for words of intermediate cloze probability, and barely detectable for words with high cloze probabilities.

Another ERP component is lexical processing negativity (LPN), which is a brain potential to written words that is most evident over left anterior regions of the scalp. The latency of LPN is sensitive to word frequency and so this component is useful as an indicator that the initial stages of lexical processing have been completed (Gonzalez-Marquez & Spivey, 2007, p.409). The P600 is sensitive to syntactic and morphosyntactic processing. It has been elicited by violations of agreement, phrase structure, and subcategorisation in English, German and Dutch (Gonzalez-Marquez & Spivey, 2007, p.409). The component typically begins at 500ms post-stimulus and peaks at 600ms and its scalp distribution tends to be posterior. Coulson and Kutataz (1998) cited (Gonzalez-Marquez & Spivey, 2007, p.410) carried out an experiment and found that the P600 varied with the probability of ungrammatical trials within an experimental block. One half of their study entailed 80% grammatical sentences and 20% ungrammatical and the opposite way around in the other half. The grammaticality effect was much smaller when ungrammatical sentences were probable than when they were improbable. The P600 tool is widely used for testing when speakers find sentences to be ungrammatical (Gonzalez-Marquez & Spivey, 2007, p.410).

ERPs can be reliably measured using electroencephalograph (EEG), a model sensitive to electromagnetic activity in the brain, which can monitor how the brain changes with manipulations of particular linguistic representations (Gonzalez-Marquez & Spivey, 2007, p.401).

Schütze (1996, p.59) notes that in time, ERPs will eventually be used to filter out various sources of judgements of ungrammaticality. As mentioned, however, there are restraints with conducting ERP experiments. Firstly the subject must remain relatively still during the recording of the EEG, as the
signal in the ERPs is very small, the bioamplifiers are extremely sensitive. The subject must also try to refrain from blinking or moving their eyes during critical points in the experiment as the eye also has electrical properties and blinking can produce electrical activity that is detected by the EEG electrodes. The larynx also has electrical properties which are detectable by the electrodes and so it is not practical to record ERPs to the production of overt speech (Gonzalez-Marquez & Spivey, 2007, p.417). It is important to note however that although this method of analysing grammar is more scientific and concrete than simply asking a person to rate a sentence, the method is quite expensive and would naturally be slower than other methods, which evaluate a vast number of people in a shorter time frame. It would therefore be difficult to make mass judgements about the general population based on a random select number of people. Still, ERPs are one of the only concrete methods available which allow us to really understand what happens when a grammaticality judgement is carried out.

2.1.6 Conclusion

Therefore, the capacity to make grammaticality judgements emerges out of the ability to process language normally. It has been shown that the process of judging a sentence as being grammatical or ungrammatical is linked to sentence processing and comprehension, and to the interpretation of a sentence in context. One important property of the task is that different criteria may be used, depending on the properties of the sentences being judged (Allen & Seidenberg, 1999, p.121). In other words, the judgement of a sentence differs from sentence to sentence, as mentioned above.

In summary, judgement-giving is a skill, and judgements themselves are a kind of performance data. The ability to judge grammaticality is not a logically necessary part of grammatical competence, and even native speakers differ greatly in their skill in giving consistent judgements, especially in subtle cases which will be analysed later on in this dissertation.
Chapter 3

Acceptability

“Let us use the term ‘acceptable’ to refer to utterances that are perfectly natural and immediately comprehensible without paper-and-pencil analysis, and in no way bizarre or outlandish. Obviously, acceptability will be a matter of degree along with various dimensions.” (Chomsky, 1965, p.10, 11)

3.1 Introduction

At the beginning of the previous chapter the question was put forward as to what exactly grammaticality is. A natural progression is to now define what exactly the notion of acceptability entails. In effect, this chapter is a mirror of the previous chapter, except for the fact that it commences with an account of the various literatures on acceptability, its consequent definitions, and the various ways in which acceptability judgements are useful. The present chapter then continues to describe further differences between grammaticality and acceptability which have not already been mentioned, as well as undertaking the problem as to why sentences are judged grammatical or acceptable.

3.1.1 What is Acceptability?

Recall from the last chapter that (Chomsky, 1965) states that whether a sentence is grammatical or not is a question of competence. In a similar manner, he also states that whether a sentence is acceptable or not is a question of performance and there are three ways which can demonstrate
this. Firstly, any particular instance of a speaker accepting or rejecting a sentence is an act of performance, as is any sort of generalisation across any such instances. As well as that, a judgement of one’s disposition towards accepting or rejecting a sentence is itself a type of performance (Schütze, 1996, p.26).

Acceptability does not refer to the grammatical unit of a sentence; rather, it refers to the production and interpretation of an utterance in its linguistic and social contexts. It is important to clarify this information at the beginning of any work regarding acceptability. Acceptability is an empirical concept and the acceptability of an utterance is evaluated with regard to the nature of the connectedness between its linguistic form, sequential position and social context (Fetzer, 2004, p.19). In other words, an acceptable sentence is consciously accepted by the speaker as part of his or her language upon hearing it (Schütze, 1996, p.20). Bever and Carroll (1981, p.229) also state that the processes of perception must be involved to some extent in rendering acceptability intuitions, as a sentence must be apprehended in some sense in order to be judged acceptable.

However, one of the main problems with defining acceptability is deciding whether ‘to accept’ should in fact be regarded as an act. The act of accepting something is an intentional decision, for example accepting a gift implies the individual has the option of accepting or rejecting it. Therefore, the object in question must satisfy a number of properties which are judged satisfactory to the individual (Van Djik, 1977, p.41). As regards linguistics, this definition of acceptability holds, however there are a few discrepancies. The individual must have the competence to judge whether certain properties, e.g. grammatical properties, are satisfactory, yet when a person is conversing with another they do not appear to be ‘doing’ anything. Note, stating that grammatical properties are satisfactory does not necessarily mean the sentence is fully grammatical. Van Djik (1977, p.43) states the possibility that the user automatically puts the interpreted words or phrases in the appropriate logical categories of a meaning representation and this is what the user ‘does’, but they occur automatically in his mind rather than causing them intentionally. This statement agrees with Schütze cited above, who claims that judging acceptability is a conscious act. It is indeed true that when a person is judging acceptability they are relying on their intuition of that language, but the act of consulting your intuition is not a conscious act, sometimes one cannot recall why they know a given fact, they simply do. On the other hand, one could say that when a person reacts to a given
unacceptable sentence, this is a conscious act as something has triggered the attention of the person and they are then consciously focusing in on what acquired their attention in the first place, and making conscious judgements about the grammatical nature of the sentence.

There are many different factors which interact to determine acceptability. Referring back to the Chomskyan example in the previous chapter, ‘Colourless green ideas sleep furiously’, Chomsky (1965) states that an unacceptable grammatical sentence such as this is unacceptable, not for reasons of grammar, but because such a sentence would be difficult to remember or simply for stylistic reasons. Stylistic factors such as alliteration may make a sentence easier to process and hence may lead to an acceptable judgment of the phrase. Rimmer (2006, p.247) gives the examples;

? In prison she got lucky and left well alone.
In prison she got lucky and to have her own cell.

The first sentence contains two complements to the verb ‘got’, which is rarely seen as grammatical in English. Though the second sentence also contains the same sort of construction, the sentence does not appear to be as acceptable as the first. Rimmer speculates that the first sentence may be more acceptable for reasons of alliteration among the letter ‘l’, and is perhaps ‘clever language play’, whereas the second sentence is a mistake which is commonly made by language learners.

The term acceptability may also refer to the social status of the language as a whole, especially in relation to second language acquisition. Afolayan (1977, p.14) discusses the political, educational, economic and social acceptability of English in Nigeria. The English language has gradually been accepted as the nations legislature, judiciary and administration, because it is the only language neutral to all ethnic groups in the country. The syntax of English in Nigeria is generally under the same concept of grammaticality as may be found in the World Standard English, unlike in phonology, where there is a large difference between what is regarded as grammatical and what is acceptable. In Nigerian English it is quite acceptable to say ‘sleep on bed’ or ‘shoot him on the head’, i.e. the use of the preposition ‘on’ instead of ‘in’ in phrases (Afolayan, 1977, p.14). Today, there are many varieties of English as a lingua franca, and in some countries, students are actually being taught bad grammar when compared to Standard English, as it is what they accept as their lingua franca in that country.

Although new rules and phrases are generally created in this regard because of cultural differences, this does not mean that the language as a whole
does not change and evolve every day, even in countries where Standard English is spoken. Neologisms exist in every language and are important to mention under acceptability. The definition of neologisms is that they are novel lexemes that were not observed before in a language. Put in simpler terms, they are made up words that are acceptable. For example ‘skyped’ would be a neologism. Obviously before skype existed, the verb ‘to skype’ did not exist. It is quite easy to make up words that are acceptable to a native speaker, so long as they conform to the phonology, morphology and rules of the English, German or whatever the language in question is. This is another reason it is so difficult to analyse language and grammar, as new words are created every day. The neologism assimilation scenario describes what the speakers know about how neologisms are treated in their language.

In order to guess whether a made up verb like 'spling' is regular or irregular, people have to assess its similarity to existing verbs (Schütze, 2005, p.462). The wug test is an experimental method and is used for eliciting inflected forms of novel verbs and nouns (Schütze, 2005). For example;

"This is a wug. Now there are two of them. There are two wugs."

As it is possible that neologisms may simply appear one day in a given language, it is clear that human language is very creative. As mentioned, new sentences are formed and interpreted every day. There are a potentially infinite number of sentences for any given language. Hence, in support of the above argument, the native speaker is able to form acceptable sentences every moment of every day and must have the competence to evaluate and interpret a new sentence when confronted with one. It has been said that native speakers have at their disposal an internal system of rules and principles of principles which enable them to produce well-formed sentences, grammatical sentences and to determine the acceptability of a sentence (Haegeman & Gueron, 1994) cited (Fetzer, 2004, p.14). On the other hand, (Rabin, 1977) argues that in the language of Hebrew, new rules are constantly being formulated which leads to native speakers, even those who are linguistically educated, having doubts about the language when they write it, consequently leading to native speakers being reluctant to make acceptability judgments. This is a very interesting point for the topic at hand which seemingly suggests that the more complicated a language is, the more difficult it is for the persons internal system to differentiate between acceptable and unacceptable sentences and in some cases may not be able to give a judgement of a given utterance. Although it may be true that it is more difficult to formulate such judgements, it should not be impossible. If we were not designed with the
ability to differentiate new linguistic forms and words, language would not evolve and that in itself is an unrealistic notion. It appears that the evidence for an underlying built-in grammatical competence within a speaker seems to outweigh that of opposing arguments. In short, acceptable words or sentences do not need to be grammatical, or even exist in a language; they simply need to be comprehensible. Individuals often make acceptability judgments about utterances, deriving conclusions about social group, primary language, intelligence, and so forth, based upon reflective judgments of structural output (Schütze, 1996).

3.1.2 Judging Acceptability

Chomsky (1975, p.8) states that "sentences are acceptable (or perhaps acceptable under particular circumstances) if they are suitable, appropriate, adequate to the purpose at hand, etc." Acceptability judgements are used to formulate some sort of insight into everyday language, which may not necessarily be considered grammatical in written language. "Acceptability judgments allow us to examine sentences that rarely occur in spontaneous speech or corpora." (Schütze, 1996). The differences between acceptability and grammaticality can be viewed in the next section, but therein lies one of the main differences. Grammar books rarely give acceptable forms of a sentence if it is in fact ungrammatical, even though it could be completely acceptable in everyday English. In the area of second language acquisition, citeA[p.286]Odl94 concludes that advanced learners of a second language should in fact receive detailed information about acceptability. Although intermediate and advanced learners will still produce ungrammatical sentences, they should be aware of well-formed but unacceptable sentences. However, it would be difficult to teach such a notion to learners of a language, as if they are only taught sentences which are grammatical, the probability that they will make bad mistakes lessens and therefore their language improves. This would also be an area worth investigating further.

The best way to analyse acceptability judgements would be to record speech occurring in everyday life. However, Schütze (1996) cites a problem with using recordings to analyse acceptable language in that in naturally occurring speech data, it is difficult to distinguish errors (slips of the tongue, unfinished utterances, etc.) from grammatical production. Though it is true that such data would be difficult to analyse, the problem with written acceptability judgement tests is that although people may actually say such
an utterance, they may correct it if it is written in front of them. In the present experiment, the sentence 'I bought three mouses at the computer store', was put to the subjects. On one of the binary scales 50% of people judged it to be acceptable, and each other test had a lower percentage of people rating it grammatical, in comparison to the study by Rimmer, where 56.6% of teachers accepted the form. When subjects were asked to input why they chose such a response, some obviously complained about the plural form, whereas others stated they did not know what the correct form was, or were doubtful when it came to this sort of situation. Some claimed that if the word computer preceded the word mouse it would be acceptable, whereas others simply said they found nothing wrong with the sentence. Stating that you are buying three mice at a computer store does almost imply that you are actually buying the rodent in the shop, which is obviously a strange occurrence for a computer store. Although subjects stated their dislike for this sentence on paper, as (Rimmer, 2006) points out, it does not mean that they would not intuitively say 'mouses' when asking the shop assistant in a computer store. Therefore, there does appear to be a fine margin between the use of mouses and mice in this type of plural. Thus it appears that if judging acceptability is the experiment at hand, verbal contact would be the best strategy, however this forms another problem of practicality.

3.1.3 Acceptability and Unacceptability in Dialects

To speak the same dialect (or language) as someone else is to have the same intuitions about some set of sentences as (s) he has. However, Ross claims that the notion of a dialect is far more abstract. He claims that instead of studying a dialect, one should study idiolects, in as detailed a way as possible. That is, they should try to delve as deeply as possible into the structure of each speakers intuitions (Ross, 2004, p.389, 390). Acceptability naturally differs between dialects. Fetzer (2004) gives the example of a sentence, which is perfectly acceptable in British and Irish urban dialects; 'I saw them big spiders'. Depending on location, the use of 'them' as a demonstrative pronoun is not only acceptable, but also grammatical, in spite of its status as an ungrammatical sentence in English-English, because it has become so widely used. Initially I had thought of entering such a sentence as one of the ungrammatical control items in the present survey, however, if the sentence was entered as an ungrammatical control, those who actually conform to this type of grammar may indeed claim it to be grammatical, however it is more
likely that when written on paper it would be judged ungrammatical. This relates to the sentences containing 'mouses' and 'the woman sitting next to the doors shoes', though such sentences may be uttered, it is unclear as to whether this would be put on paper.

Murphy (2007, p.45) gives another example of acceptable sentences which may be acceptable in certain dialects. Murphy refers to (Jackendoff, 1972) who states a syntactic restriction that the antecedent must appear in subject position in English, and the reflexive pronoun in a post-verb position, hence the unacceptability in most English dialects of the following examples.

John shaved himself
Himself shaved John

There are of course scenarios where '*Himself shaved John' could be acceptable, under certain conditions. For example the person whom 'himself' refers to may be standing beside the speaker and the speaker could be pointing to him.

3.1.4 Differences between Acceptability and Grammaticality

The following section deals with the differences between acceptability and grammaticality. Although similar in nature, there are distinct differences between the two. Most of what has been said so far has dealt with acceptability and grammaticality in isolation, which in effect displays the differences between them; however there still remain a few points to be mentioned regarding their differences.

As mentioned, the general view is that grammaticality is a question of competence, whereas acceptability is a question of performance, however Chomsky (1965, p.11) never claimed that degrees of grammaticality would correspond to degrees of acceptability; he in fact states that they do not coincide. Chomsky also defined a distinction between competence and performance, stating that competence is the speaker-hearers knowledge of his language, and performance is the actual use of language in concrete situations. A grammatical sentence is different from a well-formed sentence and from an acceptable sentence. As mentioned, a well-formed sentence does not only have to be grammatical, but it also needs to be easy to process, whereas an acceptable sentence does not have to be grammatical; it must merely be comprehensible and easy to process (Fetzer, 2004, p.13). It is fre-
QUENTLY NOTED IN THE LITERATURE THAT ACCEPTABLE SENTENCES MAY NOT NECESSARILY BE GRAMMATICAL. INTERESTINGLY, BICKERTON (1977, P.29) SUGGESTS THAT LINGUISTIC NORMS DO NOT EVEN EXIST IN EARLY PIDGIN LEARNERS, AND THEREFORE EVERYTHING IS ACCEPTABLE AND NOTHING IS GRAMMATICAL AT THAT STAGE.

OPINION DIFFERS ON THE NATURE OF GRAMMATICALITY AND ACCEPTABILITY. LAKOFF (1977, P.73) CLAIMS THAT GRAMMATICALITY MAY IMPLY A BINARY DECISION ON WHAT IS OR IS NOT GRAMMATICAL AND THAT ACCEPTABILITY DOES NOT IMPLY THIS SAME DIFFERENCE. HE CLAIMS THAT LANGUAGE DOES NOT HAVE TO BE ACCEPTABLE OR UNACCEPTABLE; HENCE THERE IS NO BINARY DISTINCTION BETWEEN THE TWO. A CONTINUUM FROM MOST ACCEPTABLE TO LEAST MAY BE NOTED, OR INDEED ACCEPTABILITY MAY DEPEND ON FACTORS SUCH AS SOCIAL STATUS. LAKOFF (1977, P.74) ALSO STATES THAT GRAMMATICALITY JUDGMENTS ARE DETERMINABLE BY PURELY LINGUISTIC CRITERIA, WHEREAS ACCEPTABILITY JUDGMENTS TOUCH ON AREAS SUCH AS PSYCHOLOGY AND SOCIOLOGY. NATURALLY THIS IS A VALID POINT. IT IS FAR MORE DIFFICULT TO CONSTRUCT CONCRETE ANALYSED RESULTS ON ACCEPTABILITY THAN ON GRAMMATICALITY. BOTH REPRESENTATIONS ARE BASED ON TACIT KNOWLEDGE AND AS A RESULT WHAT MAY BE TRUE OF ONE PERSON MAY NOT BE TRUE OF ANYONE ELSE. AT LEAST THERE ARE GRAMMAR BOOKS AND KNOWN RULES FOR GRAMMAR, THERE ARE NO SUCH CONCRETE NOTIONS ABOUT ACCEPTABILITY. AS MENTIONED IN THE PREVIOUS SECTION, NATIVES, OR EVEN NON-NATIVES FOR THAT MATTER, ARE NOT TAUGHT ABOUT ACCEPTABILITY OF FORMS, THEY SIMPLY KNOW WHETHER A SENTENCE IS ACCEPTABLE OR NOT, WHICH IS OUSTAINLY A MATTER OF PSYCHOLOGY MORE SO THAN SCIENTIFIC LINGUISTIC TRUTH.

3.1.5 Why is a sentence judged to be ungrammatical or unacceptable?

AS WELL AS ANALYSING THE DIFFERENT METHODS AVAILABLE FOR JUDGING GRAMMATICALITY AND ACCEPTABILITY, THE PRESENT STUDY Sought TO EXAMINE THE DIFFERENT REASONS AS TO WHY ONE JUDGES A SENTENCE THE WAY THEY DO. A FORTUNATE FACT IS THAT SOME NATIVE SPEAKERS OF A LANGUAGE OFTEN HAVE THE ABILITY TO SAY, DIRECTLY, WHETHER A SENTENCE IS POSSIBLE, AMBIGUOUS, POSSIBLE UNDER A CERTAIN INTERPRETATION, AND SO ON. IN THE PRESENT EXPERIMENT, PARTICIPANTS WERE ASKED TO TRY TO GIVE REASONS FOR THEIR RESPONSES. LEVELT AND NEIJERS (1977, P.89) CLAIM THAT IF A SUBJECT IS ASKED HOW HE OR SHE PERFORMS THE JUDGMENT TASK, A USUAL ANSWER GIVEN WOULD BE SOMETHING LIKE "I TRY TO IMAGINE A SITUATION IN WHICH THE PHRASE OR SENTENCE CAN BE SAID. THEREFORE THE SUBJECT APPEARS TO BE USING IMAGERY IN ORDER TO JUDGE THE GRAMMATICALITY OR ACCEPTABILITY OF A QUESTION AND
is hence a performance factor which may affect intuitions of grammaticality and acceptability. This ability to imagine a situation where the sentence could be plausible is also related to the context of the sentence, which has been discussed in the previous chapter.

Bialstok and Ryan (1985) proposed that grammaticality tasks can be ordered by increasing the amount of analysed knowledge required, for example grammaticality judgements, locating ungrammaticality, correcting ungrammaticality, explaining ungrammaticality or stating a rule that is violated. Schütze looks at some methods researchers used to induce subjects to express their feelings on the grammaticality of sentences. Schütze believes that subjects should be asked why they judged a sentence the way they did. For instance if a sentence was judged as a bad sentence, they could be asked why they feel this way and this helps in creating reliable results. Other methods include requesting rank orderings of sentences by grammaticality, asking for a comparison of the type of violation in bad sentences or indeed the use of ambiguity. If a sentence exists that is O.K. under one reading but perhaps not in another, one could ask the subjects if it was ambiguous (Schütze, 1996, p.55, 56, 57). Fetzer (2004, p.42) also claims that native speakers are often able to provide more precise information about what is grammatically incorrect in a sentence than simply saying it is ungrammatical. They can state what must be modified in order to assign the sentence the status of being grammatical. A general question was asked in this experiment requesting why the speaker chose the response they did. There were a few linguistically tuned responses, however most people replied in the way Fetzer states above; by simply providing what they believed to be a correct grammatical version of the sentence in question.

However, like almost all fields of linguistics, there are opposing arguments to asking participants such questions. Some linguists feel that people may not actually know the reasons for their answers and asking them to justify their answers is pointless. I agree with both sides of this argument. It is important to request the reasons one gave the response they did in order to eliminate the possibility that they simply gave a response without putting any thought into the reasons behind their thinking. However, if the subjects participating are non-linguists, it is indeed possible that they may not be able to put their reasons into words. Perhaps it simply sounds correct to them and they cannot contribute any further detail than that. An answer given in this study for one of the questions actually was 'I would say ”themselves”. Not sure why though...’ Therefore, in my opinion if the question is to be
asked, the response field should not be mandatory and written in an open ended text field.

How the questions were to be phrased was a major question at the beginning of this experiment. A further analysis regarding the control items will be talked about later, but for the ‘why’ response, initially I had considered offering a further list of possibilities for the subjects to chose from, in order to entice more people to answer that question, however in the end, the subjects were not given a list of possible answers to choose from. The initial method I considered would have been modelled on an experiment carried out by (Svatvik & Wright, 1977, p.183, 84), who conducted an experiment whereby they asked participants to rate a set of utterances which were rated with ‘yes’ (I would say this), ‘? (I might say this but probably in a different way) and ‘no’ (I would not say this). A rationalization test was then used on the answers which were rated ‘?’ or no’, in order to figure out why they rated them as such, and so asked them to choose from;

- ‘Old-fashioned’: Thou art foolish.
- ‘Ungrammatical’: I is English.
- ‘American’: He sure can run.
- ‘Never heard anything like it’: Time is walking slowly.

Naturally different possible responses would have been used in this experiment, however it was felt that supplying possible answers may steer the participants response to an answer they may not have come up with themselves, hence, not reflecting their true feelings on the subject. Though this method may have made computability easier to handle, in the end it was decided that an open ended text field would be the most consistent option and would offer the most valuable information.

3.1.6 Conclusion

This chapter concerned itself mainly with the altering descriptions regarding the nature of the notion of acceptability present in everyday contexts. In summary, acceptability is far more complex than it may initially appear. It is almost impossible to analyse fully, as new words are formed and accepted day after day. This chapter also provided a further account of the different forms of language which are acceptable in different dialects, as well as a brief
summary of the remaining differences between grammaticality and acceptability and a brief discussion as to why one may judge a sentence they way they do. Although it is a difficult area to conduct research in, the area of cognitive research on grammaticality and acceptability is a very interesting area and the present study tried to formulate some evidence for the best experimental methods to analyse grammaticality and acceptability. The remaining chapters in this dissertation concern themselves mainly with the experiment at hand and the varying results obtained in the analysis.
Chapter 4

Experiment Background

4.1 Introduction

This chapter aims to introduce the experimental background for the study at hand. The different scales used in the experiment and the reasons such scales were chosen are discussed in the first few sections, followed by the reasons the experiment was carried out online, what factors were taken into account and how the surveys were created. The experimental tool used for distributing each survey was Limesurvey and an analysis of this tool is also present in this chapter.

4.1.1 Types of Scales

The process of assigning numbers to attributes or characteristics according to a set of rules is called measurement. There are four main measurement scales, Nominal, Ordinal, Interval and Ratio. The nominal scale is used to classify different kinds of data (Gonzalez-Marquez & Spivey, 2007, p.94). Items categorised on a nominal scale can be categorised but not ordered. A binary judgement of grammar, i.e. whether a sentence is grammatical or ungrammatical, is in many ways a nominal scale (Bard & Sorace, 1996, p.39). Once order is in the equation, the scale becomes ordinal and this type of scale is widely used in surveys. An example of such a scale is that of the ranked Likert scale. Interval measurements have a fundamental property of preserving differences at equal intervals, as well as reflecting differences of degree, and finally ratio measurement gives the possibility of comparing two observations not only in terms of their difference, but also in terms of ratios.
(Gonzalez-Marquez & Spivey, 2007, p.95, 96). The technique of magnitude estimation is a method which borders interval and ratio scale data and will be explained in more detail in this chapter. The three techniques mentioned, binary judgment, use of the likert scale and magnitude estimation are the three techniques under investigation in this dissertation.

**The Binary Scale**

The binary scale does not demand as much cognitive thinking as the other scales to follow. The task does not involve a judgement relative to another sentence, nor does it involve more than two possible answers, it merely entails a yes/no type question, which intuitively is rather simple.

**The Likert Scale**

Graded acceptability and grammaticality may offer a clearer view of what the subject would actually like to say. Caroll (1979) suggests that graded acceptability can result from a discrete grammar plus performance rules of some sort. The Likert Scale is a common scale used for such judgements. The Likert scale was developed by the American psychologist Rensis Likert in 1932 and is defined as “a psychometric response scale primarily used in questionnaires to obtain participant’s preferences or degree of agreement with a statement or set of statements.” (Bertram, 2011)

There are two characteristics of the Likert scale. Firstly, it must have an equal number of positive and negative descriptors and also equal appearing intervals between the rating points. It cannot however be assumed that the difference between adjacent data will be perceived as equal, i.e. there is no way to ensure the perceived difference between 'agree' and 'strongly agree' is the same as that between 'agree' and to 'neutral', or whichever responses are being used in the experiment (Bertram, 2011) It has been stated (Chynoweth, 2003), that when creating surveys there must not be an unbalanced or unequal number of responses available. This view has been widely argued as regards the Likert scale. A summary of such arguments is to follow. Chynoweth does state that the amount of 'burden' inflicted on the subject must be kept to a minimal, which is an obvious step to take into account. In other words the survey must be kept as simple and user friendly as possible.

When using n-point scales for judging grammaticality the polar ends of
the scale should reflect both maximal grammaticality and maximal ungrammaticality, in order to represent the entire range of grammaticality (Bader & Häussler, 2008, p.6) Weskott and Fanselow (2009, p.232) argue that using n-point scales may lead to unwanted influences from the surrounding items in the course of the experiment. In other words if a subject marked a sentence a certain value, is the probability of rating the next sentence with the same value lower as a result? This argues in favour of binary judgments, however it is widely accepted that binary judgements do not allow for other ranks which the subject may wish to convey. Snow and Meijer (1977) simply asked subjects to rank the acceptability of six sentences in order and suggest that ranking may be preferable because subjects can distinguish more levels of acceptability.

As mentioned, opinions differ on the amount of levels which should be differentiated on a Likert scale. A five point Likert scale may offer more variability between groups than a three point, which is important for statistically significant results, however in everyday usage, people generally either agree, disagree or sit on the fence when asked to rate something. Is it therefore worthwhile having a five point scale if the answer the respondent wishes to express is going to be either end of the scale? Nonetheless, the 5 or 7 point Likert scale is generally used in statistical tests and occasionally people use even numbered tests, to force people to make a decision.

In a paper by (Infosurv, 2010), the advantages of 5 point and 6 point Likert scales are addressed. Support for the former includes the fact that participants may in fact be neutral on the topic and forcing them to make a decision leads to bias in the results. Allowing for mid-point response also allows for comparisons which are statistically significant, i.e. above and below neutral. On the other hand forcing the respondent to take a positive or negative side is seen as an advantage for the supporters of the 6 point scale, in the sense that it the results obtained will give clear cut answers. There are clearly advantages and disadvantages of such scales, however allowing for a neutral, or indeed 'get out clause' as it may be viewed as, may entice more people to complete the experiment as if participants must chose a response before they can proceed, they may lose interest in completing the survey.

A further study by (Garland, 1991) investigated the inclusion and exclusion of having a neutral point on a Likert scale. They conclude that the desire to please the researcher is minimised when the mid-point is not present. They also conclude opposing results to (MS & Jacoby, 1972) however, who found that as the number of points on the scale increased, the number of
mid-point answers decreased. Therefore it appears that the way in which people respond to a scale without a mid-point is content specific.

A famous theory proposed by (Miller, 1956) claimed that the human mind is capable of distinguishing about seven items, (plus or minus two). As regards to response scales, if this claim is adhered to, rating scales with more than about seven possibilities are a waste of time and cannot be comprehended by the human mind. However, there are many experiments which disregard this claim. (MS & Jacoby, 1971) claim that as few as two response possibilities may even be adequate. They claim that using two or three point scales would not decrease the reliability or validity of the study. A further remarkable experiment carried out by (Preston & Colman, 2000) tested the reliability of scales ranging from two to eleven points scales, and one 101 point scale. Their results show that the highest reliability factor came from the 7-10 point scales, with a slight decline in those with 11 and 101 response categories and the lowest reliability was obtained from the 2, 3 and 4 point scales. (p.5) also analysed the respondents preferences for each scale. For ease of use and speed, 101 and 11 were judged the worst, with the best for ease of use being five, seven and 10, and two, three and four came out on top for speed. Expressing feeling was also analysed and, as assumed, two, three and four point scales were judged the worst and 10, 11 and 101 the best. (?, p.8). They conclude that different scales may be needed for different purposes. If it is speed and ease of use the researcher is after, lower point scales are better suited for this type of research, however increased validity would require larger scales. Therefore it is up to the researcher to perform a trade-off between respondent’s preferences and what it is they intend to measure. The present study decided on the use of a five point scale, as it was felt that five appeared to be the average amount used in experiments and offered enough variability for the study at hand. The main purpose of the experiment was also to compare scales for their ability to differentiate different degrees of ungrammaticality and so any lower than five would have been verging on the binary scale, which was of course, already being analysed.

**Magnitude Estimation**

"Whatever subjects do when magnitude-estimating linguistic acceptability, and however odd they find the whole process at first, they clearly have the ability in their psychological repertoire, just as they have the ability to give proportionate judgments of brightness or prestige." (Bard & Sorace,
Magnitude estimation was originally developed to provide better than ordinal scales for measuring impressions of physical stimuli in the psychophysics domain (Stevens, 1975). As mentioned, Magnitude estimation does not restrict the number of degrees which can be differentiated by the subject, it is open ended. The method as regards linguistics involves rating a given sentence with a value greater than, less than or equal to a reference sentence. For example, if the experimental item sounds twice as good as the reference item, subjects should multiply the value they gave the reference sentence by two; if they feel it is twice as bad, they divide by two, etc. (Fukuda Shin & Beecher, 2010, p.1). The instructions given to participants can be found in the appendix. In some experiments (e.g. (Bard & Sorace, 1996), (Weskott & Fanselow, 2009), (Bader & Häusser, 2008), and (Sprouse, 2008)) subjects were told to rate the lengths of lines on a page with respect to each other, prior to commencing the study. They were then shown an example of the technique in use with sentences and finally invited to make their judgements on the sentences to be studied. The present study however did not begin in the same way as it was felt that the line examples were not needed. (Murphy, 2007, p.93) also carried out this line experiment with his practise participants and concluded that the exercise was not needed and was simply more of a hassle for the subjects to carry out, as the pilot respondents stated that such an act made the instructions "longer, less coherent, and did not aid in their understanding of the method". In the present study participants were merely given a number of examples which conveyed how to use magnitude estimation on sentences, as it was felt that it was more in keeping with what was being asked of them in the study. Subjects then continued from the example to the experimental items. "Unlike other scales used to analyse grammaticality and acceptability, magnitude estimation allows us to treat linguistic acceptability as a continuum and measures differences between items." (Keller, 2003, p.1).

Differences between Binary, Likert and Magnitude Estimation scales

Naturally, there are advantages and disadvantages associated with each scale. The binary judgement method is easy for participants to understand, however the results are thought of as being relatively coarse-grained. In comparison, the Likert scale allows for finer-grained results while still being easily understandable, yet it is unclear as to whether the chosen scale will allow for as many distinctions or levels as the subject wishes to display. Bard and Sorace
(1996, p.25) also state that a graded five point scale limits the respondents answer, as it is impossible to know in advance whether or not their answer lies within this five point scale. Magnitude Estimation has widely been regarded as the best judgment scale to use for acceptability judgements, however it is by far the most difficult scale for subjects to understand. If a scale is too difficult to use, subjects may become demotivated and the quality of their responses may decrease. Still, Magnitude Estimation overcomes the disadvantage of the Likert scale in that subjects chose the number they wish to assign to a given rank and so the scale is therefore open ended (Fukuda Shin & Beecher, 2010, p.1).

(Weskott & Fanselow, 2009) also investigated the differences between using binary, Likert or n-point scales and Magnitude Estimation. They argue that the binary judgment task may be the most natural judgement as it the closest to our natural way of metalinguistic judgement in everyday situations. They do agree, however, that the use of n-point scales is arguably better in acceptability judgements, as the differences between two sentences may differ from that between two different sentences. They state that the main difference between the three types of measures lies with their respective variability. Binary judgments only have two outcomes, in comparison to n-point scales which generally have five or seven points and thus a larger degree of variability. Finally ME judgments have no restriction on the possible values used to express the acceptability judgment. (Weskott & Fanselow, 2009, p.243) conclude that they did not find much variation overall between each of the three measures. They claim that binary and, in their case, 7-point judgment scales are as informative as magnitude estimation. Fukuda Shin and Beecher (2010) also conclude from their experiment that all three methods capture the same results, in the majority of cases.

4.1.2 The Experiment

This section introduces the study itself and what factors were taken into account when creating the questionnaires. A major part of any experiment is to decide what items to test. There have been a number of articles and books written in this area and the present experiment is based on an article by Wayne Rimmer, ‘Grammaticality Judgment Tests: Trial by Error’. The 10 sentences in his study were reused, and control sentences were added in as they were nonexistent in his work. Rimmer (2006) conducted his experiment with English teachers allowing only for binary judgements on the sentences,
whereas the present study used magnitude estimation and the likert scale, on
the sentences with the general population, as well as two binary scale judg-
ments differing in the way the question is put to the participant. The items
used in Rimmer’s study appealed to me as in some cases, the ungrammat-
icality of the item was quite subtle and the sentences appeared to be more
interesting than certain other items tested in articles. In Rimmer’s conclu-
sion, he does suggest that the experiment could be enhanced by requesting
the subjects reasons for their answers and perhaps conducting the experi-
ment with the general public instead of language teachers. I agree with this
opinion and so such factors were taken into account in my study.

Factors (Age, Gender, Native Language)

Individual differences such as gender, age and education should be taken
into account on a personal questionnaire so that variability attributable to
them can be examined in the analysis (Schütze, 1996, p.187). However, in the
present study education was not taken into account, but native language was.
It was felt that this factor would be more important to note than education,
as the majority of participants were friends of mine on the social network
Facebook, and the majority of that group did not have any background in
linguistics. It was felt that factors such as age and native language were
more important factors to take into account. Age may make a difference
to a study as what is accepted by one generation may not necessarily be
accepted by another, in the same way native language may have an effect
on results as if learners of a given language are not taught about acceptable
forms that are perhaps ungrammatical this may hinder the results. There is
ongoing research into the validity of second language learners’ responses on a
grammaticality judgement test (see e.g. (?), (Ellis, 1991)). Grammaticality
judgement tests delve into the persons underlying syntactic competence. The
problem with using non-native speakers on a grammaticality judgement test
is that “in the case of second language [grammaticality] judgements, one is
asking learners to make judgements about the language being learned at a
stage in which their knowledge of that system is incomplete” (? , p.305). In
the end, the number of non-natives was minimal in this study (n=5). In total
a survey was sent to 11 non-natives, however only five replied. In the present
case, the results were not omitted from the analysis as the non-native English
speakers did not appear to hinder the results. All participants appeared to
have good English and their results seemed on par with the native responses.
A further discussion of the results can be seen in the following chapter.

It has also been noted that handedness may be an important factor to take into account when conducting an experiment about language. Most right-handed people process language with the left side of their brain while left-handed people process language with a mixture of left and right, sometimes just one or the other but mainly both. Therefore handedness does seem to give us an indirect insight into the workings of the brain and the organisation of language in the brain. One cannot deny that there is a correlation between handedness, as it has been noted that there is a correlation between language difficulties such as dyslexia, and left-handedness and left-handed people are thought to be more accepting of weirdness than right handed people (Schütze, 1996). However, because left-handed people may be right or left handed, or both, judging results on handedness may not be as evident as it may initially appear. For this reason handedness was not taken into account in this study.

Gender, however, was taken into account as it has been documented that there may be differences between female and male speech. Lakoff (1977, p.77) gives a list of what he concluded to be the language of ‘women’.

- Special Vocabulary women seem to discriminate linguistically among colours with more precision than men do.
- Use of adjectives that express feelings, e.g. charming, adorable, divine etc.
- Use of empty intensifiers like so, such.
- Greater use of standard ‘correct’ forms, avoidance of slang and neologisms, both lexically and grammatically. Lakoff found that even in nursery school, young girls adhere to correct grammar more so than boys. For example they drop the ‘gs’ in endings (runnin’) much less than boys and use less double negation (I didnt do nothin’) and fewer substandard forms such as ‘aint’.
- Use lexical, grammatical or phonological devices to suggest hesitancy or deference, e.g. silences or interjections like ah, um etc.
- Greater variation in pitch and intonation.

However, the ratio of male participants to female participants who completed a survey in the current experiment was 38:45 and there did not appear
to be any difference between the results given by male and the results given by female participants. Further results regarding gender, and as mentioned age and native language can be seen in the results section.

Phrasing the Question

Phrasing the question was one of the most difficult parts in creating the surveys. When conducting psycholinguistic experiments it is important that the participants do not know what exactly is being analysed, as this may affect the results. Participants may answer in a way they feel mirrors what they should answer and this may not represent their actual thoughts on the topic. A method which has been used, and which exists in the present study, is phrasing the question as regards foreign speakers of English, i.e. asking the participant if they would correct an English language learner had they uttered such a phrase. A complete account of the instructions for this experiment can be found in the appendix. As the experiment was to analyse different methods for analysing grammar, this information was not made available to the participants. They were simply told it was an experiment regarding grammaticality judgements and they also did not know that more than one survey was being circulated. The test items which were used in the present study can be seen here. The grammatical control items were taken from a paper by (Bard & Sorace, 1996).

List of questions

1. Who did you quit college because you hated? (Rimmer, 2006)
2. She arent care about me.
3. Either you or I are wrong. (Rimmer, 2006)
5. John angered while Susan amused the woman. (Rimmer, 2006)
7. What did you bring that book to be read out of for to?
8. The plane that the pilot that the police questioned flew crashed. (Rimmer, 2006)
9. John was bought the book. (Rimmer, 2006)
11. John announced a plan to steal Bills car late tomorrow. (Bard & Sorace, 1996)
12. The woman sitting next to the doors shoes are like mine. (Rimmer, 2006)
13. You should lay down on the bed. (Rimmer, 2006)
15. John teached me how tie my shoes.
16. I bought three mouses at the computer store. (Rimmer, 2006)
17. Theres only one person who thinks of themself in that light. (Rimmer, 2006)
18. That is the sort of up with which will not put I.
19. Which man did Bill go to Rome to visit (Bard & Sorace, 1996)
20. Susan trained like shed never done before. (Rimmer, 2006)

The reference sentence for the Magnitude Estimation study is as follows; When do you know the man whom Mary invited?

As mentioned, an additional question requesting the knowledge as to why the participant judged a sentence the way they did followed each question. The question put forward for the reason for their choice was the following:

If you are able to state the reason (s) as to why you chose the response you did, please state it in the box below, if not please progress to the next question.

Selecting Materials and Participants

"Acceptability judgments no matter who makes them require meta-linguistic sophistication and recruit a wide variety of cognitive capacities (including perhaps folk beliefs about language and internalized prescriptions) whose
CHAPTER 4. EXPERIMENT BACKGROUND

relation to the language faculty remains obscure.” (Culbertson & Gros, 2009, p.723).

When deciding what materials and participants to use in a study one should keep several things in mind. Statistical methods demand that the data at hand is a random and representative (i.e. unbiased) sample of the phenomenon of interest, in order to be able to reliably estimate the strength of any conclusions drawn (Murphy, 2007, p.61). Some classic studies suggest that native speakers, who are left without proper guidance, may stray far from what we intend to be asking them to do. The authors of such studies concluded that very little can be assumed in advance about responses to language. (Schütze, 2005, p.457). There were a number of tests done on morphological complexity by (Hay, 2001). Her experiment dealt mainly with the complexity of words and wanted the subjects to decide which word was more complex when presented with pairs of words and to rate them on a scale. Schütze criticises this experiment saying that in order for the participants to complete the task correctly it is assumed that they have previous knowledge or intuition of what a complex word is as one would not be able to remember all of the information given to them in the beginning, as the instructions given to candidates were highly detailed. This suggests that the instructions given to candidates should be kept to a minimum. Murphy and Vogel (2001) support this claim by stating that if instructions are not kept down to a reasonable length, it may become an essay on linguistics that only a sophisticated informant can understand. Therefore relatively simple guidelines can be followed to improve the robustness of evidence gained. These include providing clear illustrative examples in the instructions which span the range of acceptability that will be encountered in the experiment and using filler sentences and randomisation of the order of utterances to factor out effects of habituation, strategising, and varying attention over the course of the elicitation exercise. (Murphy and Vogel, 2011, p.9) The present experiment provided brief examples to candidates and this information appeared to provide the subject with the information needed to complete the task. As mentioned, filler sentences were also used, however in the current study, an attempt to randomise the data failed as the tool used did not allow for random generations of the sentence groups. In order to overcome this, the order of the question groups in each of the four surveys was reversed, creating a further four surveys. Consequently the total number of surveys analysed in this experiment came to eight.

Culbertson and Gros (2009) investigated who the best people to carry out
linguistic grammaticality judgements are. They focus on work by (Devitt, 2010), who claims that linguists themselves are the best subjects, however many linguists have argued against this case. Culbertson and Gros (2009) conclude that the distinguishing factors between linguists and non-linguists is not what one should be investigating in such a study, rather they should look at the differences between subjects with and without task specific knowledge. In their study, there were no clear advantages between those with linguistic backgrounds and those with minimal knowledge in the area. Using linguists as your subjects does not, however, yield satisfactory results if it is the initial reaction of the general population one wishes to investigate. Haegeman and Gueron (1994, p.8) claim that 'The native speaker who judges a sentence cannot decide whether it is grammatical. He only has intuitions about acceptability'. However, if the grammar of non-linguists coincides with the grammar of linguists, which may indeed by the case, this claim must be false. Murphy and Vogel (2001, p.7) agree that non-expert informants supply more naturalistic, theory-neutral responses, i.e. the spontaneous, gut-reaction of an average speaker.

This is on par with the views of (Schütze, 1996), who claims that the subjects being used in an experiment must have no prior linguistic training, as if it is the competence of normal native speakers to be investigated, random samples of normal native speakers needs to be studied. This is almost never done by theoretical linguists, according to Schütze. Speed is also of the essence for many psycholinguistic experiments, as much research has shown that our memory for the form of an utterance decays extremely quickly as compared to our memory for its content (Schütze, 1996, p.86).

As mentioned, the experimental items used in this experiment were taken from the article listed above. In his experiment, Rimmer used teachers as his subjects, however, the use of teachers judgements verges on that of using linguists. Everyday learners expect that if they query their teacher about some grammatical phenomenon, that the teacher will provide an accurate response or judgement. As teachers are generally trained in noticing incorrect grammar, they may also be prone to judging an ordinary sentence unacceptable or ungrammatical (Odlin, 1994, p.279). As the experimental items were taken from this study, it aided in a comparison between teachers and non-linguists, i.e. a random sample of the population, as the same items were used and the binary scale judgements could be compared to that of Rimmers study. It also appears that much of the literature indicated non linguists would be the best participants, and thus my friends list on Facebook appeared to be
an acceptable group of subjects to analyse.

After it had been decided what group of subjects to analyse, the surveys had be distributed evenly. Participants were recruited by an n-point sequence extraction from my friends list on the social network Facebook. Put in simpler terms, the friends list was divided into 8 lists and a different survey was sent to each one, there were consequently around 40 people in each list. Posters were also displayed around the University of Dublin, Trinity College, with information about each experiment in order to entice more people to participate (see appendix). There were eight different posters, each containing the same one link on the page, in order to disallow repeated participation. Fukuda Shin and Beecher (2010) raise the point that using the same participants across each test may lead to problems such as the possible effects of one experiment on the other, or changes in judgments after repeated exposure to the items. For this reason, the present online survey had a cookie in place which disallowed repeated participation. This was done in order to generate a more significant result. As mentioned, Schütze (1996) claims that participants should be given examples and trial runs of how to carry out the test and the test data should contain roughly equal numbers of grammatical and ungrammatical sentences. The instructions for candidates can also be seen in the appendix for this experiment. Examples of the methodology of each scale was given in each case and the test data contained the 10 items under investigation, along with 10 control items, five of which were grammatical and five of which were drastically ungrammatical. Again, trial runs were not carried out in this case, as it was thought that the examples would provide enough information for participants to know what to do. Trial runs would also have elongated the study and there would have been a higher probability that participants would opt out of the survey before even starting the test items. On this note, the subjects did not have to give their names in any part of the survey and their results were kept confidential and anonymous, so as to encourage more people to complete the surveys. Allowing the identity of the participants to be known may hinder the results as the subjects may be more inclined to answer how they think they should answer rather than what they would actually say or believe to be grammatical or acceptable. The candidates were ensured that there were no right or wrong answers and that it was their reactions and reasons as to why they deemed a certain sentence more grammatical than another that was of interest. Obviously everybody thinks differently so their interpretation of where they think a given sentence lies on a scale is very interesting.
Benefits of Online Experiments

As mentioned, the present study was carried out online, but what are the benefits of online experiments? Murphy (2007, p.64) cites a review by (Wygant & Lindorf, 1999) who investigated the differences between postal and online surveys. They conclude that an unpaid survey had a response rate of 50% online, with 80% of the results being returned within two days, whereas the same survey conducted by post had a lower response rate (32%) and the responses took 22 days to arrive. Hence, it seems online surveys are much more practical than postal ones. However, the experiment at hand had a significantly lower rate of responses than 50%, with an estimated average of 25% of participants completing the survey assigned to them. Taking into account those who entered the survey but existed with completing it, the average partake was 32%. The values are estimates as there was no way of knowing how many people read about the survey from a poster and completed it. Perhaps more people would have taken part had there been a small reward or token of gratitude for doing so, or even the possibility of winning a reward by entering into a draw at the end of the survey. Be that as it may, I do believe that I would have received fewer replies by post than online. C. and Stewart (2000) also describe the benefits of using computers to conduct experiments. The main points being that:

- They provide extended access to participants
- Reduce costs
- Make large amounts of data easy to handle
- Provide a comfortable environment for participants.

As well as that, participation is more likely via the web as opposed to face-to-face communication, especially when sensitive accounts are being asked of the participants. It was felt that distributing the experiments online would reach a larger sum of people than simply handing out surveys and hoping that people would return them. It is also practical to use online experimentation, if open ended text fields exist for people to write in, as it may be difficult to decipher what exactly the participant was trying to convey if their handwriting is difficult to read. Such open ended fields existed in this survey and thus the online method was chosen.
Limesurvey: Online Experimentation Tool

Limesurvey is a free, open source and easy to use tool available for download online. Initially, Limesurvey seemed like a foolproof tool to use, however, it did not come without its problems and after some time, my opinion of the online survey tool began to alter as it caused problems before, during, and after experimentation. Limesurvey requires a database and a server to run off, which did cause a number of problems in the set-up within college because of certain college policies with using the undergraduate servers. After this problem had been rectified and Limesurvey was set up and installed, the work began. The tool is very user friendly and there is an online forum full of frequently asked questions should the need to view them arise, as well as an online user manual. However, in spite of the ease of use, many problems arose. Firstly, after I had created all eight surveys, they began to mix together. Generally the problem remained within the same survey, except in one case where subjects were answering on the likert scale and a 'yes/no' option appeared. The magnitude estimation surveys became completely jumbled, instead of saying 'only numbers may be entered in this field' after the numerical input for the subjects answer in each case, units of measurement appeared such as 'kWh' and 'Kg'. Other problems came in the form of a duplicated 'why' question instead of the question on the sentence followed by the 'why' question, or indeed in some case no question at all appeared from that question group. Where a subject was asked to input their gender, the answer field was appearing as numerical input, and again the question regarding the subjects age was followed by random units of measurement. The problem only seemed to emerge when the survey was run, as within Limesurvey the questions appeared to be fine. In an attempt to overcome this strange obstacle, I had to delete the question group in which the problem was occurring, in the case of magnitude estimation that meant deleting every question, and redoing it. This seemed to solve the problem and after I had run through each survey several times to check there were no such problems, I activated them and sent the link in a mail over Facebook. People then started to inform me that the text field after the 'why' question would only allow them to input up to 5 characters in some cases. I figured this question would be poorly answered regardless but the fact that this was occurring resulted in fewer responses in this field. The text field for inputting your native language also stopped working after the first day on one of the binary scales and, as mentioned, one of the likert scale responses became of
the 'yes/no' variety. There also appeared to be a problem with one of the magnitude estimation questions, as no one answered on it, but when I ran through the survey, it seemed fine.

A consent form was also a requirement for carrying out an online survey for this dissertation. There are a number of limitations of Limesurvey, one being that the binary answer of yes or no could either entail a further 'no answer' response or not, however this had to be set in the settings and could not be customised to individual questions. As a 'no answer' response was needed for the experimental items, this option had to be set. This in turn created a problem with the consent page, as the participants response needed to be 'yes' or 'no'. In order to overcome this set back, eight html pages were created, each with a link to the subsequent experiments. Participants were told to click the link entitled [name of survey] if they consented to participation. The html pages are all identical except for the different paths to the survey and the different names for the surveys at the bottom of the page. Names used include 'Survey, Judgement, Experiment, Exploration, Questionnaire, Study, Responses and Research', and would appear on the other pages. As mentioned, subjects should not know what is being tested when carrying out an experiment. It was felt that these titles did not indicate the process at hand, that being a comparison of judgement scales. Again, the code for one of the html pages and how it looked on screen can be seen in the appendix as a screen shot and also as plain text. The information sheet which appeared on the first page of the experiment can also be viewed in the appendix. The links to each of the html pages are as follows:

The problems which occurred afterwards can be seen in the next chapter as it relates to R, the statistical platform which was used. In short, the data did not export how it was supposed to and a file compatible with R had to be manually created in Microsoft Excel. Initially this was one of the advantages enticing me to use Limesurvey, as the data could be used in R, however it became a hindrance by the time the data had been collected. However, in spite of the above mentioned problems, Limesurvey does have its advantages. Statistical tests can be carried out within the tool, which was practical for receiving a quick overview of the results, however the pie charts and graphs created within Limesurvey cannot be exported to other applications, only the raw data may be exported to applications such as R and SPSS but as mentioned, this did not go according to plan in the present case. One of the main reasons I chose to use Limesurvey initially was for presentation. There are a number of different templates saved within the tool and I felt that it was an advantage not to have a plain black and grey background, in order to entice more people to complete a survey. As I also mentioned earlier, limesurvey allows you to set a cookie which disallows repeated participation, which is indeed a plus in any experiment. There was also the possibility of using different question groups, so that the question would appear on different pages after the participant had submitted their answer and there was no limit to the amount of questions you could input.

4.1.3 Conclusion

Creating the experiment proved more difficult and required more time than I had initially thought. This unexpectedly set the entire process back a few weeks and I do not think I will ever use Limesurvey again. Nonetheless, one cannot complain too much about a free online experimentation tool and once the initial problems had been rectified, the tool did appear to work well most of the time. This concludes this section regarding the background to the experiment. The three types of scales used have been discussed, as well as the factors taken into account when creating a survey, my motivation for using online experiments and the problems encountered with using Limesurvey.
Chapter 5

Results and Discussion

Without performing statistical tests of significance, we cannot know whether trends in our data are likely due to chance or to actual facts about grammars [...] The more levels of grammaticality we try to distinguish, the less unanimity we find, and the more we need to rely on statistics (Schütze, 1996, p.195).

5.1 Introduction

The following section produces the results of the present experiment and discusses their implications for the question at hand. This section begins with a number of problems which resulted from using Limesurvey to export the results to the statistical platform R, followed by a description of how the data was manipulated in order to be computable. The chapter then continues to discuss the results as regards their significance in the area, as well as a note on why subjects judged the items the way they did. The results of the grammatical and ungrammatical control items is also put forward in this chapter. To begin with, the following is a list of the main research questions set out in this dissertation;

- Does the order in which the questions are asked make a difference to the results?
- Does the wording of the question make a difference to the results?
- Are the results for Magnitude Estimation different from that of the Likert, and Binary?
• Is there a correlation between gender or age and grammaticality?
• Is there a difference between natives and non-natives?

5.1.1 Statistical Background

Statistics, in particular inferential statistics, is a method of observation and data gathering used in the areas of behaviour, performance, and mental activity. Linguistics and psychology differ in what they constitute as empirical evidence for falsifying or supporting a given theoretical claim. As regards linguistics, if one counter example of a proposed theoretical statement can be found, this falsifies the statement and it must be consequently reviewed and changed. However in psychology, the elements for falsifying or supporting a statement are composed of empirical data observed at the level of peoples behaviour or performance (Gonzalez-Marquez & Spivey, 2007, p.89). Statistics has two main areas, descriptive and inferential statistics. Descriptive statistics deals with organising and summarising information about a collection of actual empirical observations. These tools serve to describe actual observed data, whereas inferential statistics provides tools for generalising conclusions beyond actual observations. The idea is to draw inferences about an entire population from the observed sample. Therefore the goal of inferential statistics is not to provide absolute certainty, but to provide robust tools to evaluate the likelihood of the generalisations you would like to make (Gonzalez-Marquez & Spivey, 2007, p.91). Statistical analyses are generally conceived to be performed on collections of numerical data, where numbers are supposed to represent the properties or attributes under investigation.

When properties or attributes can have different values, e.g. ‘degree of ungrammaticality’, they are technically called variables. Variables are at the core process of investigating cause-effect relationships (Gonzalez-Marquez & Spivey, 2007). Variables are important for correlation studies. There are specific statistical tools that can be used to establish precisely the strength and degree of relationship between variables. The statistical tests used in this study will be discussed in more detail later on, but in summary the main test used for establishing the strength of correlation between two tests was the Pearson test for correlation and the test used for establishing the level of significance was the Mann-Whitney U test. The statistical analysis was carried out in the statistical software program R. The book used for help with R was ‘The R book’ by (Crawley, 2007). As mentioned, Limesurvey offers
the possibility of exporting the data obtained as .csv and .R files, however, there were a number of problems with the data which was exported to R. One of the prerequisites of R is that the data entered into the system needs to be of the form of tables or aligned columns if using Microsoft Excel, in order to be analysed. However, this was not the case of the data in the present case. When loaded into R, the columns were not aligned under their subsequent question headings and as there had been an open ended text field for participants to input their responses as to why they chose the response they did, this data had been split into a separated string words, one word under each column. This obviously caused problems for the analysis as the correct answer was not under the correct question. There was also no way of simply deleting the answer questions from this part of the data either, as the number of the responses had become mixed up between some words. As well as this, when the data was imported into Excel, inverted commas appeared around the number which had been given as a response, the data was spread out over the entire sheet and where the participant did not give any response for a given question, this was recorded by two commas. Hence the problem was the same in the .csv file and the R file and the only way to rectify the problem was to manually delete every occurrence of a word or letter and relocate the numbers to their own individual cell of the Excel sheet, under the correct question number and personal data. This took hours to complete and consequently pushed the analysis of the data back, as, before the data could be properly analysed, certain other measures had to be carried out first.

5.1.2 Working with the Data

To begin with, the Magnitude Estimation data could not be computed and compared to the other scales, or indeed within itself, as the given responses for each question depended on what value the participant gave to the reference sentence. In order to comprehend the data in a more meaningful way, it was normalized with respect to the grammatical controls, as it was thought that they would score a higher value in most cases than the reference sentence. Therefore the average of the grammatical control responses was divided into the probe answers in each case. For example, in one case a participant gave the grammatical controls the values of, 1000,200,1000,1000,1000. These were then averaged to produce the value of 840, which was divided into each response, that participant gave. The inverse log of this value was then com-
puted as it had been discussed (Murphy, 2007) that Magnitude Estimation is a logarithmic scale. This also made the scale comparable to the other scales, as in the .csv file, answers on the binary scale which were judged grammatical were given a score of 1, and those which were ungrammatical a score of 2, as for the 1-5 Likert scale, 1 also meant grammatical and 5 ungrammatical, and then after this analysis had been done on the Magnitude Estimation scale, the values ranged from -2 indicating grammaticality and +2 indicating ungrammaticality. Hence, each scale was comparable as the higher the number, the more ungrammatical the sentence. As well as having to analyse the Magnitude Estimation data before computation began, the order of the question had to be taken into account. Each scale had two types of survey, an ordered survey, and a survey which was in reverse order to its counterpart. Though I use the word ordered, this does not mean that the questions were ordered grammatical and ungrammatical, as the question distribution was random, with grammatical control items mixed in with the ungrammatical controls and test items, rather, that the order in which the sentences appeared was one experiment and the reverse of that ’order’ was the other. However, for ease of reading and computation, these types of scales will henceforth be known as ordered and reversed. Naturally, the answer for a given question on both the ordered and the reversed scale had to coincide, i.e. the answers given by the participant had to be representative of the question they actually gave. When the data was inputted into the Excel file, naturally, the question order of 1-20 was inserted for each scale, however this meant that question one on the ordered scale was in the same column of question 20 on the reverse scale. Consequently, both answers did not refer to the same item. In order to overcome this, Excel offers a transpose function and this was used on the reverse scales. The data was firstly transposed, then ordered in descending order with respect to the question and finally transposed back onto itself, resulting in the question which appeared in position 20 in the reverse experiments, now being in position one and representing the same question as the other values in the same column. The data was then ready to be analysed.

5.1.3 Analysis

This brings me to the first research question listed above; does the order in which the question is asked make a difference to the results? Each scale was analysed in order to answer this question. As mentioned, there were
eight questionnaires in total sent out in this Experiment. In order to test if there was a significant difference between the order in which the questions were presented to participants, the Mann-Whitney U test was used to test for rank significance between the two scales and a further test was done by means of Pearson correlation to test the strength of correlation between the two. If the outcome happened to circulate an insignificant p-value and a positive correlation, it would suggest strong evidence that ordering does not matter, and that is what the outcome was in this case. The level of significance was set at 0.05 for each experiment, i.e. a value greater than 0.05 would indicate that there was not a significant difference between the two conditions. How the analysis was carried out in R can be seen in the following paragraph. It may be worth mentioning that when one wishes to compute the Mann-Whitney U statistic in R, the wilcox.test function is used, however 'paired=FALSE,correct=FALSE' must be inputted as an attribute in order to stop the Wilcoxon test from being carried out. The Mann-Whitney test is the unrelated design counterpart of the Wilcoxon test. Though both are used to test for rank summation, the Wilcoxon test is used when the same subjects are tested over two conditions, whereas the Mann-Whitney is used when different subjects are being analysed, which was the case in this experiment (Greene & d’Oliveria, 1999, p.51). The data entered in each case is representative of the average results obtained overall for each question in that data set. In the case of Magnitude Estimation, the data set was the resulting average of the normalized values, and their subsequent inverted logs as described above. The results of the ordered and reversed binary scales in which yes indicated grammatical can be seen here. 

```r
> data_pos<-c(2,2,1.0833333333,1.0833333333,1.6153846154,1.0833333333,2,2,1.8181818182,1.9166666667,1.1,1.7,1.3636363636,1.0909090909,1.9,1.6363636364,1.0909090909,2,1.3636363636,1.0909090909)
> data_pos_r<-c(1.9,2,1.2,1.4,1,2,1.9,1.7,2,1.1,1.5555555556,1.2,1.2,1.8,1.7,1.4,2,1.4,1.3)
> wilcox.test(data_pos,data_pos_r,paired=FALSE,correct=FALSE)
```

Wilcoxon rank sum test

data:  data_pos and data_pos_r
W = 197.5, p-value = 0.9457
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(data_pos, data_pos_r, paired = FALSE, correct = FALSE):
cannot compute exact p-value with ties

> cor.test(data_pos,data_pos_r)

    Pearson's product-moment correlation

data:  data_pos and data_pos_r
t = 12.0119, df = 18, p-value = 4.967e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
   0.8586863 0.9775448
sample estimates:
cor
      0.9429129

As the above result for Mann-Whitney contains a p-value of 0.9566 and there is a positive strong correlation of 0.9429129, it is quite evident that there is no significant difference between the scales in this case. Although it may not be surprising that a significant difference did not exist as regards ordering, it is nonetheless important to take such a factor into account when carrying out experiments. As many factors as possible should be taken into account when conducting an experiment. When sentence position is taken into account, it is easier to form concrete conclusions about the area of interest, as it in turn alleviates some of the doubt which may be associated with such a test. If two different groups of participants, who are subsections of the general population agree on the status of a sentence regardless of when the item appears on screen, or on paper to them, it provides further evidence that the data obtained is not done so by mere chance and is thus more representative of the general consensus of the sentence or utterance.

The above tests were also carried out on the other binary scales in which yes indicated ungrammaticality, the two likert scales and both of the Magnitude Estimation scales. As mentioned earlier, there appeared to be a problem with question three in one of the Magnitude Estimation scales, and so the answers to this question were removed from the set of figures in both ME sets of data. The results of each of these tests are summarized in the table on the following page.

As you can see, each scale results in an insignificant p-value and a strong positive correlation in each case. Therefore, it appears that the order in which the questions were asked did not make a significant difference in any of the scales examined in this dissertation. This resulted in the data for
Table 5.1: Table of Order relations

<table>
<thead>
<tr>
<th></th>
<th>Mann-Whitney U</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary: Yes = Grammatical</td>
<td>0.9457</td>
<td>0.9429129</td>
</tr>
<tr>
<td>Binary:Yes = Grammatical</td>
<td>0.3078</td>
<td>0.9204709</td>
</tr>
<tr>
<td>Likert Scale</td>
<td>0.6262</td>
<td>0.9186828</td>
</tr>
<tr>
<td>Magnitude Estimation</td>
<td>0.4523</td>
<td>0.9265716</td>
</tr>
</tbody>
</table>

each scale being bulked together as one in each case, consequently creating four sets of data for further analysis, irrespective of the order in which the question was asked.

The next research question to answer was then that of whether the way in which the question was asked made a difference to the results. The analysis was carried out using the same tests to test for significance and correlation and it was carried out on the new set of binary data, which was independent of order. In other words the average figures used in this analysis were that of the two binary scales where yes meant grammatical and the other two where yes meant ungrammatical. The tests were carried out on the binary data as it was only that scale which was under investigation for this part of the experiment. The results showed that the way in which the question was worded is irrelevant. The p-value in this case was well over 0.05 at 0.8074 and there was again a positive correlation between the two, at 0.942421. Tables 1, 2 and 3 in the appendix display the aggregate values for both of the binary scales and the Likert scale results for each individual question. Though there are a few divergences, it is clear that as a whole the question order, and the phrasal of the question did not make a difference, which is backed up by the significant results given in this section. Hence, all four binary data surveys could be analysed as a whole, creating one large binary file, alongside the Likert and ME scales, which were also irrespective of order. Consequently, there were then three scales to compare overall.

The overall comparison of the three scales, binary, likert and Magnitude Estimation, is the main question under investigation in this dissertation. It is widely assumed that magnitude estimation is the best test to use in conducting such research, as the participant may distinguish as many degrees of ungrammaticality as they wish, however, on inspection of the bare data received for ME, it appears that only two people out of 16 differentiated 10 degrees, i.e. they gave their reference sentences a value of 100 and 200
respectively and rated some questions with 1000 and 2000 respectively. This is obviously one of the advantages of the ME scale, however, besides these individuals, the third highest differentiated degree was five. This indicates that the vast majority of those who judged the scales of ME (87.5%), only needed a five point scale, i.e. that of the Likert scale. Therefore, is the hassle of Magnitude Estimation worth the trouble if participants do not make use of the entire scale? The following table shows the statistical results, again the Mann-Whitney U and Pearson correlation tests, for the three tests. It was not possible to compute all three at once, as the test is limited to two values, hence ME was analysed against the Likert, the Likert also against the Binary scale and to complete the loop, ME was examined against the binary scale. Though the scales were comparable, a further effort was made to ensure that the rank sum was tested the individual section available for that scale. In order to do so, the linear vectors of data were subtracted from each other and the resulting mean value of this data was then set as an attribute within the test. As an example, the following R code displays exactly how this was carried out in the Mann-Whitney test and the code for the correlation test is also present.

```r
> lik_magnitude <- likert - mag
> summary(lik_magnitude)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.  
   1.391  1.759  2.066  2.152  2.603  2.995

> wilcox.test(likert, mag, mu=2.152, paired=FALSE, correct=FALSE)

Wilcoxon rank sum test

data: likert and mag
W = 190, p-value = 0.7994
alternative hypothesis: true location shift is not equal to 2.152

> cor.test(likert, mag)

Pearson's product-moment correlation

data: likert and mag
t = 15.0896, df = 18, p-value = 1.167e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.9061925 0.9854074
sample estimates:
cor
```
From the above information, it appears that there is no statistical difference between the Likert scale and the Magnitude Estimation scale, as the p-value is 0.7994 and there is a positive correlation. It is a similar story for the Likert and binary scales, with a p-value of 0.8498 and a positive correlation of 0.9654007, as well as for the binary and Magnitude Estimation scale, there is a positive correlation between the two, 0.9466252 and an insignificant p-value; 0.8924.

A further test for significance which could be carried out in order to double check the results given here, is the commonly used t-test.

> t.test(binary, mag, mu=0.815)

Welch Two Sample t-test

data: binary and mag

t = 2e-04, df = 29.706, p-value = 0.9998
alternative hypothesis: true difference in means is not equal to 0.815
95 percent confidence interval:
0.4761018 1.1539766
sample estimates:
mean of x mean of y
1.5398315  0.7247923

Again, there is no significant difference between the binary and ME scale, and neither is there a difference between the binary and Likert, (p-value = 0.9985) or the Likert and ME scale (p-value = 0.9987).

Therefore it appears that there is no significant difference between using Magnitude Estimation, the Likert scale or a binary scale. Naturally, this is an interesting result to obtain from this data, as Magnitude Estimation has been seen as the best test to use in acceptability judgements. This study concludes that there is no significant difference and if that is the case, it appears that requesting subjects to rate scales using Magnitude Estimation may not be worth the hassle after all. The data is more difficult to analyse than the other two scales and thus it appears there is no outstanding advantage of using Magnitude Estimation. Indeed, this also draws the question of whether the Likert scale is needed. The present study did not find any difference between it and the other two scales and really, if the binary scale is as affective as any other scale and it mirrors our natural way of thinking, maybe it is the best test after all.
The fourth and fifth research questions stated above questions whether factors such as age, gender and native language make a difference to the judgements people make on certain items. In order to investigate this claim, all of the data was converted into an identical scale, of ones and twos. This was done as there was a problem with the magnitude estimation results when they were loaded into R in a text file. As the results had been obtained by dividing and retrieving the log value within Excel, there was a problem with transferring the single values on their own out of that excel sheet. The values were inserted into Emacs and then loaded into another Excel sheet and this appeared to work on the outside, however an error message appeared in each cell within R. Therefore converting everything to one and two seemed like the other reasonable option for comparing age and gender. Naturally the binary scale was already in such a format, as for the Likert scale, everything from 1-2.5 was converted to 1 (grammatical), in this case all of the ones and twos in each response, and anything between 2.5 and 5 was converted to 2 (ungrammatical), in this case 3, 4 and 5. As regards Magnitude Estimation, all those values which were below zero were regarded as 1, and those above as 2. Tables 4, 5 and 6 in the appendix contain pie charts of the distribution of age categories, male and female participants and differing native language participants respectively. The figures were automatically rounded up within R and hence the summed figures on the pie chart are slightly over or slightly under 100% by 1% as a result of this automatic rounding. Nonetheless, the distributions can be seen clearly. As you can see, the distribution of age is by far dominated by the 18-24 age bracket at 82%. This was expected as most of my friends on Facebook would fall into this category and the posters put up around college may indeed have been picked up by lecturers, but also students, more than likely falling into this category. As such, the results given relating to age cannot be seen as a representative sample of the general population, however the results are inserted for completeness sake. The same can be said for the distribution of natives and non-natives in the study. 79% were native English speakers, 2% German and 2% Irish, 1% Latvian and 15% did not give their language, though this may be because in one of the likert scales the option for inputting your native language only seemed to work randomly. In bare figures, there were only 5 non-native English speakers from a sample of 88. Again, this is not surprising as most of the people who received the survey were native speakers, therefore the data is again present for completeness sake. The gender distribution was almost 50:50, with slightly more females (52%) than males (44%) taking
part in experiment, 5% did not give their gender. In order to run a test for significance and analysis of deviance of the rating given in relation to age and the question, the relation of gender to those same factors and finally the relation native language had on such factors, a quasipoisson generalized linear model was used. As mentioned, the data used in this case is of the form of one and two; however that was not the only manipulation that was needed to carry out this test. The data had to be transposed so that the question name ran as a column instead of a row, as well as an individual id for each set of data had to be inserted. Instead of having one person’s answers running horizontally, they had to be transposed vertically and each item in a cell relating to this person was given the same id. This Excel sheet was then saved as a .csv file and imported into R.

The results are not surprising. Naturally, the question made a difference to the rating. For example questions thirteen, nine and three each yielded significant results, with question 13 being highly significant, amongst others, whilst other questions, e.g. Question 5 did not yield a significant result. Having said that, only seven questions out of 20 did not yield a significant result.

```
question[T.q_13]   -0.5962967  0.1024188  -5.822  7.01e-09  ***
question[T.q_3]    -0.3378877  0.1147504  -2.945  0.00328   **
question[T.q_9]    -0.2364496  0.0978092  -2.417  0.01574   *
question[T.q_5]    -0.1194231  0.1023372  -1.167  0.24340
```

As regards age, the significance levels are a lot less than that of the question:rating ratio. Only question 13, 17 and 20 yielded significant results.

```
question[T.q_13]:age  0.0095128  0.0037600   2.530   0.01150   *
question[T.q_17]:age  0.0088437  0.0037355   2.367   0.01803   *
question[T.q_20]:age  0.0089821  0.0038031   2.362   0.01831   *
```

Again, it would be quite astonishing if age was an indicator of the type of rating one question received over the other, although as mentioned in the previous chapter, perhaps if the age differences had been more even, one could perhaps analyse certain teenage sayings for acceptability amongst adults and teenage/young adults. Perhaps this would yield more significant results.

There were no significant results for gender. As the distribution I obtained for male vs. female was relatively equal, it is possible to conclude that gender does not appear to make a difference, under the present circumstances.
For completeness sake I will include the results for native language, though it is important to again point out that I cannot conclude from the following results that native language had an effect as the samples were not evenly distributed by any means. Naturally, one might expect a difference between the ratings of native and non-natives. In this case, there were significant deviances for question 11, rated by a German, question 6, again rated by a German, question 16 was just above the significance bracket at 0.065, and their native tongue was, interestingly, Irish. Question four yielded a significance of 0.087 and question 5 of 0.083727, again borderline cases rated by a Latvian and native English speaker respectively. The results obtained here were representative of every participant over each scale. The following graphs seek to display in visual form the average ratings received for each question and where they reside on the scale of grammatical and ungrammatical with respect to gender and language and the second graph includes age also.

It was also possible to obtain information on the separate scales in relation to age. The boxplot in Figure 5.1 on the next page indicates that the highest average ratings of four and five fell in and around the 21 year old mark.

### 5.1.4 Grammatical Controls

The following section delves into the reasons as to why one judged a sentence the way they did, with comparisons to the article by Wayne Rimmer. Rimmer (2006) only used the binary scale in his study and as a result only the binary
scale results are comparable in this case. Naturally, the information given in the following text does not supply scientific facts or significance results, as the participants were merely asked to state why they chose the response they did and thus this section is looked on as more of an aid to the statistical results recorded above. The section is split into a discussion of the grammatical and ungrammatical controls, and the test items.

The grammatical control items were taken from a paper by (Bard & Sorace, 1996), who copied their sentences from (Haegeman, 1991). Overall, the controls appeared to carry out their task at receiving a grammatical status, however, having said that, no one grammatical control received a 100% grammatical rating across each test. 'Who did John invite' was the closest with only two people giving it a rating of '2' on likert scale, i.e. almost

Figure 5.1: Age corresponding to Rating on Likert Scale
grammatical. However there was a problem with this question in that for the ordered scale, the answers yes and no appeared instead of a five point likert scale. Nonetheless, most people stated whether they thought it to be grammatical or ungrammatical by clicking yes or no. All those who answered on the binary scale judged the sentence to be grammatical. The sentence, ‘which book would you recommend reading?’ was also one of my grammatical control item and again, almost 100% of participants in each case judged this sentence to be grammatical. The responses given generally included ‘Perfect’, ‘valid’ etc.. A few responses did imply that perhaps to read would be better than reading, but overall this sentence appeared to be acceptable.

The final three grammatical controls included, ‘John announced a plan to steal Bill’s car late tomorrow.’, ‘I wonder whether John can solve the problem’, and ‘Which man did Bill go to Rome to visit?’. The first sentence was again generally perceived as grammatical, however a minute few believed that the phrase late tomorrow did not sound good and thus instead of having 100% of participants judge the sentence as grammatical, 80% and 88.89% judged the sentence to be grammatical in both binary cases. The reasons given, as mentioned, resided with the final phrase; ‘late’ is slightly out-of-place, but still completely understandable’, ‘Would just point out that ‘late tomorrow’ doesn’t sound good, e.g. ‘tomorrow night’ would work better. One participant, who appears to be a non-native, also had a problem with the use of the verb ‘announced’; ‘announced sounds funny here. If you were saying this to someone else, you’d more likely say; ‘John has a plan to steal Bill’s car, late tomorrow. ‘Late tomorrow’ also sounds wrong. Someone would be more likely to saying ‘evening’ or ‘night’.’ On the other hand, the majority of people did accept the sentence, with the majority of the likert scale responses also residing in the 1-2, i.e. grammatical categories.

The second sentence again had its problems, however in this case the order in which the sentence appeared seemed to make a difference. When it appeared in position 14, 90% judged it to be grammatical, however in position 7, 77.78% judged it to be free of errors. The reasons given were that people felt if would work better than whether, or that ‘or not’ could be inserted at the end, but that it did not affect the understanding of the sentence.

Finally the third sentence, ‘Which man did Bill go to Rome to visit?’ was the least successful in its job as a grammatical control. There appeared to be a lot of doubt in its nature in the case of yes meaning grammatical, with only around 60% judging the sentence to be grammatical, which was
unexpected. The consensus for judging the sentence to be ungrammatical seemed to reside in the phrase 'which man', with people stating that they would say 'who' rather than which man. The other responses which were received however claimed the sentence was grammatical and valid. Interestingly, where yes indicated ungrammaticality, 88.5% judged the sentence grammatical. Though no significant difference was found between the phrasing of the question overall, it is clear that in certain cases, this did in fact appear to make a difference.

5.1.5 Ungrammatical controls

The ungrammatical controls included 'She aren’t care about me', 'What did you bring that book to be read out for to?', 'Bill sent London package', 'John teached me how tie my shoes' and finally, 'that is the sort of up with which will not put I'. The interpretation of the first sentence was common across almost all answers, indicating that the verb 'aren’t' should be 'doesn’t'. Some participants gave further insight into their reasons in stating that the verb 'to be' cannot precede 'care' in this sense. Naturally this is the case as the phrase in common convention is 'I don’t care'. Though the sentence is drastically ungrammatical, most people did claim to understand the sentence, however not one person judged to be acceptable under any circumstance.

The second sentence was also judged ungrammatical by almost all participants. Participants again claimed to understand the sentence, however there were too many prepositions and phrases misplaced for it to be judged as a natural sounding sentence. The third sentence listed above, 'Bill sent London package', was regarded in much the same manner, ungrammatical but comprehensible. Many claimed that the preposition 'to' should be inserted, and many others that preposition and determiner needed to be inserted. This would be the correct case as you cannot send the city of London a package, one must send a person a package. One participant made the clever observation that either the sentence is 'missing a preposition, or else there exists a type of package called a London package'. It is interesting that only one person came to this conclusion, however even in that case an indirect object would be missing as 'to send' is a dative verb. Regardless of how in depth one wishes to analyse this sentence, it is clearly ungrammatical and has been judged as such.

'John teached me how tie my shoes' was again judged ungrammatical by the majority of people. The main reason given was that of the verb form
teached being used instead of 'taught'. Very few commented on the fact that the word 'to' was also missing, which may be a result of fast reading. Sometimes people read what they want to read and the fact that there was a mistake early in the sentence may have meant that people just skimmed over the rest of it, but this is mere speculation. There were however two answers that stated the full sentence should be 'John taught me how to tie my shoes'. All the rest merely stated that 'taught' should replace teached', with one answer stating it 'sound[ed] like something a child would say. The final ungrammatical control listed above was also judged to be 100% ungrammatical across the board, with the expected reasons such as 'put up with' is the phrase and should not be separated. It was also claimed that such a sentence would 'never be used by a native speaker' and that there was a lot of collocation present. A few people tried to make sense of the sentence; 'This is something up with which I will not put', but it is clearly difficult to understand the semantics of the sentence. There was a humorous response to this sentence also, stating that this is 'something like Yoda would say', which is actually, funnily enough, where I got my inspiration for finding such an awkward English sentence. Yoda is a character from the film Star Wars, who speaks in such a manner.

5.1.6 Test Items

This brings me to the test items, a number of which have been discussed earlier in the dissertation and so I will not repeat them here. The figures given in Rimmer’s study are also present in this section.

'Who did you quit college because you hated?' was the first, and indeed last sentence which appeared in the case of the reversed order sentences. In general this sentence was perceived as very ungrammatical and the main reason for such was to do with the semantics of the sentence. Many people stated that it took a few readings to decipher the meaning of the sentence, with some trying to formulate what they believed the correct form should be such as 'Why did you quite college? Was it because you hated it?', or 'Who did you hate in college that made you quit? It is evident that there is a general consensus as to how the sentence should have been formed, or how it was understood by some people, but why is this the case? This question may be answered by consulting other responses people gave regarding syntax. Some people recognised the incorrect syntax of the sentence, such as 'Quit already has a direct object - college - and therefore there cannot be another
direct object here (who).’ and that it is impossible to ‘quit someone’. It was also stated that the problem may reside in the fact that the phrase ‘because you hated’ does not contain an object. This result is in keeping with that of the article by (Rimmer, 2006), in which he also claims that the adverbial ‘because you hated’, requires an object which is not present. Consequently, as an average over all four binary cases, only 2.5% of participants judged the sentence grammatical in the current study, which is in keeping with Rimmer’s study of 5.7%.

'John angered while Susan amused the woman’ is the next question under investigation in this section. Rimmer (2006) describes this sentence as conveying 'John made the woman angry at the same time Susan amused her’, or 'John made the woman angry, but Susan on the other hand amused her. This interpretation however, was not so clearly evident in my analysis. There were a few responses which indicated this conclusion; ‘John angered the woman’ and ‘Susan amused the woman’ are valid, and this could be a conjunction of the two, with the extra information that the two events were concurrent’ and 'John angered the woman, while Susan amused her’. The majority of responses actually understood the sentence to be that John was becoming angry, while Susan was amusing the woman. For instance, Doesn’t show the anger is in John. One starts thinking that John is getting someone angry but when it is immediately followed by while we realise that it is John who is getting angry. Therefore it should be "John got angry”, This response is of interest in this context, as the subject clearly interpreted the sentence to initially convey that of what is described by Rimmer, but then changes their opinion upon further inspection. The use of 'whilst' instead of 'while’ was also questioned by one participant who said I think john angered is technically correct but I might use whilst instead of while....dont know why, indicating another problem with native English speakers, i.e. the difference of use between 'while' and 'whilst'.

Interestingly, the order in which this question was asked also seemed to make a difference to the results in the case of yes indicating grammaticality. When the item appeared as the 5th sentence, only 27.27% judged it to be grammatical, with 60% judging it grammatical when it appeared in 16th position. In a similar way to the opposite case, where 58.33% judged it grammatical in 5th position and 33.33% in the reversed position. However, when all of the grammatical results are averaged together, 44.73% of participants judged the sentence grammatical, which is almost a direct match to Rimmer’s study of 43.4%.
This brings me to the next sentence of 'John was bought the book.' This sentence appeared to be valid for some participants but not for others. The general thought was that the passive construction in English is usually something like "The book was bought for John by X", and so an initial reaction to such a sentence would be something like its O.K, but this would be better, which was shown in the following answers; Unusual sentence structure for English; you’d expect it to mention who bought the book for John.’ In Rimmers study 39.6% judged the sentence to be grammatical, in close correlation to the binary scale participants in this study, in which an average of 34.96% judged the sentence to be grammatical. Again, a clear view of the results can be seen in the table 1.

'The woman sitting next to the door’s shoes are like mine' received an average score of 33.48% in the present study and 30.2% in Rimmer’s study. The problem here obviously resides in the possessive phrase, with many people claiming this sentence indicates that the shoes belong to the door; shoes are not the doors, 'door’s’ here implies that the shoes belong to the door.’ Others claimed that if the sentence contained more punctuation it would make it more grammatical. Some people also gave what their interpretation of the sentence was; should be 'the women sitting next to the door has shoes like mine', 'Should be 'woman’s shoes next to...', this contrasts to what Rimmer claims the correct sentence structure would be; 'the shoes of the woman sitting next to the door are like mine', i.e. the use of the ‘of’ phrase to imply the genitive case. Most claimed to understand the sentence regardless, e.g. 'A door can’t have shoes nor can the verb 'to be' go between two nouns door and shoes’. Nonetheless I would understand the comparison being made, with a few saying they found it to be O.K; 'one day the English will accept this....I'm pretty sure I say this but its wrong I guess. I would not correct a foreigner anyway....or myself for that matter'. The latter response relates to that of the problem with stating computer mouses. As mentioned in chapter two, although people may claim such a sentence to be ungrammatical on paper, this does not disregard the fact that such an utterance may be said by a native speaker in everyday speech. Native speakers generally do not have to exert much cognitive effort to imply what it is they wish to say, unlike learners of a language, who may constantly be thinking about what they are saying. It is quite easy to imagine a situation where a person is talking to his or her friend about the woman in the corner and uttering such a sentence without thinking about what it is they are actually saying, or without their friend even noticing the mistake in their sentence. Again, it would be
interesting to carry out an oral experiment on sentences such as these items. This brings me to a common mistake amongst native and non-native speakers of English, the use of 'lay' in the sentence 'You should lay down on the bed.' This sentence was accepted by an average of 59% of participants on the binary scales, again in close comparison to Rimmer’s study, in which 54.7% judged it to be grammatical. The problem lies with the verb 'lay'. A number of participants claimed they believed this construction was allowed in American English, although it sounded 'completely wrong to [them] and 'Not something I would say, but have heard it said and it sounds fine'. Rimmer notes that this construction causes confusion as lay is the past tense form of lie, but he claims this is not the issue here, rather, lie needs an intransitive verb phrase which is present here and lay takes an object. Nonetheless, the fact that lay is the past tense of the verb seemed to be the reason some people deemed the sentence ungrammatical in this case; 'lay is past tense, lie would be correct here'. The other responses simply stated that the verb should be 'lie' but did not give their reasons for stating so.

Another interesting construction which was analysed in this study was 'there's only one person who thinks of themself in that light.' There is only an acute problem with this sentence, the problem being that of the reflexive form of themself. This, however, did not appear to be a major problem for participants on the binary scale in this study, with an average of 65.5% judging it to be grammatical, in comparison to Rimmer’s study in which 47.2% judged it to be grammatical. The reasons included that of the expected, that themselves should exist in place of 'themself', some not knowing why they thought that; 'Would say 'themselves”. Not sure why though...’, and another said it should be 'theirself'. All of the other responses appeared to be in favour of the sentence.

The final sentence, 'Susan trained like she’d never done before' did not, surprisingly, receive one reason response throughout the entire study. The results show that 78.43% of the binary subjects judged the sentence to be grammatical, again, similar to Rimmer’s subject in which 71.7% judged the sentence to be grammatical.

5.2 Conclusion

It is interesting that so many sentences in the present study received such close percentages of grammaticality to that of Rimmer’s study, as his experi-
ment was carried out using teachers and the present study generally contained students with no linguistic background knowledge. As stated in (Odlin, 1994, p.279), the act of using teachers as subjects may verge on that of using linguists as they are generally trained in noticing incorrect grammar and may be prone to judging an ordinary sentence unacceptable or ungrammatical. Although I do agree that native speakers from the general population are perhaps better indicators about what is and is not acceptable, the present study did not appear to differ widely from the results produced by teachers on the same set of test items. There are no significant results available in this study to back up this claim, however by merely viewing the aggregate differences between both groups on some answers, the difference appears to be minimal. This concludes the section on results and discussion. The main results found in this section conclude that there is no significant difference between the order in which the question is put to the participant and that the way in which the question is asked also does not appear to make a difference. It can also be concluded that gender does not appear to make a significant difference, nor do age or native language, though as mentioned above, age and language cannot be considered a fair representation of the entire population.

Despite commonly cited literature, it does not appear that Magnitude Estimation makes a significant difference to the results in relation to the likert or binary scale. Although Magnitude Estimation does not limit the respondents answer, I conclude from this study that if participants do not make use of this fact, and if it does not appear to make a significant difference to the other available scales, then Magnitude Estimation does not appear to be worth the hassle of explaining to participants or having to normalize and interpret the data before one can work with it. Simply using the likert or binary scale would make life a lot easier and as the results of this study suggest, they are as effective any other scale.
Chapter 6

Conclusion

At the beginning of this report a number of questions were put to the reader. It is hoped that during the course of this report such questions have been answered and readers will be able to make their own decisions about whether or not they agree with some of the literature available in present day. Stimulus factors such as context and meaning play an important role in deciding whether or not a sentence is grammatical or ungrammatical and the person’s internal grammar appears to be one of the reasons as to why some people believe a certain utterance is grammatical and others believe it to be ungrammatical. The report then continued to describe the experiment which is the topic of interest in this dissertation and provided an analysis of the results.

I believe the experiment conducted by (Rimmer, 2006), though containing interesting sentences to analyse, was lacking in many areas. The subjects were teachers and therefore not a good representation of the general consensus as they have received training in English grammar and should be more aware of the sometimes minute mistakes in the experimental sentences. The reasons subjects chose the response they did was also not analysed and any contributing factors such as age or gender were not taken into account. However, one of the biggest problems with the study was that it did not contain any control items, a factor which is necessary when conducting such an experiment. If subjects begin to realise that there is a problem with each sentence and there exists only the option of grammatical or ungrammatical, with no space to explain their choice, they may simply tick ungrammatical for each item, or look for a mistake they intuitively did not find at a first glance. Although the present experiment included the problems listed here,
it in itself could also do with some enhancements. The problems encountered throughout this experiment have been discussed. However, a further problem worth mentioning is that the tool Limesurvey did not record the time it took for a subject to complete any given survey. This would be an interesting factor to take into account in a further study as it would provide an indication as to how much cognitive effort was needed to decide the status of any given sentence. Further enhancements which could be made would naturally include a wider range of subjects, in order to form conclusions about whether or not factors such as age and native language have an effect on the results, or the use of a tool which randomises question order for each participant and not just one reverse order. However, the present study did not find any difference between those who answered in one given order over another and so in this case order did not make a difference. Nonetheless it would be interesting to compare the results of this experiment with one done using bare randomisation.

In summary, the notions that grammaticality occurs on a continuum from grammatical to ungrammatical and that it is a dichotomous notion have been reviewed in this study. Factors such as how the wording of a question may affect the results have also been analysed, concluding that there were no significant results in this area. Although it appears to be a fairer description of grammaticality to state that there are many different degrees of ungrammaticality which can be differentiated, if our natural way of cognitive decision making is based on either one state or another, i.e. grammatical or ungrammatical, it may not be necessary to differentiate these different points on a scale. Still, a sentence such as '*Mary talk on the phone' is clearly not as ungrammatical as '*Mary up the phone talk', and thus provides a further argument for supporters of the Likert and Magnitude Estimation scales. Basing my conclusions on the present experiment, there does not appear to be any significant difference between the binary, Likert and Magnitude Estimation scales and thus indicates that even though Magnitude Estimation has been considered one of the best tests for analysing grammar, the results of this study cannot arrive at the same conclusion.
Bibliography


Bertram, D. (2011). Likert Scales..are the meaning of life. unpublished manuscript.


Infosurv (2010). White Paper 5-point vs. 6-point Likert Scales.


Appendix A

Instructions To Candidates

TRINITY COLLEGE DUBLIN
CONSENT FORM FOR PARTICIPANTS

• I am aged 18 years or over and competent to give consent.

• I understand that there are no compulsory questions and that I may skip any question I do not wish to answer.

• I understand that I may opt out of the survey even when I have completed it by not submitting it.

• I understand that because the survey is being carried out using computer monitors, if I or anyone in my family has a history of epilepsy then I am proceeding at my own risk.

• I understand that in the extremely unlikely event that illicit activity is reported, it will be reported to the appropriate authorities.

• I will not name third parties in any open text field of the questionnaire, as any such replies will be anonymised.

• I understand that all data received is anonymous and that participation is completely voluntary.

• I understand that my answers will be used in aggregate form with those of others to inform an undergraduate final year dissertation and other scholarly works.
If you wish to receive a copy of the results of my analysis please do not hesitate to contact me at scallanf@tcd.ie
If you agree to the above conditions please click on the link entitled Survey, alternatively you may return to Facebook!
The main purpose of my dissertation is to explore different methods used in grammaticality judgement tests. The research is being carried out as part of my final year undergraduate dissertation in Computer Science, Linguistics and German.

In this survey you will see 20 different sentences, each followed by a question requesting the reasons for your answer. Some of the sentences will appear perfectly natural to you. The task in each case is to decide how native-like the sentence sounds. Another way of looking at it is if you encountered an individual who was learning English as a second language, would you correct them or believe it to be OK. There are no right or wrong answers for any of the questions. Please do not linger on a sentence; it is your initial reaction that I am interested in.

There are no compulsory questions; you may skip any question you do not wish to answer. You may also opt out of the survey even when you have completed it by not submitting it.

Participants were chosen by randomly from my friends list on the social network Facebook. Alternatively you may have seen information about the survey on a poster.

The survey will take no longer than 10 minutes to complete.

In the extremely unlikely event that illicit activity is reported I will be obliged to report it to the appropriate authorities.

Please do not name third parties in any open text field of the questionnaire. Any such replies will be anonymised.

All data received will be anonymous; you do not have to give your name in any section of the survey and participation is completely voluntary. The survey is open to all volunteers aged 18 years and over.

Finally thank you for your participation in this survey.
Binary Instructions Experiment One

If you feel a sentence sounds like something you would not correct an English language learner for saying, i.e. it sounds native-like, please click Yes, indicating it is acceptable and does not need correction, or if you feel the opposite, please click No, indicating it is unacceptable and needs correction.

For example, you may feel that the following sentence sounds native-like, so you click ‘Yes’.

Which man did Bill go to Rome to visit?

However, you may feel the second sentence does not sound native-like, so you click ‘No’.

Where did Bill goes to Rome to visit?
Binary Instructions Experiment Two

If you feel a sentence sounds like something you would correct an English language learner for saying, i.e. it does not sound native-like, please click Yes, indicating it is unacceptable and needs correction, or if you feel the opposite, please click No, indicating it is acceptable and does not need correction.

For example, you may feel that the following sentence sounds native-like, so you click 'No'.

Which man did Bill go to Rome to visit?

However, you may feel the second sentence does not sound native-like, so you click 'Yes'.

Where did Bill goes to Rome to visit?
Likert Scale Instructions

If you feel a sentence sounds like something you would most likely not correct an English language learner for saying, i.e. it sounds native-like and does not need correction, please click 1, indicating it is acceptable, or if you feel the opposite, please click 5, indicating it is unacceptable. If you feel the sentence is between those extremes, please click 2, 3 or 4 as appropriate to your graded scale.

For example, the first sentence below is OK but not great, so we will give it a rating of 3.

Which man do you wonder when to meet?

You may think that the next sentence sounds perfectly native-like, so we will give it a rating of 1.

Which man did Bill go to Rome to visit?

You may think that the following sentence sounds unnatural, so we will give it a rating of 4.

Where did Bill go to Rome to work?

Finally you may think that the fourth sentence is completely unnatural, as it is difficult to make any sense of at all. So it gains a score of 5.

Where did Bill goes to Rome to work?

Remember, you are rating the sentences on a scale of 1-5, 1 being the most native-like sounding and 5 being the most unnatural.
Magnitude Estimation Instructions for Candidates

In this experiment you will have to give each sentence a numerical rating. A reference sentence will be provided and you will assign this sentence a numerical value and then compare all other sentences to it. If you feel a sentence sounds, in comparison to the reference sentence, some amount more in need of your correction for an English language learner, (i.e. it sounds less native-like than the reference sentence) then assign it with a number that many times smaller than your reference number; if you feel it is some amount less in need of correction, (i.e. it sounds more native-like than the reference sentence) then assign it a number that many times larger. Finally if you feel a sentence is equal to that of the reference sentence, please assign it the same numerical value you gave to the reference sentence. You may choose any number you wish; perhaps large even numbers would be easier to work with.

Here is an example. The first sentence below is the reference sentence. Let's give it a rating of 100.

Which man do you wonder when to meet?

You may think the next sentence is twice as good as the reference sentence, because it sounds more native like, so we'll give it a rating of 200 (100 \times 2).

Which man did Bill go to Rome to visit?

You may feel that the third sentence is twice as bad as the reference sentence, so it gains a rating of 50 (100 / 2).

Where did Bill go to Rome to work?

Finally you may feel that the fourth sentence is three times as bad as the reference sentence, as it is difficult to make any sense of at all. So it gains a score of 30 (roughly 100 / 3).

Where did Bill goes to Rome to work?

You may use any positive number you wish for the reference quote, however
please do not use negative numbers or zero. It is important to remember that you are not rating a sentence in comparison with the previous one, rather, you are rating it in comparison to the reference sentence.

1. The reference sentence is: When do you know the man whom Mary invited?

Please assign a number of your choice to this sentence.
My name is Fiona Scallan and I am a final year undergraduate student of Computer Science, Linguistics and German. My final year project entails grammaticality judgments and I am conducting an online survey as part of my analysis. If you would be interested in completing a survey for me, I would be grateful if you would follow one of the links given below.

Alternatively, feel free to contact me at scallanf@tcd.ie to obtain further information.

Thank You!

www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
www.cs.tcd.ie/clg/FS/experiment.html
Appendix B

Tables
TABLE 1
Binary scale: yes meaning acceptable and No meaning unacceptable.

<table>
<thead>
<tr>
<th></th>
<th>Order Gramm (%)</th>
<th>Order Ungram (%)</th>
<th>No answer (%)</th>
<th>Reverse Order Gramm (%)</th>
<th>Reverse Order Ungram (%)</th>
<th>No answer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who did you quit college because you hated?</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>10</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>She aren’t care about me.</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Either you or I are wrong.</td>
<td>81.82</td>
<td>9.09</td>
<td>9.09</td>
<td>80</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>Which book would you recommend reading?</td>
<td>90.91</td>
<td>9.09</td>
<td>9.09</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>John angered while Susan amused the woman.</td>
<td>27.27</td>
<td>72.73</td>
<td>0</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Who did John invite?</td>
<td>90.91</td>
<td>0</td>
<td>9.09</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>What did you bring that book to be read out for to?</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>The plane that the pilot that the police questioned flew crashed.</td>
<td>18.18</td>
<td>81.82</td>
<td>0</td>
<td>30</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>John was bought the book.</td>
<td>9.09</td>
<td>90.91</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>John announced a plan to steal Bill’s car late tomorrow.</td>
<td>81.82</td>
<td>9.09</td>
<td>9.09</td>
<td>90</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>The woman sitting next to the door’s shoes are like mine.</td>
<td>27.27</td>
<td>63.64</td>
<td>9.09</td>
<td>40</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>You should lay down on the bed.</td>
<td>63.64</td>
<td>36.36</td>
<td>0</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>I wonder whether John can solve the problem.</td>
<td>90.91</td>
<td>9.09</td>
<td>0</td>
<td>80</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>John taught me how tie my shoes.</td>
<td>9.09</td>
<td>81.82</td>
<td>9.09</td>
<td>20</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>I bought three mouses at the computer store.</td>
<td>36.36</td>
<td>63.64</td>
<td>0</td>
<td>30</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>There’s only one person who thinks of themself in that light.</td>
<td>90.91</td>
<td>9.09</td>
<td>0</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>That is the sort of up with which will not put I</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Which man did Bill go to Rome to visit?</td>
<td>63.64</td>
<td>36.36</td>
<td>0</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Susan trained like she’d never done before.</td>
<td>90.91</td>
<td>9.09</td>
<td>0</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
## APPENDIX B. TABLES

**TABLE 2**

Binary scale: Yes meaning unacceptable and No meaning acceptable.

<table>
<thead>
<tr>
<th>Question</th>
<th>Order Gramm (%)</th>
<th>Order Ungram. (%)</th>
<th>No answer (%)</th>
<th>Reverse Order Gramm (%)</th>
<th>Reverse Order Ungram. (%)</th>
<th>No answer (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who did you quit college because you hated?</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>She aren’t care about me.</td>
<td>8.33</td>
<td>91.97</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Either you or I are wrong.</td>
<td>91.67</td>
<td>8.33</td>
<td>0</td>
<td>44.44</td>
<td>55.56</td>
<td>0</td>
</tr>
<tr>
<td>Which book would you recommend reading?</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>88.89</td>
<td>11.11</td>
<td>0</td>
</tr>
<tr>
<td>John angered while Susan amused the woman.</td>
<td>58.33</td>
<td>41.67</td>
<td>0</td>
<td>33.33</td>
<td>55.56</td>
<td>11.11</td>
</tr>
<tr>
<td>Who did John invite?</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>What did you bring that book to be read out for to?</td>
<td>8.33</td>
<td>83.33</td>
<td>8.33</td>
<td>11.11</td>
<td>88.89</td>
<td>0</td>
</tr>
<tr>
<td>The plane that the pilot that the police questioned flew crashed.</td>
<td>16.67</td>
<td>83.33</td>
<td>0</td>
<td>0</td>
<td>88.89</td>
<td>11.11</td>
</tr>
<tr>
<td>John was bought the book.</td>
<td>58.33</td>
<td>41.67</td>
<td>0</td>
<td>33.33</td>
<td>66.67</td>
<td>0</td>
</tr>
<tr>
<td>Bill sent London package.</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>John announced a plan to steal Bill’s car late tomorrow.</td>
<td>83.33</td>
<td>16.67</td>
<td>0</td>
<td>66.67</td>
<td>33.33</td>
<td>0</td>
</tr>
<tr>
<td>The woman sitting next to the door’s shoes are like mine.</td>
<td>33.33</td>
<td>66.67</td>
<td>0</td>
<td>33.33</td>
<td>44.44</td>
<td>22.22</td>
</tr>
<tr>
<td>You should lay down on the bed.</td>
<td>58.33</td>
<td>33.33</td>
<td>8.33</td>
<td>44.44</td>
<td>55.56</td>
<td>0</td>
</tr>
<tr>
<td>I wonder whether John can solve the problem.</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>88.89</td>
<td>11.11</td>
<td>0</td>
</tr>
<tr>
<td>John taught me how tie my shoes.</td>
<td>0</td>
<td>91.67</td>
<td>8.33</td>
<td>11.11</td>
<td>88.89</td>
<td>0</td>
</tr>
<tr>
<td>I bought three mouses at the computer store.</td>
<td>50</td>
<td>33.33</td>
<td>16.67</td>
<td>22.22</td>
<td>77.78</td>
<td>0</td>
</tr>
<tr>
<td>There’s only one person who thinks of themself in that light.</td>
<td>66.67</td>
<td>33.33</td>
<td>0</td>
<td>44.44</td>
<td>55.56</td>
<td>0</td>
</tr>
<tr>
<td>That is the sort of up with which will not put I</td>
<td>8.33</td>
<td>83.33</td>
<td>8.33</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Which man did Bill go to Rome to visit?</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>77.78</td>
<td>22.22</td>
<td>0</td>
</tr>
<tr>
<td>Susan trained like she’d never done before.</td>
<td>75</td>
<td>25</td>
<td>0</td>
<td>77.78</td>
<td>22.22</td>
<td>0</td>
</tr>
</tbody>
</table>
Pie Chart of Native Languages

- English: 79%
- Not Given: 15%
- German: 2%
- Irish: 2%
- Latvian: 1%
Figure B.1: Information sheet for Participants

Figure B.2: Example of Question One
Figure B.3: *Example of Question Two*

![Example of Question Two](image)

Figure B.4: *Consent Form*

![Consent Form](image)

Trinity College Dublin

Consent Form

1. I am aged 18 years or over and competent give consent.
2. I understand that any upset or distress caused by the survey is at my own risk.
3. I understand that the data obtained whilst participating in this survey is anonymous.
4. I understand that participation in the survey is voluntary.
5. I understand that consent for this survey will remain in effect throughout my participation.
6. I understand that I am free to withdraw from the survey at any time.

Consent Form completed by: [Name]
Date: [Date]

Consent Form: [Signature]

Consent Form: [Signature]

Consent Form: [Signature]
# Appendix C

## Html Code

```html
<html><head><title>TCD CS Fiona Scallan</title></head>
<body bgcolor="#ffffff" width=600>
<center>
<TABLE BORDER=0 CELLPACING=1 CELLPADDING=1>
<!-- First row -->
<TR>
<!-- Embedded Table -->
<TABLE BORDER=0 CELLPACING=4 CELLPADDING=4>
<!-- First row -->
<TR>
   <TD ALIGN=left>
      <IMG SRC="https://www.cs.tcd.ie/wwwresources/du_cols.gif" ALT="Dublin"
   </TD>
   <TD ALIGN=center>
      <font size="-2"><strong>
         <a href="https://www.cs.tcd.ie">Computer Science Department</a>,
         <a href="https://www.tcd.ie">Trinity College</a>
      </font>
   </TD>
   <TD>
      <img src="https://www.cs.tcd.ie/wwwresources/tcd_crss.gif" ALT="Trinity"
   </TD>
   <TD>
      <img src="https://www.cs.tcd.ie/wwwresources/tcd_crss.gif" ALT="Trinity"
   </TD>
</TR>
</TABLE>
<!-- The University of Dublin<br> -->
<!-- Trinity College<br> -->
</TABLE>
</center>
</body></html>
```
<hr width=600>

<!-- -------- End header -->

<!-- -------- End header -->

<center><Table width=600 border=0><P><h1>Trinity College Dublin</h1>
<TR><TD ALIGN=center><strong>Researcher’s Name</strong>: Fiona Scallan</TD></TR></P>
<br>
<TR><TD ALIGN=center><strong>Consent Form</strong>: <li> I am aged 18 years or over
    <li> I understand that there are no compulsory questions and that I may skip any
    <li> I understand that I may opt out of the survey even when I have completed
    <li> I understand that because the survey is being carried out online, if I complete
    <li> I understand that in the extremely unlikely event that illicit activity
    <li> I will not name third parties in any open text field of the questionnaire.
    <li> I understand that my answers will be used in aggregate form with those of others.
    <li> I understand that my answers will be used in aggregate form with those of others.
    <li> If you wish to receive a copy of the results of my analysis please do not hesitate
    <li> If you agree to the above conditions please click on the link entitled ’Survey’
</table>
</center>
<hr width=600><font size="-1">
<!-- 
[<a href="https://www.cs.tcd.ie/dept/personnel/">Personnel</a>]
[<a href="https://www.cs.tcd.ie/cgi-bin/ldap/directory.cgi">Update your</a>]
[<a href="https://www.cs.tcd.ie/">Computer Science</a>]
[<a href="https://www.tcd.ie/">Trinity College Dublin</a>]
[<a href="http://www.facebook.com/">Facebook</a>]
[<a href="http://www.scss.tcd.ie/disciplines/intelligent_systems/clg/clgcl"]

<!--<br> Last modified: Fri, 26 Sep 2008 12:39</font> -->
<br>
<br> <font size="-2"><address>WebAdmin at cs.tcd.ie</address></font></center>