A multivariate study of explicitation in English translations: optional ‘that’ in complement clauses

Word count: 15,259

Eline Vermeire
Student number: 01402664

Supervisor: Prof. Dr. Gert De Sutter

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Arts in Multilingual Communication

Academic year: 2017 – 2018
De auteur en de promotor(en) geven de toelating deze studie als geheel voor consultatie ter beschikking te stellen voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting de bron uitdrukkelijk te vermelden bij het aanhalen van gegevens uit deze studie.

Het auteursrecht betreffende de gegevens vermeld in deze studie berust bij de promotor(en). Het auteursrecht beperkt zich tot de wijze waarop de auteur de problematiek van het onderwerp heeft benaderd en neergeschreven. De auteur respecteert daarbij het oorspronkelijke auteursrecht van de individueel geciteerde studies en eventueel bijhorende documentatie, zoals tabellen en figuren.
Abstract

This study investigates whether there is a difference between original English and translated English (source language: Dutch) in including or omitting that in complement clauses. In doing so, it is also scrutinised which factors are decisive for the choice between that and zero in original, non-translated English and whether these factors are similar in translated English. The data is extracted from the Dutch Parallel Corpus and the MuPDAR approach is applied, which means that a general linear mixed-effects model is fitted for the non-translated data and subsequently tested on the translated data in order to verify to what extent translated English differs from original English. The results indicate that in general that-complementation functions very similarly in translated and non-translated English. However, a slight tendency towards using a that-complementiser more frequently in translated English could be found. In order to find explanations for the small differences in that-complementation between translated English and original, three hypotheses are tested. The hypotheses are adopted from Kruger (2018) and consist of the transfer hypothesis, the cognitive complexity hypothesis and the risk-aversion hypothesis. According to this study, both the cognitive complexity hypothesis and the risk-aversion hypothesis seem to provide a plausible explanation for the higher that-rate in translated English.
Acknowledgements

First of all, I would like to express my sincere gratitude to my promotor Gert De Sutter for guiding me through this process and for patiently explaining the working of the statistical program. In addition, I would like to thank Fieke Van der Gucht for proofreading my dissertation and providing suggestions to improve my writing style. Furthermore, I want to thank Sonia Vandepitte for checking parts of my dissertation on language mistakes. Lastly, I would like to thank my family and friends for their support during this period.
Table of contents

List with tables and figures.................................................................................................................. 1
1 Introduction ........................................................................................................................................ 2
2 Theoretical framework....................................................................................................................... 4
2.1 Explaining explicitation in translation studies............................................................................... 4
2.2 Previous research on that-explicitation in original English.......................................................... 8
2.2.1 Explanatory principles on that-realisation in original English.................................................. 8
2.2.2 Multifactorial corpus studies on that-realisation in original English ....................................... 10
2.3 Multifactorial corpus studies on that-realisation in translated English......................................... 13
3 Methodology ...................................................................................................................................... 16
3.1 Corpus ........................................................................................................................................... 16
3.2 Data extraction............................................................................................................................... 17
3.3 Variables.......................................................................................................................................... 17
3.3.1 Dependent variable ....................................................................................................................... 17
3.3.2 Independent variables ................................................................................................................... 18
3.4 Statistical analysis.......................................................................................................................... 21
4 Results ................................................................................................................................................ 24
4.1 Factors influencing that-realisation in non-translated English ....................................................... 25
4.2 The application of R1 to the translated dataset .............................................................................. 27
4.3 Factors leading to deviant choices ................................................................................................ 28
4.4 Direction of the deviant choices .................................................................................................... 30
5 Discussion ......................................................................................................................................... 32
6 Conclusion ......................................................................................................................................... 36
7 References .......................................................................................................................................... 38
8 Appendices .......................................................................................................................................... 40
8.1 Appendix A: Full statistical report on the fitting of the R1 model .................................................. 40
8.2 Appendix B: Full statistical report on the fitting of the R2 model .................................................. 41
List with tables and figures

Table 1 (p. 20): Overview of the factors with the corresponding factor levels

Figure 1 (p. 25): Plots of the significant factors for *that*-realisation in non-translated English: MCVerbToCCLength, LemmaConstrFreq100kLOG, aspect, subject, text type and tense/modality

Figure 2 (p. 29): Plots of the factors influencing the correct choice for *that*-realisation in the translated dataset: Text Type, MCVerbToCCLength, LemmaConstrFreq100kLOG, subject and aspect
1 Introduction

This study probes into the alternation between *that* and *zero* in English complement clauses, illustrated in the following examples (Dutch Parallel Corpus, 2018):

1. He insisted *that/ Ø* I repeat that after him.
2. He said *that/Ø* he wanted to see how strong he was.
3. In the painting you see *that/ Ø* the Tsar was the only blue-blooded guest.
4. I don’t think *that/Ø* media education will be done in schools.

When *that* is present, the relation between matrix clause and complement clause is explicit whereas *zero* (*Ø*) allows the complement clause to carry the dependency information. In other words, the only difference between *that* and *zero* is the explicitation of syntactic information. The advantage of an explicit *that* is that it “permits the complement clause to be immediately recognizable as subordinate, and distinct from the following main clause” (Hawkins, 2004, p. 154). This phenomenon has already been widely researched in variational linguistics, language learning and translation studies. The latter focusses on the difference in *that*-realisation between translations and non-translated original texts. In language learning, on the other hand, the difference between natives and non-native learners of English is examined and in variational linguistics, the factors which seem to influence the choice between *that* and *zero* in original English are investigated.

Many researches in translation studies have already indicated that a *that*-complementiser is more frequent in translations than in non-translated texts (Olohan & Baker, 2000). Overall, translators have a tendency to make implicit meanings and information explicit in the target text. According to Blum-Kulka (1986) and Olohan & Baker (2000), this tendency is inherent to the translation process, which includes that the explicitation bias can be seen as a “universal of translation” for which no specific other reason can be found. Becher (2010), on the other hand, rejects the translation-inherent explanation of Blum-Kulka (1986) and assigns the explicitation tendency to the fact that translators want to overcome cultural distance or that they want to make sure that the text is understood by the target audience. Additionally, Becher (2011) argues that translators try to follow conventional target language systems or intent to comply with the communicative norms of the target language. The most common explanation is the one of risk-avoidance in a sense that translators prefer explicitation to ensure understanding between the author and the readers (Becher, 2011; Pym, 2005 and Pym, 2015). Translators rather play safe than venture not being understood. In section 2.1, these explanations will be discussed in more detail.
That-realisation has not only been examined in translation studies. Also many researchers devoted their studies to finding the reasons for omitting or adding *that* in original English. For instance, *that* tends to be omitted when language users were asked to repeat a complement clause sentence and several cues concerning the complement clause were given (Ferreira & Dell, 2000). This experiment was conducted in order to examine whether the availability of lexical items affects the choice between *that* and *zero* in language processing. Moreover, *that* tends to be added in a target sentence when this sentence is primed with a lexically and syntactically similar *that* (Ferreira, 2003). Furthermore, combinatorial and/or dependency relations between matrix clause and complement clause influence the choice for *that* or *zero* in a sense that *that* is preferred when complement clauses cannot be adjacent to their heads (Hawkins, 2003). These researches will be discussed in depth in section 2.2.1. To further examine the reasons for the explicitation or implicitation of *that*, multifactorial studies were conducted to research several factors at the same time. These multifactorial studies found that also factors within the sentence condition the *that/zero* alternation (Tagliamonte & Smith, 2004; Torres Cacoullos & Walker, 2009; Jaeger, 2010; Thompson & Mulac, 1991 and Wulf, Gries & Lester, to appear). A few examples are the subject of the main clause, intervening material between the main clause and the complement clause, the frequency of the main verb, etc.

This study attempts to combine these two research strands, i.e. variational linguistics and translation studies, following the example set by Kruger (2018). While her multifactorial study focusses on the difference between original English and translations from Afrikaans to English, this research will examine optional *that* in complement clauses in original English texts and in English texts translated from Dutch. This study addresses two research question:

1. In what respect (and to what extent) does the choice between *that* and *zero* in translated English differ from that in non-translated English?
2. Which factors tend to be more decisive for the choice between *that* and *zero* and which factors are less important in non-translated English? And to what extent do these factors have a different effect for translated English?

By means of a multivariate analysis both research questions will be investigated in order to corroborate three hypotheses formulated by Kruger (2018): the transfer hypothesis, the cognitive complexity hypothesis and the risk-aversion hypothesis. These hypotheses will be discussed thoroughly in section 2.3. For the analysis, all the sentences including a public verb or private verb (based on Quirk et al., 1985) were extracted from the Dutch Parallel Corpus (DPC) and the relevant sentences were distinguished from the
irrelevant ones, i.e. only the sentences which included a complement clause that was introduced or could be introduced by the complementiser that. When the sentences turned out relevant, the data was annotated manually according to the same eight factors as Kruger (2018), which will be discussed in depth in section 3.3.2. Next, the MuPDAR approach was applied: a generalized-linear mixed effects model for that-realisation was fitted on the non-translated sentences first and subsequently tested on the translated sentences to answer the above-mentioned research questions. Before we discuss the methodology in greater detail, more attention will be paid to the theoretical framework that has been used and previous studies on related issues. This paper is structured as follows:

In section 2 existing research in translation studies will be considered as well as previous research on explicitation in original English. The study of Kruger (2018) and her main findings will be discussed in section 2.3. Section 3 elaborates on the methodology including a discussion of the corpus and data extraction, the annotation of the dependent and independent variables and the statistical analysis. The results of the analysis will be discussed in section 4. Section 5 presents a discussion and in section 6 the conclusion and suggestions for further research can be found.

2 Theoretical framework

2.1 Explaining explicitation in translation studies

The hypothesis that translations tend to be longer than their corresponding source texts is traditionally associated with Blum-Kulka (1986). One of the reasons for this hypothesis is that translators often make information explicit that is implicit in the source text in order to clarify the meaning or the sociocultural background of the source text to the target audience (Becher, 2010). Another reason is that a translator has to make implicit constructions of the source text explicit in the target text when there are grammatical differences between source text and target text (Becher, 2010). Baker (1993, p. 243) referred to examples from her previous study in which sentences were extended in the target text in order to give more background information. For instance, a short clause like “the example of Truman was always present in my mind” was translated into Arabic as:

In my mind there was always the example of the American president Harry Truman, who succeeded Franklin Roosevelt towards the end of World War II. At that time – and after Roosevelt – Truman seemed a rather nondescript and unknown character who could not lead the great human struggle in World War II to its desired and inevitable end. But Truman – faced with the
challenged of practical experience – grew and matured and became one of the most prominent American presidents in modern times. I imagined that the same thing could happen to Sadat.

The most well-known explanation for the explicitation bias in translations is Blum-Kulka’s Explicitation Hypothesis (1986), which postulates “an observed cohesive explicitness from SL to TL\(^1\) texts regardless of the increase traceable to differences between the two linguistic and textual systems involved. It follows that explicitation is viewed here as inherent in the process of translation” (Blum-Kulka, 1986, p. 19). This hypothesis was tested by Olohan & Baker (2000) by analysing the patterns of the explicitation or implicitation of optional *that* for the verbs ‘say’ and ‘tell’ in a corpus of original English (BNC) and a translated English corpus (TEC). According to Olohan & Baker (2000, p. 143), “a higher incidence of the optional *that* in translated English would provide evidence of inherent, subliminal processes of explicitation in translation”. In other words, it would confirm the hypothesis of Blum-Kulka. The results of their research indeed indicated that the complementiser *that* was used more frequently in TEC and a zero-connective was more frequent in BNC. Moreover, it appeared that a *that*-connective was used more with simpler syntactic constructions in translated English, even though simple syntactic constructions prefer a zero-complementiser in non-translated English since they are easier to process (Rohdenburg, 1996). This reveals that translators overuse the complementiser in situations where normally zero would be preferred.

While Olohan & Baker (2000) try to find evidence to support the Explicitation Hypothesis, Becher (2010, p. 1) argues that the hypothesis “rests on fallacious theoretical considerations and premature interpretations of empirical data”. He claims that the hypothesis is unparsimonious, not sufficiently motivated and vaguely formulated. Firstly, according to Occam’s Razor, hypotheses have to be parsimonious, which make them easier to handle and more trustworthy. “Blum-Kulka assumes the existence of a new entity, namely a new, translation-inherent type of explicitation\(^2\). This means that any other hypothesis that might explain an observed tendency of explicitation in translation without assuming a new type of explicitation would be more compatible with Occam’s Razor and thus preferable to the Explicitation Hypothesis” (Becher, 2010, p. 7). Secondly, the hypothesis is not sufficiently motivated because the preference for explicitation on behalf of translators could have other causes than “the process of interpretation performed by the translator on the source text” (Becher, 2010, p. 7). Thirdly, in her paper, Blum-Kulka (1986, p. 21) paraphrases the Explicitation Hypothesis as positing that “explicitation is a universal strategy inherent in the process of language mediation”. The term ‘strategy’ was interpreted

---

\(^1\) Source language to target language

\(^2\) Obligatory explicitation, optional explicitation and pragmatic explicitation can be regarded as the “old” types of explicitation.
differently by researchers who based their study on this hypothesis as it is unclear whether she means a conscious or unconscious strategy. Therefore, Becher (2010) concluded that the hypothesis is vaguely formulated. In addition, studies defending the Explicitation Hypothesis, such as the research of Olohan & Baker (2000), fail to give convincing evidence for the translation-inherent nature of explicitation (Becher, 2010). Consequently, Becher prefers the use of an adapted version of Klaudy’s (2009) Asymmetry Hypothesis instead of the Explicitation Hypothesis. The Asymmetry Hypothesis posits that “obligatory, optional and pragmatic explicitations tend to be more frequent than the corresponding implicitations regardless of the SL/TL constellation at hand” (Becher, 2010, p. 17). On this view, Becher accepts the explicitation preference of translators while abandoning the idea of a translation-inherent type of explicitation. He motivates this hypothesis explaining that a translator produces translations that are more explicit than their source texts to overcome the cultural distance between the source language and the target language, not because he somehow feels an inherent, subconscious need for explicitation. Furthermore, a translator must make sure that the source language author is understood by the target language reader and he wants to avoid the risk of being misunderstood (Becher, 2010). Thus, he or she has a stronger tendency to translate explicitly rather than implicitly.

Becher (2011) goes more into detail on the reasons why translators have a bias towards adding connectives. In an English-German corpus, he identified and counted the omissions and additions of connectives. He concluded that the majority of the alternations of connectives should not be attributed to a “universal strategy” of explicitation. Instead, he offered four reasons for the different use of connectives in the two languages. Firstly, the German language generally uses more connectives and therefore translators could be biased towards explicitation in order to “comply with the communicative norms of the target language” (Becher, 2011, p. 32). Moreover, they want to avoid a “stylistically marked way of expression” (Becher, 2011, p. 39). Secondly, a translator could prefer adding connectives to exploit the syntactic and lexical features, which are offered by the target language system. Thirdly, sometimes a connective is necessary because the target language does not offer a similar construction, e.g. German does not have an equivalent for the English ing-form. Ultimately, since the main goal of translators is to ensure understanding between the author and the target audience, they have a tendency to optimise the cohesion of the target text. This last explanation can be linked to the fact that translators act as “risk-avoiding mediators between cultures proposed by Pym (2005)” (Becher, 2011, p. 41). This last observation is in line with the Asymmetry Hypothesis of Klaudy (2009).
Similar to Becher (2010 & 2011), Pym (2005) attributed the arguments for the explicitation preference within “the frame of risk management”. Pym (2005) argues that translators tend to be risk-averse because they are prudent. Furthermore, Pym (2005) alleges that translators, due to the fact that they are both readers and writers of the text, notice the difficulties of a text more easily and try to make the meaning more explicit by adding markers. Additionally, he argues that a translator gives communicative clues because the receivers of the target text have less knowledge about, for instance, the cultural references than the receivers of the source text.

Pym (2015) further elaborated on translation as risk management in a sense that translators tend to avoid taking risks in order to provide the expected product. However, Pym (2015, p. 78) argues that translators do not necessarily have to avoid risks all the time because “a culture that relies on translations tends to produce and consume quite anodyne texts, unable to surprise or inspire”. He divides risks into three types: the credibility risk, the uncertainty risk and the communicative risk. The credibility risk refers to translators losing their credibility in the eyes of their clients, for instance by making a minor translation mistake. As a result, translators tend to be risk-averse in order to prevent them from losing credibility. Translators are considered to take risks when they are not certain about the meaning of an element in the source text, this is the so-called the uncertainty risk. It implies that when the text is rather difficult, translators tend to guess or figure out the right translation. Nevertheless, Pym (2015) also mentions the gains from risk-taking and emphasises that there is a difference between guessing and trying to find a neologism for an non-existing word in the target language or adapting a slogan culturally for a new target audience. The communicative risk indicates that high-risk items, such as the name and date of birth on a birth certificate, can be crucial for the communicative success and, therefore, translators should invest a higher effort in translating high-risk items than in translating low-risk items.

All the above studies aim to find reasons why explicitation is more frequent in translated English than in original English by focussing on the translation process. In other words, the researchers assume that the explicitation tendency is typical of translators and they suggest several arguments ranging from the Explicitation Hypothesis to risk management on behalf of translators. This study, however, does not assume that the implicitation or explicitation of that is only the result of the translation process. Alternatively, this study hypothesises that the inclusion of optional that will also be dependent on other factors, which are the same for both original English and translated English. Within the area of variational linguistics and language learning, many studies have already indicated that the addition or omission of the complementiser that is subject to syntactic and semantic factors and it remains to be seen how important
the factor translation status (translated vs. non-translated) is vis-à-vis other language-external and -internal factors. The next section provides a brief overview of this type of studies.

2.2 Previous research on that-explicitation in original English

2.2.1 Explanatory principles on that-realisation in original English

Explicitation and implicitation preferences have been largely scrutinised in original English in a sense that many researchers examined several principles, which they assumed would have an influence on the addition or omission of that. On the one hand, it was examined whether that-realisation was influenced by ambiguity and lexical availability or by prime sentences which included the word that. On the other hand, it was researched why complement clauses, which were not introduced by that, preferred to be close to the main clause.

Ferreira & Dell (2000), for example, conducted six experiments to verify whether the mention of optional that is guided by ambiguity and/or availability-based factors in language production. For their research Ferreira & Dell (2000) relied on two approaches to language production. The first approach, which is the lexical availability prediction, stipulates that “language production processes choose syntactic structures that ease the task of creating sentences” (Ferreira & Dell, 2000, p. 296). The second approach, which is the ambiguity prediction, states that “optional words like that are used to avoid temporarily ambiguous, difficult-to-comprehend sentences” (Ferreira & Dell, 2000, p. 296). To test the lexical availability prediction, several sentences were presented to the subjects and they had to recall them with visually presented cues. These cues consisted of the complement clause subjects or complement clause material. Ferreira & Dell (2000) argued that if the subject of the complement clause was shown, there would be a greater tendency to omit the complementiser that. To test the ambiguity prediction, the sentences presented were temporary ambiguous sentences with and without that, which also had to be recalled by means of visual cues. An example of a temporary ambiguous sentence is ‘the coach knew you…’. When only this part of the sentence is given, it is not yet clear whether you is a direct object (e.g. the coach knew you well) or whether you is the subject of the main clause (e.g. the coach knew you missed practice). If a complementiser would be inserted, no doubt can arise about the function of you (e.g. the coach knew that you...). The results demonstrated that first of all, speakers do not have a tendency to disambiguate sentences by adding the complementiser that. Yet, the use of that is subject to the availability of spoken material, which means that when complement clause subjects were given as visual cues, speakers were
more likely to omit that. This last observation leads to the question as to whether the addition or omission of that is, on the one hand, determined by the inclusion or exclusion of the lexical item which corresponds with the word that or, on the other hand, the choice between a syntactic structure with that and another one without that (Ferreira & Dell, 2000). Nevertheless, “since the choice of a syntactic structure varies with the activation of only one lexical item”, speakers tend to choose between syntactic structures rather than between lexical items (Ferreira & Dell, 2000, p. 328). Therefore, Ferreira & Dell (2000) concluded that syntactic structure mechanisms mediate lexical availability effects.

A follow-up study by Ferreira (2003) researched whether the explicitation of a that-complementiser in complement clauses could be primed by the previous production of a sentence which included a lexical similar that or a lexical and syntactical similar that. His results showed that prime sentences with complement clauses including that lead to target sentences also mentioning the optional complementiser. In other words, “target that-mention was influenced by primes with lexically and syntactically similar thats” (Ferreira, 2003, p. 179). However, primes with only a lexically similar that, such as transitive primes with the determiner that (e.g. the company ensured that farm for…) and noun-complement primes with complementiser that (e.g. the theory that…) had no effect on the mention of that in complement clauses. In order to corroborate this finding, Ferreira (2003) introduced neutral primes, which are primes where no that could be added or omitted. The results pointed out that in 73% of the time speakers did produce a that-complementiser after these neutral primes. “This reveals speakers’ bias to mention thats in targets more often than they omit them independent of any effect of priming” (Ferreira, 2003, p. 391).

Hawkins (2003), in his turn, tried to find an explanation for Rohdenburg’s complexity principle (1996, p. 151), which postulates that “explicitly-marked phrases are preferred over zero-marked counterparts in cognitively complex environments”. Hawkins (2003) introduced a general hypothesis for adjacency which supposes that the more combinatorial and/or dependency relations exist in a phrase, the more these phrases tend to be close to their heads, which makes it easier to process them. When the subordinate phrases are zero-marked, they are more dependent on their heads and therefore favour to be adjacent. Yet, explicitly-marked phrases show the link between the head and the phrase. Therefore, they are easier to process and do not have the necessity to be adjacent to their heads.

To sum up, on the one hand, zero-marked and explicitly-marked phrases are determined by the availability of spoken material. On the other hand, ambiguous sentences do not have a large effect on that-realisation. Additionally, primes with that influence target sentences, but only when the prime includes a syntactical and lexical similar that. When looking closer at the sentence itself, it appears that phrases which are
dependent on their heads want to be adjacent to their heads. If this is not the case, a complementiser is strongly favoured. In other words, when there is intervening material between the matrix clause and the complement clause, a *that* is more likely to be added. This matter is explained in more detail in the following section, which focusses on multifactorial studies. These studies each consider several factors within the main clause or the complement clause of sentences which can include an optional *that*.

2.2.2 MULTIFACTORIAL CORPUS STUDIES ON *THAT*-REALISATION IN ORIGINAL ENGLISH

The principles mentioned above were further researched in multifactorial corpus studies. According to Rohdenburg’s complexity principle (1996), more complex complement clause sentences should prefer *that* since they are more difficult to process if a zero-complementiser is used. Several researchers tried to identify complexity factors in complement clause sentences and scrutinised whether these factors influenced the choice between *that* and *zero*. According to Rohdenburg (1998) there are two major complexity factors: intervening material between the matrix clause and the complement clause and the subject of the complement clause. Other researches (Jaeger, 2010 and Torres Cacoullos & Walker, 2009) added subject of the main clause and tense of the matrix verb to these complexity factors. Next to complexity factors, also the factors frequency of the matrix verb and register (spoken or written mode) were examined. What follows is a brief summary of the researches into these factors and the main findings concerning each factor.

Four multifactorial researches are presented here, which discuss the mechanisms that determine the variation between *zero* and *that* in spoken English. Multifactorial refers to “having, involving, or produced by a variety of elements or causes” (Merriam-Webster Online Dictionary), which means that several factors determine individually or together the *that*/zero alternation. Each study takes into account a different type of spoken English. Tagliamonte & Smith (2004) analysed their factors in an archive of British dialects, while Torres Cacoullos & Walker (2009) focussed on spoken Canadian English. Both Thompson & Mulac (1991) and Jaeger (2010) concentrated on spontaneous speech in respectively American English and all varieties of English.

The first factor considered is the **frequency of the matrix verb**, that is how frequent a verb occurs with complement clause constructions in the corpus. In Tagliamonte & Smith (2004), the most frequent matrix verbs favoured *zero* more than other matrix verbs. Also Thompson & Mulac (1991) and Jaeger (2010) observed similar tendencies in their research. However, the analysis of Torres Cacoullos & Walker (2009)
indicated that the frequency of the main verb did not have a clear effect on the choice between that and zero. Yet, when the idea of frequency was extended to frequent collocations (i.e. subject + verb), the rate of that was significantly lower for those frequent collocations than for the infrequent collocations. The most frequent collocation was ‘I think’, followed by ‘I guess’, ‘I remember’ and ‘I find’. Interestingly, all these frequent collocations have a first-person singular subject. The latter observation can be linked to the influence of the matrix clause subject on that-explicitation.

The subject of the main clause is the second factor which has been frequently researched. According to Jaeger (2010), the complexity of the matrix subject affects that-realisation in a sense that more complex subjects such as noun phrases prefer that more frequently than simple first-person singular subjects. When the matrix subject is a first-person singular pronoun, a zero-complementiser is preferred in 60% of the cases, and as a result, it is the most attractive subject for that-omission (Tagliamonte & Smith, 2004). Jaeger (2010) aligns with this observation adding that the second-person-singular you is the second most attractive subject for the omission of that, followed by other pronouns and lexical noun phrases. Nonetheless, Torres Cacoullos & Walker (2009) argue that the rate from subjects of the main clause favouring to disfavouring that goes: noun phrase, you, other pronoun, I. Although the order in which the subjects prefer that or zero differs slightly, all researches agree that I is most likely to occurs with a zero-complementiser.

Thompson & Mulac (1991) showed that a zero-complementiser tends to occur frequently in a combination of the most frequent main verbs (think, guess, know, mean, …) and the most frequent main subjects (I and to a lesser extent you). The fact that these collocations favour zero confirms Thompson & Mulac’s hypothesis which posits that in contexts with these frequent collocations, the matrix clause no longer exists but rather functions as an epistemic phrase. “Epistemic phrases are phrases that clarify the speaker’s stance in relation to what is being said. A major question addressed through phrases is whether or to what extent the speaker is claiming knowledge (I know)” (Avantlive, 2010). Additionally, Torres Cacoullos & Walker (2009) tested whether the frequent collocations acted as autonomous discourse formulas by adding adverbials in the matrix clause. If they acted as autonomous discourse formulas, the factors conditioning that should be different for these forms than for regular matrix clauses. They concluded that “highly frequent subject-verb collocations behave as fixed units, since that is most likely to occur precisely when the formulaic nature of the collocation is annulled by a co-occurring adverbial” (Torres Cacoullos & Walker, 2009, p. 19). In conclusion, when frequent matrix verbs are combined with frequent matrix subjects, it is more likely that that will be omitted because of epistemicity.
It seems that next to the verb and subject of the main clause, also other elements in the main clause should be taken into account. Firstly, the tense of the matrix verb seems to influence that-realisation. When the tense of the matrix verb is the present simple, more than half of the time zero is preferred (Tagliamonte and Smith, 2004). However, Torres Cacoullos & Walker (2009) argue that non-finite forms of the matrix verb prefer that since it decreases the capability of the main subject and main verb to act as an epistemic phrase or because the non-finite form includes more syntactic complexity. Secondly, Torres Cacoullos & Walker (2009) also scrutinised whether adverbials in the main clause affected that-realisation. They found that only when the adverbial appears after the subject, a zero-complementiser is disfavoured. According to Thompson & Mulac (1991), adverbials or other elements that occur between the main subject and the main verb reduce the ability for them to function as an epistemic phrase and therefore the complementiser tends to be included.

As already noted by Hawkins (2003) zero-marked complement clauses depend more on the matrix clause and therefore favour to be adjacent. Intervening material between the matrix clause and the complement clause results in a tendency towards using that since it is more difficult to identify the relation between both clauses (Tagliamonte & Smith, 2004). This inclination is also supported by Jaeger (2010) and Torres Cacoullos & Walker (2009), who argue that intervening material is one of the most significant factors for the choice between that and zero as it is more difficult to process the sentence.

Besides the factor intervening material, also the subject of the complement clause is regarded as one of the most influential elements for that-realisation (Torres Cacoullos & Walker, 2009). When the subject of the complement clause is a pronoun, the use of zero is preferred because the subject is considered topic of the discourse (Thompson & Mulac, 1991 and Tagliamonte & Smith, 2004). Torres Cacoullos & Walker (2009) specified the different types of subjects and ranked the complement clause subjects from most favourable to least favourable to the complementiser zero. The most favourable complement clause subject is it/there, followed by I. Other pronouns and noun phrases as complement clause subjects favour that.

Both Jaeger (2010) and Torres Cacoullos & Walker (2009) examined whether co-referentiality between the matrix subject and the complement clause subject influenced the that-rate. In both studies, however, this factor did not turn out significant. Furthermore, co-temporality or similar polarity between matrix clause and complement clauses appeared to be not significant for the addition or omission of that (Torres Cacoullos & Walker, 2009). Thompson & Mulac (1991) concluded that also interrogation, negation or passive morphology of the main clause were not significant for predicting that-complementation.
In conclusion, shorter and easier sentences with frequent matrix verbs in the present tense, first-person-singular matrix subjects, it/there as the complement clause subject and no intervening material between matrix clause and complement clause are easier to process. As a result, there is a greater likelihood to omit that in these sentences. The study of Jaeger (2010) sums up that when the complement clause onset carries too much information, complementiser that would be favoured. By inserting the complementiser, the information which would be carried by the complement clause onset is made explicit.

Most of the studies on optional that-explicitation are conducted using spoken corpora. Nevertheless, that occurs more frequently in written English than in spoken English (Shank, Plevoets & Van Bogaert, 2016). In fact, Shank, Plevoets & Van Bogaert (2016, p. 229) indicated that the difference between spoken and written English plays a more important role on the zero/that alternation than one would expect since “the strength of various other factors depends heavily on mode, i.e. some factors may be better predictors of the zero form in one mode as opposed to the other”. For instance, the factor intervening material between main clause and complement clause can better predict the presence or absence of that in a written mode than in a spoken mode in a sense that the that-rate will be higher in the written mode when intervening material is present (Shank, Plevoets & Van Bogaert, 2016).

In the domain of language learning, Wulf, Gries and Lester (to appear) conducted a multivariate study to investigate which factors condition that-realisation for intermediate-level German and Spanish L2 learners of English and if these factors differed much from the ones governing native speakers’ choices. They examined both written and spoken corpora. The results indicated that in general, the “intermediate-advanced German and Spanish learners are quite well aligned with the native speakers’ norms overall” (Wulf, Gries and Lester, to appear, p. 115). Yet, the L2 learners adopt a more conservative attitude towards the omission of that since they tend to only leave out the complementiser in ideal contexts. Based on the results of Wulf, Gries and Lester’s research, this study hypothesises that the differences in that/zero alternation between translators and non-translators will be minimal.

2.3 Multifactorial corpus studies on that-realisation in translated English

That-realisation has been widely researched in various study domains. However, it seems that in translation studies the focus lays only on the difference between original English and translated English while in variational linguistics and language learning, the main topic of research consists of the effect of other factors, as operationalizations of higher-order principles such as Rohndenburgh’s complexity
principle. These factors appear to have a significant influence on the decision between *that* and *zero* in original English. Yet, in translation studies, the effect of these factors still remains poorly examined. Only the recent study of Kruger (2018) integrates the factors conditioning *that* in original English into translation studies. In her research, Kruger (2018) aims to verify common explanations for the tendency towards the mention of *that* in translated English by means of a multivariate analysis of the factors conditioning *that*-realisation in four corpora. The corpora that were used were: a corpus of English translated from Afrikaans, a corpus of written Afrikaans, corpora of written British English and corpora of native South African English. The multivariate analysis she used consisted of a logistic regression analysis and conditional inference trees, which is a related regression-based method.

The factors Kruger (2018) chose to research were register, corpus, matrix verb semantic class, matrix verb lemma-in-construction frequency, matrix clause verb tense and modality, matrix clause verb aspect, matrix clause subject, matrix clause polarity and distance between the matrix clause verb and the onset of complement clause. She uses the analysis of these factors to corroborate the following three hypotheses:

- **The transfer hypothesis:** next to English, the Afrikaans language also offers the choice between *that* and *zero* in complement clauses. Therefore, the transfer hypothesis presumes that the higher the *zero*-rate is in Afrikaans, the higher the *zero*-rate will be in its translations into English.

- **The cognitive complexity hypothesis:** *that* omission or inclusion is based on complexity-related factors in a sense that the complementiser tends to be included in more complex sentences due to an increased cognitive effort. The cognitive complexity hypothesis argues that in translated texts these complexity-related factors will be more important for the *that*/*zero* alternation than in non-translated texts.

- **The risk-aversion hypothesis:** translators have an inclination towards explicitation in order to avoid misunderstandings, disregarding the factors that favour a *zero*-complementiser. As a result, translated English should respect less register and frequency effects than non-translated English since translators always have a tendency to opt for the safe option and include *that* in any register and with any verb.

The results of Kruger (2018) indicated that the transfer hypothesis could be ruled out since translated English texts used the complementiser *that* more frequently than their corresponding original Afrikaans texts or than original English texts. In order to scrutinise the other hypotheses, the factors mentioned above were analysed to see whether they had the same effect in non-translated English and in translated
English. As a result of the analysis, only the factors register, semantic class, lemma-in-construction frequency, subject, corpus, matrix verb lemma-in-construction frequency and distance between matrix clause verb and onset of complement clause were included for the creation of the conditional inference tree. The multivariate analysis shows that the factors determining the use of *that* and *zero* are in large part similar for translated and non-translated English. Complexity factors, such as intervening material between matrix clause and complement clause and more complex subjects, procure a higher *that*-rate in both translated English and non-translated English, confirming the cognitive complexity hypothesis. To test the risk-aversion hypothesis, the factors register and lemma-in-construction frequency were taken into account. The results revealed that translators tend to be less sensitive to register, meaning that even in registers that prefer the omission of *that*, the complementiser was mentioned. This tendency could be extended to the cognitive complexity hypothesis in the sense that even in cognitive less complex sentences, translators opt for the complementiser.

The present research is largely based on Kruger (2018) and can be seen as an extension to her research. The same factors conditioning *that*-realisation will be analysed. Using the MuPDAR approach, it will be examined which factors determine *that*-explicitation and whether these factors are the same for both translated and non-translated English. In this research another statistical analysis is adopted since the statistical analysis used in Kruger (2018) has the limitation of leaving out “the contributions of individuals’ idiosyncrasy to the observed variation” (Paolillo, 2013, p. 89). Therefore, in this research it is opted for the use of the MuPDAR approach which is more suitable for detecting multivariate differences between translated English and non-translated English. “The MuPDAR approach outperforms classical techniques such as ‘plain’ logistic regression and generalised linear mixed-effects modelling in answering questions about differences in several linguistic varieties, since it is better at pointing out where exactly varieties diverge” (Kruger & De Sutter, to appear). Another important difference between this study and Kruger (2018) is that this research considers translations from Dutch into English instead of Afrikaans into English. The Dutch language is interesting to probe since *that*-complementation is different compared to Afrikaans or English. In Dutch *that* (Nl: dat) in complement clauses is obligatory, while in both Afrikaans and English it is optional. According to Kruger (2018) the omission rate of *that* for Afrikaans is higher than for English. For this reason, the transfer hypothesis as known in Kruger (2018) cannot be considered in this research. However, the difference in *that*-realisation between English and Dutch leads to a new transfer hypothesis: translations from Dutch into English will have a higher *that*-rate as a result of copying the Dutch obligatory complementiser into English. This *that*-rate will then be compared to the one of Kruger (2018). The
cognitive complexity hypothesis and the risk-aversion hypothesis will be tested in a similar way to the study of Kruger (2018).

3 Methodology

3.1 Corpus

The corpus used for this study is the Dutch Parallel Corpus (in short: DPC), which is a high-quality corpus that contains over 10 million words and was created within the STEVIN programme by the University of Leuven (Campus Kortrijk) and the Faculty of Translation Studies of Ghent University College. It is a sentence-aligned corpus which provides sentences of the following translation directions: English – Dutch, Dutch – English, French – Dutch and Dutch – French.

Macken, De Clercq & Paulussen (2011) set out the details and the advantages of the DPC. Now a brief overview of the advantages for this particular study is presented. Firstly, the main objective of the creation of the DPC was to provide a representative and balanced corpus. As a result, the corpus texts were divided into five text types: literature, administrative texts, instructive texts, journalistic texts and texts for external communication. As mentioned above, there are four translation directions incorporated which “implies that each text type should contain about two million words and that within each text type each translation direction should contain about 500,000 words” (Macken, De Clercq & Paulussen, 2011, p. 378).

Secondly, two common forms of corpus annotation were added. The first one is part-of-speech tagging, which means that the morpho-syntactic class (noun, verb, adjective, etc) of each word is determined. The second one is lemmatisation, that is the process which generates the base form of each orthographic token. For instance, the base form of a verb is the infinitive and the base form of a noun is its singular form. This facilitates the research process because one only has to search for the base form in order to find all forms of the verb or noun. Thirdly, the corpus is well-structured in order to improve the data exploitation. For instance, “the source and the target texts are coded as monolingual XML files and contain the annotated sentences; the sentence alignments are stored in a separate index file, in which the indexes point to the sentences of the monolingual files” (Macken, De Clercq & Paulussen, 2011, p. 385).

Furthermore, the corpus is easily accessible via http://dpcserv.ugent.be/comure.
3.2 Data extraction

Since the corpus consists of translations from Dutch to English and vice versa and from Dutch to French and vice versa, the French part was excluded. As a result, the sentences were extracted from that part of the corpus which only included the translations from Dutch to English and from English to Dutch. Of the latter translation direction only the original English texts were used since the Dutch translations were irrelevant for this research. There was an option to search only in the database for American English or British English. However, this study included both as there was no intention of examining the difference between both varieties. The search terms used were all the private and public verbs listed by Quirke et al. (1985). These verbs are regarded as verbs that are frequently followed by complement clauses and therefore they are very suitable for this research. In total, 123 verbs were searched, that is when the American and British way of writing the verb was considered different. For instance, the verbs ‘recognise’ and ‘recognize’ were seen as separate verbs. In the search term bar, the infinitive form of the verb was filled in in order to extract all the sentences which incorporated this term. Thanks to the lemmatisation included in the corpus, all the other forms of the verb were immediately searched for as well. As a result, all the sentences of the corpus which included any form of the verb were shown and could be exported into an excel file. In the excel file, the sentence itself was given together with all the extra information that was annotated in the corpus. The only downside was that when a given verb form could also be regarded as a noun, both all forms of the verb and all forms of the noun were comprised in the excel file. This is an indication that the part-of-speech tagging was not always executed properly in the corpus since this would normally ensure that verbs and nouns would be categorised separately. Yet, the data had to be further annotated manually so that irrelevant sentences could be discarded. Every private or public verb could be found in the DPC, except for the verb ‘to prophesy’, which produced no search results. In the next section, the dependent and independent variables are discussed.

3.3 Variables

3.3.1 Dependent variable

One of the main goals of this research is to determine which factors condition that-realisation in original English and translated English. Consequently, the dependent variable is whether a complement clause includes a that-complementiser or a zero-complementiser. After exporting the excel file for a particular verb, an extra column ‘complementiser’ was added in the document and all sentences were manually
checked on relevance. Only the sentences which contained complement clauses that could be introduced by the complementiser *that* were considered relevant. For the relevant sentences, the column ‘complementiser’ was filled in. When the sentence involved the complementiser, ‘yes’ was written and when the sentence did not involve the complementiser, ‘no’ was written. When all sentences were inspected, the relevant sentences of each verb were put together in one excel file in order to further analyse them in RStudio.

Only the sentences in which the matrix clause was in initial position and a complement clause followed were considered relevant sentences. The following sentences were considered irrelevant and therefore they were not further annotated (ICE-GB, ICE-SA & TR used in Kruger, 2018):

- Sentences with the matrix clause in final position: ‘It was almost like being at a funeral, Vaselinetjie thought.’
- Sentences with the matrix clause in medial position: ‘Life, he eventually decided, was a verse you had to learn over and over until you knew the words by heart and started to understand them.’
- Cleft sentences: ‘Of course what I found was that almost all his poetry is filled with expressions of his need for and love of other people.’
- Wh-extractions: ‘So what do you think JP was doing on the tippy-top of the Petronas Towers.’
- Relative clauses: ‘In a different context Hambidge has noted what she refers to as “a blueprint for all travel experiences”, namely “that which one thinks one knows about a country, that which one ultimately experiences, and that which one will never really be able to...”’
- Direct speeches: ‘He said she could just put any date that suited, and added, ”You can just send the form to me, personal.”’
- Non-verbal use of the verb: ‘The political thought and legal theory of the late medieval period and the Renaissance, which are hardly distinguishable from one another,...’

### 3.3.2 INDEPENDENT VARIABLES

An independent variable is “a factor or phenomenon that causes or influences another associated factor or phenomenon called a dependent variable” (Business dictionary online). The independent variables of this study are similar to the ones of Kruger (2018) which allows comparison with her research. The variables consist of both numerical and categorical variables. What follows is a brief summary of the
different variables and the corresponding levels; in general, the guidelines found in Kruger (2018) were followed.

Two of the factors which are investigated are numerical variables. The first one is the distance between the matrix clause and the onset of the complement clause. The distance was calculated not by the number of words but by the number of characters. This was done by manually placing a vertical bar after the matrix clause verb and another one just before the onset of the complement clause so that the distance could be calculated in excel. The second one is matrix verb lemma-in-construction, which examines the frequency of the verb within complement clauses constructions. These frequencies were copied from the research of Kruger (2018) and added to the excel file.

The other factors examined are categorical variables. Firstly, the factor tensemodality was divided into present, past, non-finite, modal and quasi-modal. Secondly, the factor aspect was broken down into simple, perfect and progressive. Thirdly, the divisions of the factor subject were: noun, expletive it, no overt subject, clause, pronoun and relative pronoun. Fourthly, the levels of factor polarity were positive and negative. Fifthly, the factor text type consisted of the levels administrative texts, instructive texts, fictional literature, non-fictional literature, external communication and journalistic texts. The last factor examined was the semantic class of the matrix verb. Just as Kruger (2018), this category was split into private verbs, public verbs and suasive verbs according to the classification of Quirk (et al., 1985). However, due to the extent of this research and the time limit, it was not accomplishable to investigate all the sentences in the DPC which contained one of these verbs. Consequently, it was decided to focus only on the private and the public verbs.

In order to include the former factors, a column was added to the excel file and the levels corresponding to each sentence were manually filled in. Only the factors and factor levels of text type were already included in the DPC. Table 1 below presents an overview of all the factors and the corresponding levels.

---

3 The possible modality of the main verb prevailed over the tense of the main verb. Consequently, if the main verb consisted of a quasi-modal, a modal or a non-finite form, only these factor levels were included in the column of tensemodality and the tense of the verb was not included. Conversely, the factor levels present and past were only included when the main verb did not consist of a quasi-modal, a modal or a non-finite form.
<table>
<thead>
<tr>
<th>Factor</th>
<th>Factor levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance between the matrix clause and the onset of the complement clause</td>
<td>A numeric variable of the number of characters between the end of the matrix clause and the onset of the complement clause.</td>
</tr>
<tr>
<td>Matrix verb lemma-in-construction frequency</td>
<td>A numeric variable of the frequency of the matrix verb in complement clauses, per 100 000 words.</td>
</tr>
</tbody>
</table>
| Tense and modality of the matrix clause                             | • Present  
  • Past  
  • Non-finite  
  • Modal  
  • Quasi-modal                                                   |
| Aspect of the matrix clause                                         | • Simple  
  • Perfect  
  • Progressive                                                    |
| Subject of the matrix clause                                        | • Noun  
  • Expletive it  
  • No overt subject  
  • Clause  
  • Pronoun  
  • Relative pronoun                                                  |
| Polarity of the matrix clause                                       | • Positive  
  • Negative                                                                 |
| Text type                                                            | • Administrative texts  
  • Instructive texts  
  • Fictional literature  
  • Non-fictional literature  
  • External communication  
  • Journalistic texts                                                |
| Semantic class of the matrix verb                                   | • Public verbs  
  • Private verbs                                                      |

*Table 1. Overview of the factors with the corresponding factor levels*
3.4 Statistical analysis

To analyse the influence of the above mentioned factors on that-realisation in original English and translated English, the MuPDAR approach was applied. MuPDAR stands for Multifactorial Prediction and Deviance Analysis with Regressions and was established by Stefan Th. Gries and fellow researchers. In Gries & Deshors (2015) the approach is explained in depth.

The first step of the analysis was importing both datasets (translated and non-translated English) and checking the frequency distributions of all factor levels. The translated dataset consisted of 2866 sentences and the non-translated dataset comprised 2910 sentences. If a factor level did not occur sufficiently in the dataset or if there was a huge difference in numbers between the factor levels, some adjustments had to be made. Both numeric variables matrix verb lemma-in-construction and the distance between the matrix verb and the onset of the complement clause were very skewed. Therefore, their distributions were logarithmically transformed in order to obtain a normal distribution. Concerning the categorical variables, some sparseness issues were detected and consequently some levels needed to be put together and covered by a more general term. The following adjustments were made:

- For the factor aspect, the level simple was highly frequent, while the levels perfect and progressive occurred in smaller numbers. More specifically, perfect occurred 134 times in the non-translated dataset and 222 times in the translated dataset while progressive occurred 89 times in the non-translated dataset and 139 times in the translated dataset. As a result, it was decided to combine both levels in perfectprogr as they are both more complex verb forms.

- When inspecting the factor subject, it was clear that the level relative pronoun was infrequent since it only occurred 49 times in the non-translated dataset and 96 times in the translated dataset. Consequently, relative pronouns were no longer considered a separate level and they were integrated in the more general level pronoun. Also the level clause appeared in low frequencies, that is 27 times in the non-translated dataset and 44 in the translated dataset. However, no other factor level seemed to have enough in common to be put together with clause. For this reason, it was decided to omit all sentences including a clause as the subject of the matrix clause and to not further investigate this factor level.

- The frequency distribution of the levels of the factor tensemodality also contained some sparseness issues. Firstly, there were very few frequencies of the level quasi-modal (6 occurrences in the non-translated dataset and 7 occurrences in the translated dataset) and therefore it was combined with the factor level modal. Moreover, imperative did not occur sufficiently because
this level only occurred 13 times in the non-translated dataset and 16 times in the translated dataset. As a result, this level was integrated in the level present.

- The level non-fictional literature of the factor text type was sufficiently represented in the non-translated dataset. However, in the translated dataset, this factor level did not occur. In order to solve this issue, the level non-fictional literature was integrated in the level journalistic texts, since both text types provide factual and objective information to the audience. Furthermore, the representation of the level instructive texts was very poor in both datasets with 136 occurrences in the non-translated dataset and 26 occurrences in the translated dataset. Therefore, it was decided to include this factor level in the level administrative texts; Delaere & De Sutter (2017) have convincingly argued that most of the instructive texts in the DPC are in fact administrative in nature.

The second step was the fitting of a generalised linear mixed-effects model (in short: glmm) on the non-translated data, since this variety is regarded as the prototypical variety and therefore seen as the point of reference in this research. This model was fitted to predict that-realisation based on the different factors of the non-translated dataset. Consequently, the addition or omission of that was the dependent variable and the other factors mentioned above were the predictor variables. Moreover, the factors File and Lemma (in this research referred to as Benaming) were integrated as random factors to control for specificities coming from the same writer or company and the different lemmas. According to Kruger and De Sutter (to appear) “the possibility of including random factors alongside fixed factors (i.e. the already mentioned predictor variables) is what distinguishes glmm from ‘classical’ logistic regression analysis”. The model was fitted by adding the predictor variables one by one. First, only one variable was added and it was checked whether one or several levels were significant for the prediction of that-complementation. If this was the case, the variable was included in the model and if not, the variable was rejected. When adding the next variables, the same method was used. Thus, the generalised linear mixed-effects model was fitted only by using the predictor variables which had a significant influence on that-realisation.

In the third step, the generalised linear mixed-effects model based on the non-translated data was applied to the translated dataset. The coefficients that were measured for each factor incorporated in the generalised linear mixed-effects model were used as predictors for that-realisation in the translated data. It was attempted to determine whether translators make deviant choices compared to non-translators by means of checking whether the generalised linear mixed-effects model could predict the inclusion or omission of that correctly for the translated dataset. If the predicted choices would be identical to the
actual choices, both varieties (translated and non-translated) would decide on the inclusion of *that* or *zero* in the exact same way. If the predicted choices would deviate from the actual choices, it would be concluded that translators have a tendency to make – to some extent – different choices opposed to non-translators. Deviance is seen as the difference between the predicted choice and the actual choice. When the predicted choice is identical to the actual choice for a particular sentence, the deviance is 0. However, when the predicted choice is distinct from the actual choice, the deviance is 1. The total number of the individual differences divided by the sum of the sentences is the overall deviance of the translated data opposed to the non-translated data.

Finally, another glmm model was fitted based on the translated dataset and the results of the above mentioned deviance test. This model measures the effect of the factors included in this research on a new dependent variable (i.e. the correctness of the predicted choice). The aim was to discover which factors prevent the predicted choice from being correct. In other words, “which predictors increase or decrease the likelihood of getting identical choices as in the central, prototypical variety” (Kruger & De Sutter, to appear). It was determined which factors prevented the predicted choice from being correct by adding each factor to the model and looking at the AICc score. When the score lowered by adding a particular factor, this meant that the factor reduced the likelihood of the predicted choice from being correct. Nonetheless, when the AICc score remained the same or increased when adding the factor, this factor did not prevent the predicted choice from being correct. After determining which factors tend to influence the correctness of the predicted choice, it was investigated which specific factor levels procure a deviant choice. Moreover, it was examined in which direction the choices were deviant, that is checking whether there is an bias towards using a *zero*-complementiser or a *that*-complementiser in translated English.

In conclusion, the MuPDAR approach used in this study comprised three principal parts. The first one was the fitting of a generalised linear mixed-effects model based on the non-translated dataset. Secondly, this model was applied to the translated dataset in order to verify whether the model could predict *that*-realisation correctly for translated sentences. Ultimately, it was investigated for the translated data which factors influence a correct prediction and lead to deviant results.

The statistical analysis was performed with R 3.3.2 (R Core Team 2016) using the packages *car, effects, MuMIn, MASS, lme4* and *Hmisc*. The cut-off point for the statistical significance was established at the 0.05-level.
4 Results

In this section, the results of the MuPDAR approach will be discussed. First, the general linear mixed-effects model, which uncovers the significant factors for that-realisation in non-translated English, is presented in 4.1. In section 4.2, we will find out to what extent the significant factors in non-translated English are capable of capturing that-variation in translated English as well; in other words, to what extent the choices for that or zero are similar for translated and non-translated English. Ultimately, the factor levels that lead to deviant choices for translated English will be displayed in 4.3 and the direction of the deviant choices will be discussed in 4.4.

The results will provide answers to the two research questions of this study:

1. In what respect (and to what extent) does the choice between that and zero in translated English differ from non-translated English?
2. Which factors tend to be more decisive for the choice between that and zero and which factors are less important in non-translated English? And to what extent do these factors have a different effect for translated English?

Based on the results of the research of Wulf, Gries & Lester (to appear), this study hypothesises that the differences in that/zero alternation between translators and non-translators will be minimal. Wulf, Gries & Lester (to appear) indicated that in general the choice between that and zero of German and Spanish L2 learners of English aligns quite well with the choices of native speakers. Consequently, this study believes that the same tendency will be found when comparing translators to non-translators.

The results will also provide insight into the correctness of the three hypotheses formulated by Kruger (2018) which are presented in section 2.4 of this study: the transfer hypothesis, the cognitive complexity hypothesis and the risk-aversion hypothesis. The transfer hypothesis is tested by examining whether the higher rate of that in Dutch leads to a higher rate of that in translated English. The cognitive complexity hypothesis is tested by investigating whether more complex sentences have a greater preference towards including the complementiser. The risk-aversion hypothesis is tested by scrutinising whether translators opt for that regardless of register or frequency effects.
4.1 Factors influencing that-realisation in non-translated English

The first step of examining if that-realisation is different in translated and non-translated English is the fitting of a general linear mixed-effects model (or R1 model) on the non-translated data. This model is created on the basis of the factors which influence the omission or inclusion of that for original English. The factors scrutinised in this study are the distance between the matrix clause and the onset of the complement (in the statistical analysis referred to as: \textit{MCVerbToCCLength}), matrix verb lemma-in-construction frequency (in the statistical analysis referred to as: \textit{LemmaConstrFreq100kLOG}), the type of the matrix verb, the aspect of the matrix verb, the tense and modality of the matrix verb, the subject of the main clause, the polarity of the main clause and the text type. Firstly, we added each factor incrementally, on a one-by-one basis, thereby excluding insignificant factors immediately. Afterwards, the accuracy and the predictive ability of the model was computed.

The full statistical report on the fitting of the R1 model can be found in appendix A (p. 40). The plots below represent the significant factors for the addition or omission of the complementiser that. The grey band in the plots of the numerical variables and the red lines in the plots of the categorical variables point out the boundaries of the 95% confidence interval.

\textbf{Figure 1.} Plots of the significant factors for that-realisation in non-translated English: \textit{MCVerbToCCLength}, \textit{LemmaConstrFreq100kLOG}, aspect, subject, text type and tense modality

First of all, it has to be noted that in non-translated written English there is a considerable tendency of incorporating that in complement clauses since the probability of using that in all plots depicted in Figure 1 approaches 1.0. The plots have to be considered as follows: the probability of that is 100% for 1.0 and the probability of zero is 100% for 0.0. For example, when the factor \textit{MCVerbToCCLength} is 0 (i.e. there is
no intervening material between the matrix clause and the onset of the complement clause), the probability of *that* is approximately 80%. Now a brief discussion on the influence of each factor on *that*-realisation is presented.

The first plot, which is the left plot of the top row, represents the effect of the factor $MCVerbToCCLength$. The plot indicates that when there is a larger distance between the matrix clause and the onset of the complement clause, there is a greater tendency to use *that*. This can be explained by the fact that intervening material makes the sentences more complex and more difficult to process (Kruger, 2018). Consequently, a complementiser is preferred to make the dependency relations more explicit and as a result the sentences will be easier to process (Hawkins, 2003). When there is no intervening material between the main clause and the onset of the complement clause, *that* is preferred in 80% of the cases. However, in the other 20% *zero* is favoured. The second plot, which is the middle plot of the top row, shows that the more frequently a verb occurs with complement clauses, the more likely it is that a *zero*-complementiser is used. When the main verb is used very frequently with a complement clause construction, *zero* is used in 20% of the cases. These observations are in line with previous studies (Tagliamonte & Smith, 2004; Thompson & Mulac, 1991 and Jaeger, 2010). The plot on the right side of the top row demonstrates that both the *simple aspect* as well as the *perfect or progressive aspect* most frequently prefer a complementiser. However, it has to be observed that the *simple aspect* has a slight preference towards omitting *that* since *that* only occurs in 90% of the sentences. When the main verb occurs in the *simple aspect*, this can be regarded as a less complex verb construction and therefore the probability of *zero* is higher (Kruger, 2018). The left plot of the lower row points out that an expletive *it* as subject has the greatest preference for including *that* while a *pronoun* as subject has the smallest tendency towards using the complementiser (in approximately 15% *zero* is preferred). This is in line with previous researches on the influence of the subject on *that*-realisation (Tagliamonte & Smith, 2004; Torres Cacoullos & Walker, 2009 and Jaeger, 2010), which already concluded that the pronoun *I* occurred most frequently with zero. Both a *noun* as subject and *no overt subjects* prefer a *that*-complementiser in circa 90% of the cases. Nevertheless, in some cases (approximately 10%) a *zero*-complementiser is opted for.

The middle plot of the lower row represents the factor *text type*. It is clear that in both *administrative texts* and *external communication* that is used more frequently than in *journalistic texts* and *fictional literature*. A plausible reason for this observation is that *administrative texts* and *external communication* are considered as more formal text types whereas *journalistic texts* and *fictional literature* are more open to the creativeness of the writer (Kruger, 2018). Finally, the last plot represents the factor *tense modality*. Both *modal forms* and *non-finite forms* tend to add the complementiser whereas the *past tense* and to a
lesser extent the present tense are less inclined to always incorporate that. More specifically, when the main verb is in the past tense, zero is favoured in nearly 15% and when the main verb is in the present tense, zero is preferred in almost 10%. This observation can be justified by the fact that verb forms in the present tense or past tense can be regarded as less complex compared to modal forms and non-finite forms of the matrix verb (Kruger & De Sutter, to appear). In conclusion, that-complementation is affected by complexity factors (MCVerbToCCLength, subject, aspect and tensemodality) in a sense that the more complex a sentence is, the more likely it is that a that-complementiser will be added. Moreover, other factors such as text type and LemmaConstrFreq100kLOG also seem to play a significant role for that-realisation.

It appeared that all factors are significant except for the factors polarity and type of the main verb. Two conclusions can be drawn based on this observation. First, it can be confirmed that that/zero alternation is similar for a positive and negative main clause, which is in line with Kruger (2018) as this factor did not result significant in her research either. Second, it can be concluded that there is no difference in that-complementation for public verbs and private verbs. This, however, does not align with Kruger (2018) since this factor was significant in her research.

The accuracy of the estimates in the R1 model is very high since it aligns in 83% of the cases with what was actually observed. Moreover, the model is significantly better compared to the null model (p < 0.001), i.e. “the model in which that-variation is modelled on the basis of no predictors at all, and which only selects the most frequent outcome in the response variable” (Kruger & De Sutter, to appear). Next to the accuracy of the R1 model, the predictive ability is also very good with a c-value of 0.90 (c-values range from 0.5 to 1; c-values > 0.8 are considered good predictive models). This means that this model is highly capable of predicting the presence or absence of that in unseen data. In the next section, the results of the application of the R1 model on the translated dataset will be discussed.

4.2 The application of R1 to the translated dataset

In order to examine whether that-realisation is the same for translated and non-translated English, the R1 model (based on the factors determining that-complementation in non-translated English) is applied to the translated dataset. Consequently, this model will predict whether a that or a zero is used in the translated sentences. If the R1 model can predict correctly when that or zero is written, it can be concluded
that both varieties (translated and non-translated English) are very similar. If the R1 model cannot predict
the *that*/zero alternation correctly, both varieties handle the use of the complementiser *that* differently.

Results indicate that in 82% of the cases, the R1 model can predict correctly whether a *that*-
complementiser or a *zero*-complementiser is included. More specifically in 2347 of the 2865 translated
sentences, the prediction model indicates *that*-realisation correctly. This means that, in general, *that*-
complementation is very similar for translated and non-translated English. Moreover, both varieties take
into account the same factors when it comes to adding or omitting the complementiser. Although both
varieties are similar in 82% of the times, there is still an important 18% in which the model predicted the
*that*/zero alternation incorrectly. In other words, there are still some differences between both varieties
which lead to deviation from the prediction of the R1 model. In the following section a detailed discussion
on the factors which lead to deviant choices for *that*-realisation in translated English will be introduced.

4.3 Factors leading to deviant choices

A new regression model was fitted which measures the effect of the factors of the R1 model on the
correctness of the predicted choice for translated English in order to find which factors provoke deviation.
Appendix B (p. 41) presents the full statistical analysis for the new regression model, which will be referred
to as R2 in this study. Figure 2 below shows the factors that influence the correctness of the predicted
choice significantly. As can be observed all factors comprised in the R1 model affect the correct choice for
*that*-realisation except for the factor *tense modality*. Similar to Figure 1, the grey band for the numerical
variables and the red lines for the categorical variables point out the boundaries of the 95% confidence
interval.
As can be noticed, not all factors have the same influence on the correctness of the predictive choice. For instance, \texttt{MCVerbToCCLength} and \texttt{aspect} have rather small effects while \texttt{text type}, \texttt{LemmaConstrFreq100kLog} and \texttt{subject} seem to have somewhat larger effects. Firstly, the middle plot of the top row indicates that intervening material between the matrix clause and the complement clause seems to have a slightly significant influence. It appears that only when there is no intervening material (i.e. \texttt{MCVerbTOCCLength} is 0) the predicted choices are deviant. This could be forecast since the AICc value only lowered with two points when adding this factor. Secondly, the factor \texttt{aspect}, presented by the middle plot of the lower row, also seems to have a rather small significant effect on the correctness of the predicted choice. Whereas sentences of which the main verb is written in the \texttt{perfect} or \texttt{progressive aspect} mostly have similar choices for \texttt{that}-realisation in translated and non-translated English, translators tend to make slightly different choices for sentences which include the \texttt{simple aspect}. Thirdly, the left plot of the top row offers the probability of the deviant choices of the factor \texttt{text type}. Both for \texttt{external communication} and \texttt{administrative texts} \texttt{that}-complementation seems to be similar for translated and non-translated texts. Yet, for the text types \texttt{fictional literature} and \texttt{journalistic texts} it seems that deviant choices are made in translated English opposed to non-translated English. The last but one plot that will be discussed is the right plot of the top row, \texttt{LemmaConstrFreq100kLog}. The more frequently a verb is
used with complement clauses, the more deviant choices are made by translators. For verbs that do not tend to occur frequently with complement clauses, the R1 model could predict the inclusion or omission of *that* easily. However, for verbs that tend to occur mostly with complement clauses, the R1 model predicted in 10% wrongly whether *that* or *zero* would be used. The final plot, which is the left plot of the lower row, indicates the significant influence of the factor *subject* on the correctness of the predicted choice. When an *expletive it* is the subject, no deviant choices are made. In addition, only in few cases deviant choices are observed when the subject is a *noun*. Nonetheless, when there is a *no overt subject* in translated sentences there is a slight tendency towards making different choices compared to non-translated sentences. The most deviant choices are observed when the subject is a *pronoun* since in more than 10% of these sentences the R1 model could not predict correctly whether *zero* or *that* was used.

To sum up, mostly the factors *LemmaConstrFreq100kLog, subject* and *text type* influence the correct or incorrect prediction of the choice between *that* and *zero*. When looking more thoroughly at the level factors, it is clear that particularly verbs which occur frequently with complement clauses, *pronouns* used as subject and *journalistic texts or fictional literature* provoke deviant choices. Interestingly, according to Figure 1, these are also the factor levels which have a greater tendency towards using a *zero*-complementiser in original English. Since it is already discernible which factors lead to deviant choices, the next step is examining in which direction these deviant choices are taken. In other words, it is scrutinised whether the differences between translated and non-translated English in *that*-complementation are in favour of adding or omitting the complementiser. This will be discussed in the following section.

### 4.4 Direction of the deviant choices

The last step consists of examining which direction the deviant choices take, i.e. analysing whether the complementiser *that* or the complementiser *zero* is used more frequently among translators in those contexts. For this matter, plots are established which represent the distribution of the deviation scores as a function of the various factors in order to see which factor levels cause deviance and in which direction the deviance is mainly oriented. If the deviant scores are lower than 0, this means that for that particular factor level the complementiser *that* is used more frequently in translated English than in original English. Conversely, if the deviant scores are higher than 0, this means that for that particular factor level a *zero*-complementiser is used in translated English while in original English it would be opted for *that*. In Figure
3, the plots of all the factors that cause deviance significantly and the direction of these deviant scores are displayed.

![Figure 3](image)

*Figure 3. Plots of the deviant scores of the factors influencing that-realisation in translated English: text type, subject, aspect, MCVeToCCLength and LemmaConstrFreq100kLOG*

Overall, it can be noticed that the deviant scores are mostly smaller than 0, which means that in translated English, there is an inclination towards using *that* while in original English *zero* would be preferred. This tendency is in line with several researches (Olohan & Baker, 2000 and Blum-Kulka, 1986), which already indicated the higher tendency of explicitation in translated English. Nevertheless, the results of this research have already shown that the majority of the choices between *that* and *zero* for translated texts aligns with the ones for non-translated text. Therefore, the inclination towards using *that* in translated English appears to be rather limited and less extreme than Olohan & Baker (2000) claimed. Now, the deviant scores of each factor level will be considered in detail.

The first plot, which is the left one of the top row, portrays the deviant scores of the factor *text type*. In general, most of the times the choices align with the non-translated choices (deviation score = 0). Yet, for *fictional literature* and *journalistic texts*, there is a significant amount of choices towards using *that* where in original English it would be opted for *zero*. In addition, there are also some choices towards using *zero* where *that* is preferred in original English but these choices are minimal. When it comes to *administrative texts* and *external communication*, some deviant choices are observed towards more frequently including a *that*-complementiser but they are sparse. The second plot, which is the middle plot of the top row, depicts the direction of the deviance for the factor *subject*. For the factor levels *no overt subject* and *expletive it*, the rare deviant scores are lower than 0. Furthermore, for *a noun* as subject the deviance is also towards using the complementiser *that* more frequently. However, the most significant factor level is
pronoun, which shows many deviant choices towards that and also a few towards zero. Consequently, translators have the greatest inclination towards explicitation when the subject of the main clause is a pronoun. The third plot is the right plot of the top row indicating the direction of the deviant choices for the factor aspect. There are few deviant choices for the factor level perfectprogr and for the factor level simple, it can be observed that there is a significant number of deviant choices towards using that in translated English whereas in original English zero would be preferred. The next plot, which is the left one of the lower row, indicates the deviation directions for the factor MCVerbToCCLength. Interestingly, almost every time when the distance between the matrix clause and the onset of the complement clause is 0, the complementiser is used in translated English whereas in non-translated English zero is used. As a result, there are more deviant choices than similar choices for translated English and original English when there is no intervening material between the main clause and the onset of the complement clause. The few times when there is intervening material between the main clause and the onset of the complement clause, the choices in that-complementation are the same for both varieties. The last plot that will be discussed is the one of the factor LemmaConstrFreq100kLOG, which is the right plot of the lower row. It can be noted that many choices concerning that-realisation in translated English are deviant from the ones in non-translated English for the factor LemmaConstrFreq100kLOG in a sense that a that-complementiser is added very frequently. In addition, the more frequently a verb tends to occur with complement clauses, the higher the rate of deviant choices towards using that seems to be.

All in all, it can be concluded that the direction of the majority of the deviant choices of all factor levels is towards the inclusion of a that-complementiser. This means that translators have a tendency to make complement clauses explicit whereas in original English texts these complement sentences would be written with a zero-complementiser. Especially when there is no intervening material between the matrix clause and the onset of the complement clause, when the subject of the main clause is a pronoun, when the sentences belong to the text types fictional literature or journalistic text, when the main verb occurs frequently with complement clause constructions or when the aspect of the main clause is simple, it seems that there is a greater inclination towards including that.

5 Discussion

In this section, we try to make sense of the results in the previous chapter by presenting a reflection of the replies to the research questions and verifying the plausibility of the hypotheses.
This study first inquired in what respect and to what extent *that*-realisation is different for translated English and original English. The results demonstrated that both varieties behave similarly since in 82% of the cases the R1 model, which was based on the non-translated data, could predict the inclusion or exclusion of *that* correctly for the translated data. In other words, *that*/zero alternation is in large part based on the same factors for both varieties. This large degree of similarity between translated and non-translated English was also found in Kruger (2018) with regard to the factors conditioning *that*-complementation. In her study, using a different statistical technique, the difference between translated and non-translated English was of secondary importance to understand variable *that*-complementation. Moreover, in Kruger & De Sutter (to appear) it seemed that both varieties (original English and English translated from Afrikaans) were very similar as well, as their model based on the non-translated dataset could correctly predict the choice between *that* and *zero* in 80% of the sentences in the translated dataset.

In spite of the considerable similarity between both varieties, the R1 model could not predict *that*-complementation correctly in 18% of the cases. In the majority of these instances, the prediction model suggested that a *zero*-complementiser would be used while in the translated sentences a *that*-complementiser was found. This shows that translators have a slight tendency towards the explicitation of *that* in complement clauses. This tendency is also confirmed by Kruger (2018), in which differences between translated English and non-translated English were found in terms of a decreased sensitivity to register and a decreased sensitivity to low complexity contexts in translated English. More specifically, in registers where a *zero*-complementiser was preferred in non-translated English, the complement clause’s onset was made explicit in translated English and less complex sentences, which would normally prefer *zero* in original English, also included the complementiser *that* in translated English. Similarly, Kruger & De Sutter (to appear) found that the complementiser tends to be omitted less frequently in translations than in original English texts.

The second research question investigates the factors that are decisive for *that*-realisation in non-translated English and whether these factors generate a different effect in translated English. The factors that influence the choice between *that* and *zero* for non-translated English turned out to be the aspect of the main verb, its tense and modality, its subject, the frequency of the main verb in complement clause constructions, the text type and the distance between the main clause and the onset of the complement clause. In general, most factor levels which show a preference towards using a *zero*-complementiser in original English, demonstrate a slight tendency towards using a *that*-complementiser in translated English. These observations are in line with Kruger (2018) and Kruger & De Sutter (to appear) in a way that
translated English is less sensitive to factor levels which usually lead to the omission of *that*, for example register and reduced complexity. In this study as well as in Kruger (2018) and Kruger & De Sutter (to appear), it can be concluded that the following factor levels tend to have a greater tendency towards using a zero-complementiser in non-translated English: less formal registers (creative writing, reportage, fictional writing and journalistic texts), a pronoun as subject, no intervening material between the main clause and the onset of the complement clause, verbs which frequently occur in complement clause constructions, main verbs in the present tense or past tense and main verbs in the simple aspect. Both Kruger & De Sutter (to appear) and this study pointed out that these factor levels all reveal a slightly higher preference towards using a *that*-complementiser in translated English. The only finding in which this study differs from Kruger & De Sutter (to appear) is the factor tensemodality, which led to a different choice in *that*-realisation for translated and original English in Kruger & De Sutter (to appear). In this study, however, the factor did not turn out to be significant which means that the factor tensemodality does not lead to differences in *that*-complementation for translated and non-translated English. All in all, where zero is preferred with less complex constructions, such as pronouns as subjects and main verbs in the simple aspect, in original English, this inclination is less outspoken in translated English.

Several researchers have tried to provide reasons and explanations for the functioning of the alternation between *that* and zero in translated English and original English. Kruger (2018), for instance, tested three hypotheses to examine the possible reasons for differences in *that*-complementation in English texts translated from Afrikaans compared to original English texts. Now, the same hypotheses will be scrutinized in order to provide insight into *that*-realisation for original English texts compared to English texts translated from Dutch.

In Kruger (2018) and Kruger & De Sutter (to appear) no evidence could be found to support the transfer hypothesis. This means that there is no correlation between the *that/zero* alternation in Afrikaans and in English. In other words, if a zero-complementiser is used in an Afrikaans sentence, which is the default case, this does not result in the use of a zero-complementiser in the corresponding translated English sentence. Instead, the use of *that* was higher in translated English than in Afrikaans or in original English. However, for Dutch sentences translated into English this tendency is different. Since the complementiser is obligatory in Dutch, the higher use of *that* in translations from Dutch to English can be due to the copying of the complementiser. To put it differently, the increased use of a *that*-complementiser in English texts translated from Dutch is possibly due to the influence of the contact variety Dutch. Nevertheless, the higher *that*-rate can also be a result of the risk-avoidance hypothesis which will be discussed below.
Further research will have to indicate whether or not the transfer hypothesis is a plausible explanation for the higher that-rate in English texts translated from Dutch. This can be done by very precisely verifying whether or not a complementiser is present in the source sentence and how this affects the target sentence. All in all, translators tend to favour the addition of that in English translations for both Afrikaans and Dutch as source language. This might indicate that translators do not take into account the source language for the choice between that and zero.

Another element that could influence that-complementation is the complexity of the sentence, which was tested in this research by the cognitive complexity hypothesis. The hypothesis postulates that the more difficult a sentence is to process due to complexity-related factors, the greater the tendency is to include that in order to facilitate the cognitive process for the reader. The results of this research indicated that complexity-related factors have an effect on that-realisation in a sense that the more complex a sentence is, the more likely it is that the complementiser that is included. Additionally, the findings of this research show that in translated English that seems to be preferred in less complex sentences as well. The same observations were made in Kruger (2018) and in Kruger & De Sutter (to appear). Kruger & De Sutter (to appear) offer an explanation which implies that the that-rate is higher in less complex sentences as a result of the extended cognitive effort that is needed in the process of translation. In addition, this can also explain why frequency effects of the main verb and the register of the text have a less outspoken influence on the choice between that and zero in translated English. To sum up, the results of this research seem to support the cognitive complexity hypothesis since there is a tendency to add a complementiser in less complex sentences. Translators find themselves in a more cognitive challenging situation than regular authors because they have to deal with two languages instead of one. Due to this increased cognitive effort, they will include a that even in less complex sentences.

Another reason for the higher that-rate in translated English opposed to original English is the risk-aversion hypothesis. Becher (2010, 2011) and Pym (2005, 2015) already indicated that translators tend to make the complement clause construction explicit in order to reduce the risk of being misunderstood by the readers. Also in this research, it can be concluded that translators have an inclination to include that where in original English zero is preferred. For instance, translators tend to be less sensitive to the register of the text because a that is added in most sentences regardless of whether this sentence belongs to a more creative or informal register, where zero is the norm. Additionally, also the frequency of the main verb in complement clause constructions is less taken into account by translators since they opt for a that-complementiser in almost every case. Complementary to the observations above for the cognitive
complexity hypothesis, it can be argued that translators add a *that* in less complex sentences to make the complement clause structure more visual and clear to the readers. Also Kruger (2018) and Kruger & De Sutter (to appear) claim that the higher *that*-rate can be attributed to the fact that translators choose for the option which has the smallest chance to lead to misunderstandings and which is closest to the normal, written standard language. In short, the risk-aversion hypothesis can be confirmed on the basis of this research.

### 6 Conclusion

The aim of this research was to use insights and methods concerning *that*-realisation in variational linguistics and translation studies in order to examine whether the choice between *that* and *zero* differed for translated English and original English. Whereas Kruger (2018) focussed on English translations from Afrikaans compared to original English, this study investigated English translations from Dutch opposed to original English. It was scrutinised whether there was a difference in *that*-complementation between translated English and original English, which factors were decisive for including or omitting *that* and whether these factors were similar for both varieties (i.e. translated and non-translated English). This was examined by using the MuPDAR approach, which consisted of three parts. The first step was the fitting of a general linear mixed-effect model (R1 model) based on the non-translated dataset. In step 2, the R1 model was applied to the translated dataset to verify whether *that*-complementation functioned similarly in both varieties. If the R1 model could predict the choice between *that* and *zero* correctly, translated and non-translated English would appear to be very similar in terms of *that*-complementation and vice versa. In the final step, a new general linear mixed-effects model was fitted to discover which factors led to differences in *that*-realisation between translated English and original English and to determine whether in translated English, there was an inclination towards using a *that*-complementiser or a *zero*-complementiser. Afterwards, the results were used to verify three hypotheses formulated by Kruger (2018): the transfer hypothesis, the cognitive complexity hypotheses and the risk-aversion hypothesis.

This study confirmed that, in general, *that*-complementation is very similar for translated and non-translated English and that the choice between *that* and *zero* depends on the same factors for both varieties. However, it has to be noted that translated English has a slight inclination towards using a *that*-complementiser more frequently than non-translated English. These results are in line with Kruger (2018) and Kruger & De Sutter (to appear).
Both Kruger (2018) and Kruger & De Sutter (to appear) could not find evidence for the transfer hypothesis since a higher zero-rate in Afrikaans did not lead to a higher zero-rate in translated English. In fact, with a few exceptions, the complementiser *that* was used almost every time in English translations from Afrikaans. In this research, however, translations from Dutch into English were scrutinised and in Dutch a *that*-complementiser is obligatory to introduce complement clauses. Results indicated a higher use of *that* in translated English which is similar to Kruger (2018) and Kruger & De Sutter (to appear). Consequently, it can be possible that the higher *that*-rate was caused by the influence of the contact language Dutch. Nevertheless, the transfer hypothesis cannot be confirmed since the higher use of *that* in translations can also be the result of the risk-aversion hypothesis. If this would be the case, it can be concluded that translators favour the complementiser *that* regardless of the source language.

The results seemed to indicate the existence of a cognitive complexity hypothesis because more complex factor levels appeared to prefer the inclusion of *that* while less complex factor levels revealed a bias towards using zero in approximately 10% to 20% of the cases. On this view, the findings of the multifactorial studies on *that*-realisation in original English were supported (Tagliamonte & Smith, 2004; Torres Cacoullos & Walker, 2009; Jaeger, 2010; Thompson & Mulac, 1991). In addition, due to the extended cognitive effort on behalf of the translators, the cognitive complexity hypothesis can provide an explanation for the higher *that*-rate in translated English.

The fact that translators tend to be less sensitive to the complexity of the sentence can be linked to the risk-aversion hypothesis. Translators have a tendency to include *that* in order to make the complement clause construction explicit and to avoid the risk of being misunderstood. In addition, results indicated that translators take less into account the register of the text or frequency effects of the main verb. They seem to add the complementiser in all situations, even when a zero-complementiser is preferred.

As it can be concluded that languages with an optional complementiser (Afrikaans) do not seem to have an influence on the *that/zero* alternation in their translations, further research is necessary to confirm whether languages with an obligatory complementiser, such as Dutch, influence the choice between *that* and zero in translations. It should be examined whether it is the contact language which affects an increased *that*-rate or whether this can rather be assigned to the avoidance of risk on behalf of the translators. In addition, it might be interesting to further examine whether L2 learners of English show similar attitudes towards *that*-complementation as translators. Based on the findings of Wulf, Gries & Lester (to appear) and this study, similar tendencies should be detected. However, it is expected that the
English L2 learners will be more influenced by the contact language, especially when in the contact language the inclusion of the complementiser is obligatory, such as in Dutch.

7 References


8 Appendices

8.1 Appendix A: Full statistical report on the fitting of the R1 model

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]
Family: binomial  (logit)
Formula: complementiser ~ log(MCVerbToCCLength + 1) + LemmaConstrFreq100kLOG + aspect + subject + Text.Type + tensemodality + (1 | File) + (1 | Benaming)
Data: data.nontrans.eng
Control: glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))

AIC  BIC  logLik  deviance  df.resid
2761.0  2850.6  -1365.5  2731.0  2894

Scaled residuals:
  Min   1Q  Median   3Q  Max
-5.4908 -0.5124  0.2102  0.4784  3.9593

Random effects:
  Groups   Name   Variance  Std.Dev.
        File  (Intercept) 0.8186  0.9047
        Benaming (Intercept) 1.3683  1.1697
Number of obs: 2909, groups:  File, 488; Benaming, 93

Fixed effects:  Estimate Std. Error  z value  Pr(>|z|)
       (Intercept) 6.2592     0.6621   9.453    < 2e-16 ***
log(MCVerbToCCLength + 1) 1.1842     0.1633   7.254    4.05e-13 ***
LemmaConstrFreq100kLOG -0.3226     0.1278  -2.525    0.011585 *
aspectSIMPLE -0.5221     0.2346  -2.226    0.026029 *
subjectno overt subject -1.6942     0.6702  -2.528    0.011474 *
subjectnoun -1.8945     0.5130  -3.693    0.000221 ***
subjectpronoun -2.3574     0.5106  -4.617    3.90e-06 ***
Text.TypeExternal Communication -0.3441     0.2969  -1.159    0.246447
Text.TypeFictional literature -1.6136     0.4951  -3.259    0.001116 **
Text.TypeJournalistic texts -1.6406     0.2194  -7.478    7.55e-14 ***
tensemodalitynon-finite -0.4739     0.4778  -0.992    0.321269

tensemodalitypast -1.0089     0.2553  -3.952    7.75e-05 ***
tensemodalitypresent -0.7547     0.2373  -3.180    0.001472 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
8.2 Appendix B: Full statistical report on the fitting of the R2 model

AIC   BIC  logLik deviance df.resid
2094.0 2165.6  -1035.0  2070.0  2853

Scaled residuals:
     Min   1Q Median  3Q    Max
-1.6110 -0.4342 -0.1923 -0.0642  12.3223

Random effects:
Groups   Name   Variance  Std.Dev.
File  (Intercept) 0.3111    0.5577
Benaming (Intercept) 1.7772    1.3331

Number of obs: 2865, groups: File, 280; Benaming, 87

Fixed effects:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)     -6.6624     0.6501   -10.248  < 2e-16 ***
Text.TypeExternal Communication -0.4656     0.2844    -1.637    0.10160
Text.TypeFictional literature    1.2028     0.3753     3.205    0.00135 **
Text.TypeJournalistic texts      1.3525     0.2083     6.493    3.89e-11 ***
log(MCVerbToCCLength + 1)       -0.2719     0.1603    -1.696    0.08988
LemmaConstrFreq100kLOG          0.3928     0.1618     2.428    0.01517 *
subjectno overt subject         1.5470     0.5318     2.909    0.00363 **
subjectnoun                    1.0280     0.4925     2.087    0.03687 *
subjectpronoun                 2.4571     0.4824     5.093    3.52e-07 ***
aspectsSIMPLE                  0.7965     0.3059     2.603    0.00923 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1