Routledge
Taylor & Francis Group

Perspectives
Studies in Translation Theory and Practice

ISSN: 0907-676X (Print) 1747-6623 (Online) Journal homepage: https://www.tandfonline.com/loi/rmps20

On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach

Gert De Sutter & Marie-Aude Lefer

To cite this article: Gert De Sutter & Marie-Aude Lefer (2019): On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach, Perspectives, DOI: 10.1080/0907676X.2019.1611891

To link to this article: https://doi.org/10.1080/0907676X.2019.1611891

Published online: 16 May 2019.

PERSPECTIVES
https://doi.org/10.1080/0907676X.2019.1611891

Routledge
Taylor & Francis Group

On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach

Gert De Sutter [a] and Marie-Aude Lefer[b]

[a]RU Empirical and Quantitative Translation and Interpreting Studies (EQTIS), Faculty of Arts and Philosophy, Ghent University, Ghent, Belgium, Centre for English Corpus Linguistics (CECL); [b]Faculté de Philosophie, Arts et Lettres, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ABSTRACT

Twenty-five years after the emergence of corpus-based translation studies the present paper offers a critical analysis of the current state of the art in corpus-based translation studies, focusing on what it has yielded in terms of description, methodology and theory. This analysis leads to the detection of problem areas which result in limitations to progress in the field. We argue that these limitations can be overcome, by adopting a revised research agenda for empirical translation studies, with a broader methodological scope and more theoretical awareness. At the very heart of this agenda is the description of translation as an inherently multidimensional linguistic activity and product, which is simultaneously constrained by sociocultural, technological and cognitive factors, leading ultimately to a better understanding of what translation exactly is, how it is shaped by varying circumstances, and how it relates to other types of constrained communication. The added value of this research agenda is illustrated in two case studies on optional *that* in English complement clause constructions.

ARTICLE HISTORY
Received 4 June 2018
Accepted 16 April 2019

KEYWORDS
Empirical translation studies; optional *that*; translated language; learner language; multifactorial statistics

1. Introduction

Anyone who is new to corpus-based translation studies (CBTS) and who wants to read his/her way into the field is very likely to first come across Mona Baker's 1990s and early 2000s papers (1993, 1995, 1999, 2004). Baker's work has influenced the field enormously, in that she founded it and proposed a specific research agenda (the so-called 'translation universals agenda') showing how corpus research in translation studies could be done, thereby paving the way for what would soon become a thriving field. Illustratively, in their diachronic meta-analysis of the dynamics of several sub-disciplines in translation studies, Zanettin, Saldanha, and Harding (2015, p. 20) note that '[i]n terms of methodologies, the impact of linguistic corpora is noticeable and is a trend that is clearly here to stay'.

Nonetheless, it seems fair to say that more than twenty years after the publication of Baker's seminal papers, the increase in understanding of translation as a product and

CONTACT Gert De Sutter gert.desutter@ugent.be Department of Translation, Interpreting and Communication, Ghent University, Groot-Brittanniëlaan 45, BE-9000, Gent, Belgium

process is relatively modest, and definitely disproportionate given the massive interest, time, money and energy invested in CBTS. Fundamental questions remain largely unanswered, such as which social, pragmatic and cognitive mechanisms shape translation, how these mechanisms interact, and to what extent this interaction functions differently than in other types of monolingual and bilingual written language production. For instance, the true nature of explicitation, a heavily researched topic in CBTS, has yet to be understood, as it is still impossible to predict when translators will (or will not) rely on linguistic devices indexing explicitness, such as cohesive markers.

As will be argued in this paper, the observed disproportion is most likely caused by over-attention to a non-crucial part of Baker's research program, namely the translation universals agenda, which was only meant as 'a very tentative list of suggestions which can provide a starting-point for corpus-based investigations in the discipline **but which do not, by any means, address the full potential of corpora in translation studies**' (Baker, 1993, p. 243, our emphasis). As a consequence, what was truly essential in Baker (1993) – the exploration of corpus-linguistic methods of scrutinizing translational products in order to find the 'principles that govern translational behaviour and the constraints under which it operates' (p. 235) – has not yet been explored to the fullest extent. Somewhat polemically, one could say that Baker's research agenda was misread from the very beginning, with translation scholars obstinately carrying on producing research findings in a research framework that was not meant as a framework in the first place, leading to much (conceptual) confusion and a general standstill in understanding the central object of research (although it must be acknowledged that recent years have witnessed a marked increase in more theory-relevant empirical translation research). In the present paper, we will argue that there are at least three other factors that hinder significant conceptual and theoretical progress, namely the strong focus of corpus-based translation scholars on finding linguistic differences rather than similarities, the lack of (advanced) statistical testing and the restricted collaboration with scholars from other fields.

A more accurate understanding of the governing principles underlying translation, and the constraints under which it operates, can be achieved, in our view, by re-adopting and updating the essential aspect of Baker's research program, i.e., looking over the disciplinary fence and carefully selecting corpus-linguistic, ethnographic, sociological, and psycholinguistic methods that are appropriate for studying central aspects of translation, as well as interpreting research outcomes in an emerging, bottom-up translation theory that builds on theories in neighboring disciplines, such as contact linguistics, second language acquisition research and psycholinguistics.

Our paper is structured as follows. Section 2 contains a re-appraisal of Baker's groundbreaking ideas and an analysis of the developments in CBTS, which leads to the detection of several problem areas. Section 2 is also devoted to the re-adoption of Baker's fundamental idea, on the basis of which we propose a new, updated specific research agenda for what we will refer to as 'empirical translation studies' (see also Ji, 2016; Oakes & Ji, 2012). This agenda calls, among other things, for more interdisciplinary awareness and more advanced statistical testing. Section 3 illustrates how this new research agenda can be put into practice by revisiting the issue of explicit and implicit *that* in English declarative complement clauses (as in examples 1 and 2), and shows the importance of applying (and possibly fine-tuning) this new research agenda as soon as possible.

(1) Business tells us that recruitment needs to be made easier and more flexible at all skill levels (dpc-erp-000443-en) [explicit *that*]
(2) I remember an enthusiastic Togolese man who came up to me after a lecture and said he wanted to conduct a prospective study in Togo (dpc-vla-001162-en) [zero complementizer]

More particularly, we show that multifactorial statistical methods are able to shed a completely different light on the alternation between explicit *that* and zero complementizer, and that patterns in translated language can (and should) be compared with other bilingualism-influenced, constrained language varieties (in this case, L2 writing). Finally, section 4 concludes this paper with a summary of the main findings of the *that* case studies and an outlook on future research initiatives.

2. The origins of CBTS: re-reading and re-applying Baker's research agenda

The start of CBTS is traditionally situated in the early 1990s with the publication of Mona Baker's seminal paper 'Corpus Linguistics and Translation Studies. Implications and Applications'. The main goal of her paper was to showcase the possibilities offered by computerized corpora and corpus-linguistic analytical techniques for the development of the descriptive and theoretical branches of translation studies.

In the first part of the paper, Baker put forward arguments as to why translation studies – at that point in time – seemed to have evolved to such an extent that it was ready to adopt the corpus-linguistic approach. In particular, she referred to a number of internal developments in the field of translation studies, such as the shift in orientation from the source text to target system (under the influence of polysystem theory) and the accompanying shift from the normative concept of 'equivalence' to the descriptive concept of 'norms' (under the influence of Gideon Toury's model of norms). As a result, translation scholars moved from the evaluative-comparative analysis of individual instances in specific translations and their corresponding source texts to the descriptive analysis of general patterns in translated texts as separate, autonomous entities in the target language. The shift from the individual to the general, from the derived to the fully fledged, and from the evaluative to the descriptive was thus seen as a turning point in translation studies, in that the field was ready to use corpora and corpus-linguistic techniques. Additionally, Baker argued, the insistence on usage-based generalizations and data-driven explanations about translational behavior in the area of descriptive translation studies, with Toury as its most important representative, helped to ensure the successful introduction of the corpus methodology into translation studies.

The second part of Baker's article was devoted to a discussion of example research questions for CBTS. There can be no doubt that this part of her paper has attracted an enormous amount of attention in the field. We would even go one step further by claiming that what has appealed to so many scholars is just one illustrative section in that second part – the *translation universals agenda* – and, in particular, one sentence within that section, namely the tentative definition of translation universals as 'features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems' (Baker, 1993, p. 243). Despite many nuances and warnings in the paper itself, this much-quoted definition has given rise to

numerous publications in CBTS, which have focused on lexical, morphosyntactic, semantic, and discursive differences between translated and original non-translated language, most of them related to the posited translation universals put forward by Baker. Crucially, the influence exerted by a range of language-internal and -external factors, such as linguistic complexity, register, language prestige and translation expertise has often been neglected (see, however, the few exceptions listed below). The role played by source language influence has not been taken into account systematically either, and this is in all likelihood inherent to Baker's definition of translation universals, which explicitly excludes interference (see, however, Laviosa, 1998).

It is worth adding here, for the sake of completeness, that a number of early studies have addressed other (i.e., non-universals-related) topics. As rightly pointed out by Laviosa (2002, p. 78), 'despite great interest in studying the specificity of translational language, empirical corpus-based studies of translation are not limited to the investigation of universals'. We might mention, for example, Laviosa (2000) on ideology and Olohan's case studies (2004, pp. 145–167) on translator style (see also Malamatidou, 2018).

However appealing the universals paradigm may have been at first sight, it is not unfair to state in retrospect that it has given rise to many of the persistent problems we are facing now, and which could have been avoided (at least in part), had corpus-based translation scholars treated Baker's seminal paper more cautiously. In our opinion, there are at least four major problem areas.

The first problem area is the preoccupation of early CBTS with finding linguistic *differences*, rather than *similarities*, between translated and non-translated texts (cf. Mauranen, 2000), and explaining these in terms of translation universals. This can of course be understood in the context of an emerging discipline that needs a *raison d'être*, but with the benefit of hindsight it is a questionable approach to assume first and foremost differences when translated texts in a given language that are produced by highly skilled, native-language professionals are compared with texts in the same language produced by presumably equally skilled language professionals (journalists, writers, spokesmen etc.), with the only obvious difference being the circumstances under which the texts are produced (bilingual vs. monolingual language activation). There is no reason to assume that professional translators would consciously or unconsciously produce texts that are *significantly* different from texts they would have produced in a monolingual setting. The only plausible reasons for assuming differences are related to the input the translator is faced with – the source text – or the specific bilingualism-influenced communicative setting of translation – expressing someone else's message in another language and culture – but these factors were often left out of the equation in early CBTS. In other words, subtle quantitative differences *are* likely to be found across translated and non-translated texts, alongside a massive number of commonalities.

The second problem area is theoretical in nature: the preoccupation with finding linguistic differences in translated texts as against non-translated texts has left the explanatory framework proposed by Baker, or any other theoretical framework, under-developed. Instead of empirical research in translation studies giving rise to the falsification, verification or adaptation of the hypotheses of universal features of translation, as initially intended by Baker, universal features have been used repeatedly and uncritically to 'explain' specific patterns observed in the corpus data. In the process, translation universals have gradually lost power in that they have only been used as *fixed, passe-partout* and

post hoc explanations: whatever linguistic phenomenon is being studied, there will always be some translation universal available which can be used to rationalize the descriptive patterns uncovered in the data. It is particularly telling that in these early studies no a priori hypotheses about translation universals are given, signaling that the main goal of these studies is descriptive in nature, not theoretical. As a consequence, a viable, encompassing explanatory framework for CBTS is still lacking, leaving the field with a plethora of descriptive results which it cannot make sense of and a set of fixed, dogmatic translation universals which do not help give a more accurate and reliable insight into what translation really is.

A third problem area is methodological in nature. Most CBTS research to date has stuck to monofactorial research designs, in which the distribution of a linguistic phenomenon is investigated with reference to only one explanatory factor – translation status/ontology (translated vs. non-translated) –, thereby ignoring other potential explanatory factors. From studies in variational linguistics we now know that linguistic choices as evidenced in corpora are the result of the interplay of a wide range of language-internal and -external factors, such as syntactic complexity, animacy, text type and gender, to name a few (cf. Bresnan, Cueni, Nikitina, & Baayen, 2007; Divjak & Gries, 2006; Gries, 2018; Pijpops & Speelman, 2017; Szmrecsanyi, Grafmiller, Heller, & Röthlisberger, 2016; Tagliamonte & Baayen, 2012). As Gries (2018) boldly puts it, 'monofactorial observational studies have virtually nothing to contribute to corpus linguistics' because

> (i) no phenomenon is monofactorial and (ii) even if one had a new *mono*factorial hypothesis of a phenomenon, it would still require *multi*factorial testing to determine either (a) whether it either adds anything to what we already know about the phenomenon (by statistically controlling for what we already know) or (b) whether it replaces (parts of) what we already know about the phenomenon. (p. 295)

Coming back to the corpus-linguistic field under scrutiny here, CBTS, it clearly remains to be seen whether *translation status* remains an important factor in understanding the variability of a given linguistic phenomenon when integrated into a multifactorial research design.

The final problem we would like to mention here is what we call the *auto-isolation* of corpus-based translation studies, which might be the basic problem underlying the other problem areas outlined above. The field can hardly be considered as an interdiscipline nowadays, with most studies narrowly charting low-level linguistic differences between translated and non-translated texts without taking into account theoretical and methodological developments in other, related, fields such as corpus linguistics (including learner corpus research), variational linguistics, contrastive linguistics, sociolinguistics, psycholinguistics and cognitive linguistics, to name but a few.

Not only have all of the above-mentioned problems, which are clearly interrelated, slowed down scientific progress in the field, standing in the way of a more accurate understanding of translation products, they have also brought about – to some extent – unreliable findings, to the point that some of the modest progress made in the field in recent years could very well be invalidated. In the present paper, we will exemplify this claim in the next section, where we show empirically that the main results obtained in the classical *that* study by Olohan and Baker (2000) cannot be sustained when a multifactorial approach is adopted.

What is remarkable is that none of the problem areas mentioned above are the consequence of an improper, narrowly focused, premature research agenda, but are mostly related to an insufficient reading of Baker's paper and an overly uncritical adoption of a 'very tentative list of suggestions' (Baker, 1993, p. 243) as a research agenda. Somewhat ironically, Baker herself has also indirectly promoted this reductionist approach in some of her empirical studies (cf. Olohan & Baker, 2000), although she has regularly warned against the 'danger of uncritical application of the methodology' (Baker, 2004, p. 169).

If we want to solve the problems mentioned above, we are in need of a new, updated research agenda for CBTS. From the discussion above, it can be deduced that modern-day empirical work in the field of translation studies should be:

- Multifactorial: as in any other communicative context, linguistic choices in a translation setting are governed by a multitude of factors, ranging across the education, experience and expertise of the translator, time constraints, the translation brief, language attitudes, the translation policy of a given target culture, the target readership, the communicative function of the target text, the type of (self-)revision and editorial intervention, the use of computer-aided translation tools, the genre and domain, the linguistic features of the source text, the source-language prestige, the translation directionality, etc. Understanding translation implies understanding its multidimensional structure, and hence multifactorial research designs are essential.
- Interdisciplinary: translation is a communicative event in which a language user mediates a message between two languages and cultures, and from that perspective translation is not as unique as it might appear at first sight. Other types of communication in language contact settings share many of the properties of the translational act, such as non-native indigenized varieties of English (Kirkpatrick, 2010), English as a Lingua Franca (Seidelhofer, 2013) or English as a Foreign Language (Granger, Gilquin, & Meunier, 2015). Instead of looking for descriptive patterns and finding accurate explanations for each of these disciplines separately, it might make more sense to build a shared theory of what Kruger and Van Rooy (2016) and Kruger (2018), following Lanstyák and Heltai (2012), call 'constrained communication', a theory which could inform new empirical investigations.
- Multi-methodological: the implementation of new corpus-linguistic methods in translation studies is certainly the first step to take, but as Baker (2004, p. 184) rightly notes, this should not 'be seen as a free-standing methodology that does not need to be complemented by other methods of research. Like any other methodology, it can only take us so far, and no further'. The cross-fertilization of different methodological approaches in the context of one research project has already been adopted in other fields of usage-based linguistics, where methodological pluralism is quite widespread (e.g., combining corpus and experimental data) (cf. Ellis & Simpson-Vlach, 2009; Gilquin & Gries, 2009; Schönefeld, 2011). Since the combination of methods is the right way to go in order to increase our understanding of translation, the term 'corpus-based translation studies' should be replaced by the more accurate and encompassing term 'empirical translation studies'.

It is certainly true to say that some scholars have already started implementing this research agenda in recent years. Multifactorial designs, for example, are

increasingly common and show that the linguistic make-up of translational products is simultaneously shaped by factors such as editorial intervention, expertise, register, source language, translation direction, translation mode (interpreting vs. written translation) and translation method (human vs. computer-aided), among others (see e.g., Bernardini, Ferraresi, & Miličević, 2016; Bisiada, 2014; Delaere & De Sutter, 2017; Evert & Neumann, 2017; Kruger, 2017, in press; Lapshinova-Koltunski, 2017). Methodological pluralism is also starting to emerge in the field: corpus data are increasingly combined with elicited data, key logging and eye tracking (cf. Halverson, 2017; Heilmann, Serbina, & Neumann, 2018; Kajzer-Wietrzny, Whyatt, & Stachowiak, 2016). There have been a few interdisciplinary studies as well, such as Gaspari and Bernardini (2010) and Kruger and Van Rooy (2016, 2018) on translated and non-native English.

In the remainder of this paper we will claim that the adoption of this new research agenda is not only important but, in our opinion, *indispensable* if we want findings in empirical translation studies to be accurate, reliable and generalizable so that we can start building solid, stable theories. More particularly, we will focus on two aspects of this agenda – multifactoriality and interdisciplinarity – by revisiting the classic topic of optional *that* complementizer in English complement clauses.

3. Two case studies exemplifying the new research agenda

In order to illustrate our claim that a multifactorial research design provides added value to our understanding of what translation is, we revisit a widely investigated topic in CBTS, namely the variation between explicit and implicit *that* in English complement clauses introduced by *say* and *tell*, weighing the explanatory value of the factor *translation status* (i.e., the difference between original and translated language) against other potentially relevant genre- and complexity-related factors. In addition, we widen the scope of our case study by investigating the same linguistic phenomenon in L2 student writing. Our objective is to identify differences and similarities between texts produced by translators on the one hand and learners of English on the other, thereby showing how interdisciplinary comparisons can increase our understanding of translation. The two groups clearly differ in terms of writing expertise (learners of English being novice writers), but they operate within similar communicative settings, namely settings in which two languages are co-activated (though arguably in different ways and to different degrees). In other words, translators and L2 novice writers are both part of bilingual contexts 'where heightened constraints operate on them' (Kruger & Van Rooy, 2016, p. 26). An important difference in this respect, however, is directionality: translators mostly produce texts in their L1 (with the source language, their L2, potentially influencing their L1 production), whereas learners of English produce texts in their L2 (with possible manifestations of L1 transfer). Before presenting and discussing the results of our two case studies in Sections 3.2 and 3.3, we first give a brief state of the art on optional *that* in Section 3.1.

3.1. Brief state of the art: optional that in CBTS and learner corpus research

As it is beyond the scope of the present paper to offer a detailed review of the literature on optional *that* complementation in translation studies and linguistics (see Kruger, in press;

Wulff, Gries, & Lester, 2018, and Kruger & De Sutter, 2018 for more extensive overviews), the present section mainly focuses on previous work that is directly relevant to our own case studies.

In a seminal paper published in 2000, Olohan & Baker investigate the use of optional *that* with the reporting verbs *say* and *tell* as an indicator indexing explicitation. The corpus data extracted from the *Translational English Corpus* (TEC) and the *British National Corpus* (BNC) reveal that the complementizer *that* is proportionally more frequent in the TEC than in the BNC with both verbs, which shows 'a tendency towards syntactic explicitation in translated English' (Olohan & Baker, 2000, p. 157). Several methodological and theoretical criticisms have been leveled at Olohan & Baker's study, most vividly by Becher (2010), who rightfully stresses (among other things) that genre variation and source language influence have been completely left out of the analysis. Following Olohan & Baker's paper, *that* complementation has been widely investigated, with corpus-based studies confirming the higher incidence of the complementizer (as opposed to its omission) in translated language (see e.g., Kruger, in press; Kruger & Van Rooy, 2012; Redelinghuys & Kruger, 2015).

The topic of *that* complementation has also been addressed – though less frequently than in translation studies – in learner corpus research, most notably by Wulff et al. (2018). With a view to uncovering the factors that govern the presence of the complementizer *that* in English complement constructions in written (argumentative essays) and spoken (informal interviews) language use by German and Spanish learners of English, the authors have analyzed a wide range of variables, such as mother tongue background, mode (writing vs. speech), surprisal, and the length of various matrix-clause and complement-clause elements (see section 3.2.2 below for an extensive overview of these complexity-related variables). The MuPDAR statistical approach they adopt reveals that, notwithstanding minor but significant differences between L2 learners and native speakers (NS) of English (with learners displaying more conservative behavior than NS), 'intermediate-advanced German and Spanish learners are quite well aligned with NS norms overall' (Wulff et al., 117).

In two recent papers, Kruger (in press) and Kruger and De Sutter (2018) have adopted Wulff et al.'s (2018) multifactorial approach in a corpus study on optional *that* in translated and non-translated South African English compared with British English. More particularly, they aim at disentangling three proposed explanations for explicitation: cognitive complexity, conventionality, and source language influence. The first explanation is operationalized by means of several length-related factors, such as the length between the matrix-clause verb and the complement clause onset, assuming that increasing length correlates with increasing cognitive complexity (and hence a higher likelihood of explicit *that*; cf. also Rohdenburg's (1996) complexity principle). The second explanation is mainly operationalized by means of register-related factors: some registers prefer explicit *that*, whereas others favor the zero complementizer. Finally, source-language influence is operationalized indirectly in both papers: given that Afrikaans, which is the source language of the translated data under investigation, has a clear preference for *dat* ('that') omission, underuse of *that* compared to original, non-translated data would point at source-language influence, whereas overuse would not. Results indicate that source language does not influence the use or omission of *that*, whereas complexity and conventionality do.

3.2 Case study 1: optional that in translated and non-translated English

3.2.1. Hypotheses

Following Olohan and Baker (2000), we hypothesize that (i) translated English shows a significantly higher inclination to use explicit *that* compared to original English, and (ii) in a multifactorial analysis *translation status* is one of the driving factors behind the *that* alternation.

3.2.2. Data

We rely on the *Dutch Parallel Corpus* (DPC), which is a bidirectional parallel corpus containing Dutch source texts translated into French and English as well as French and English source texts translated into Dutch (Macken, De Clercq, & Paulussen, 2011). Given the objectives of this case study, we only make use of the English component of the corpus, which amounts to approximately 5 million tokens in total (2.5M for original, non-translated English and 2.5M for English translated from Dutch). The corpus is stylistically and regionally stratified: it contains journalistic, touristic, legal and (non-)fictional texts, specialized and broad company communication, and political discourse taken from US and UK text providers (cf. Delaere & De Sutter, 2017 for a description of these register categories). The corpus was queried for all instances of *say* and *tell* taking a declarative complement clause that either contains explicit *that* or where *that* could be inserted. After querying the corpus and weeding out irrelevant instances, it turned out that the phenomenon under investigation was rather infrequent in some of the register and regional categories, and in order to guarantee the stability of the statistical model to be presented, we decided to leave out all categories with a frequency lower than 50. As a consequence, the dataset only contains UK texts from the following three register categories: journalistic texts, (non-)fictional literature and political discourse. The resulting dataset contains 813 occurrences, and was subsequently coded for the response variable (*implicitTHAT*, *explicitTHAT*) and the following explanatory variables (factors): translation status (*translated*, *non-translated*), verb lemma (*say*, *tell*), verb token (*say, says, said, saying, tell, tells, told, telling*) and a number of complexity-related explanatory variables, taken from Wulff et al. (2018). These are listed below, and illustrated by means of the following sentence, taken from the translational part of DPC: *As early as two years ago I said that De Post-La Poste needed an external partner with know-how in areas we have to catch up*.

- LengthComplement: the length of the full complement clause, excluding *that*, counted in number of characters (excluding spaces) (e.g., *De Post-La Poste needed an external partner with know-how in areas we have to catch up*; length = 69).
- LengthMCVerbCCSubject: the distance between the matrix-clause (MC) verb and the complement clause (CC) subject, excluding *that*, counted in number of characters (excluding spaces) (in our example sentence, length = 0).
- LengthMCSubject: the length of the matrix-clause subject, counted in number of characters (excluding spaces) (*I*; length = 1).
- LengthComplementSubject: the length of the complement-clause subject, counted in number of characters (excluding spaces) (*De Post-La Poste*; length = 13).

- LengthCIM: the length of clause-initial material (CIM) (all material preceding the onset of the main clause), counted in number of characters (excluding spaces) (*As early as two years ago*; length = 20).
- LengthCCremainder: the length of the complement-clause remaining material after the complement-clause verb, counted in number of characters (excluding spaces) (e.g. = *an external partner with know-how in areas we have to catch up*; length = 50).
- LengthMCSubjectMCVerb: the distance between the matrix-clause subject and the matrix-clause verb, counted in number of characters (excluding spaces) (in our example, length = 0).

Preliminary tests on the complexity-related variables, which are all numerical in nature, showed that the distribution was highly skewed, and for that reason we decided to logarithmically transform these variables, as is usually done in variational-linguistic research designs.

The simultaneous effect of all these explanatory variables on the presence vs. absence of *that* is measured by means of a generalized linear mixed-effects model (glmm), using RStudio 1.1.383 (R Core Team, 2016). This model reveals whether or not each of the above-mentioned explanatory variables has a significant effect on the response variable, what the relative effect of each of the variables is and what the overall performance of the model is (i.e., the joint effect of all significant explanatory variables) in terms of descriptive and predictive adequacy (i.e., to what extent the model is able to capture the variation in the current dataset, and to what extent it is able to predict unseen data).

3.2.3. Results and discussion

Before presenting the results of the glmm, we first inspect the general distribution of the response variable in the dataset (Figure 1) and the association between translation status and the response variable (Figure 2).

As can be seen in Figure 1, implicit and explicit *that* are equally distributed in our dataset, with $n = 406$ for explicit *that* and $n = 407$ for implicit *that*. Figure 2 shows that the distribution of implicit and explicit *that* across the two translation status conditions is not equal: translated English appears to have a preference for explicit *that* (58%, $n = 123$) whereas original, non-translated English (slightly) favors implicit *that* (53%, $n = 317$). The difference between translated and non-translated English is statistically significant ($X^2(1) = 6.62$, $p = 0.01$). Although this result is in line with Olohan and Baker's (2000) findings, and thereby confirms previous observations that translated English prefers more explicit structures, we still need to check whether the effect of translation status remains stable vis-à-vis the other explanatory variables. If, for instance, the distance between the matrix-clause verb and the complement-clause subject has a strong effect on the use of implicit/explicit *that* and if it appears that in our dataset translated English has a higher proportion of sentences with a long verb-to-subject distance, then the effect of translation status might turn out to be a quasi-effect of an underlying, more powerful variable.

There are three possible scenarios. First, translation status has a significant main effect on the *that* alternation, potentially alongside other explanatory variables. This scenario would confirm the importance of *translation status* as a key factor in the understanding of *that* alternation. Second, translation status does not have a significant main effect, but is included in one or more interaction effects. This would mean that translation

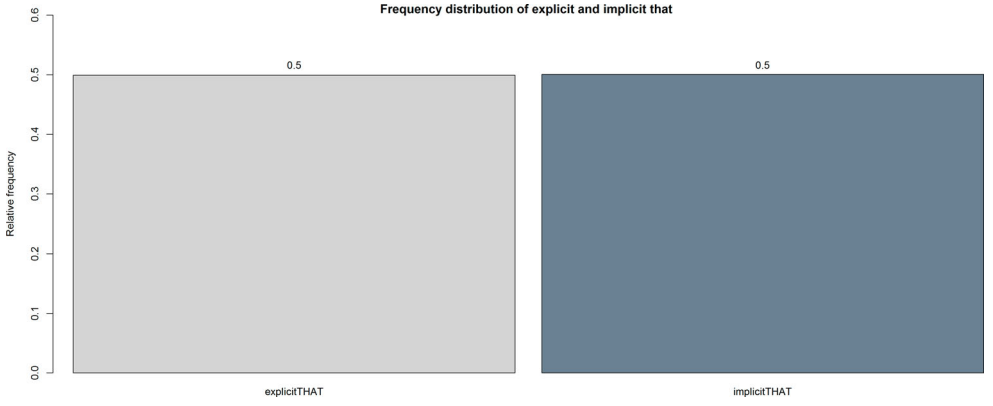

Figure 1. General distribution of implicit and explicit *that* in English complement clauses introduced by *say* and *tell* (data: Dutch Parallel Corpus).

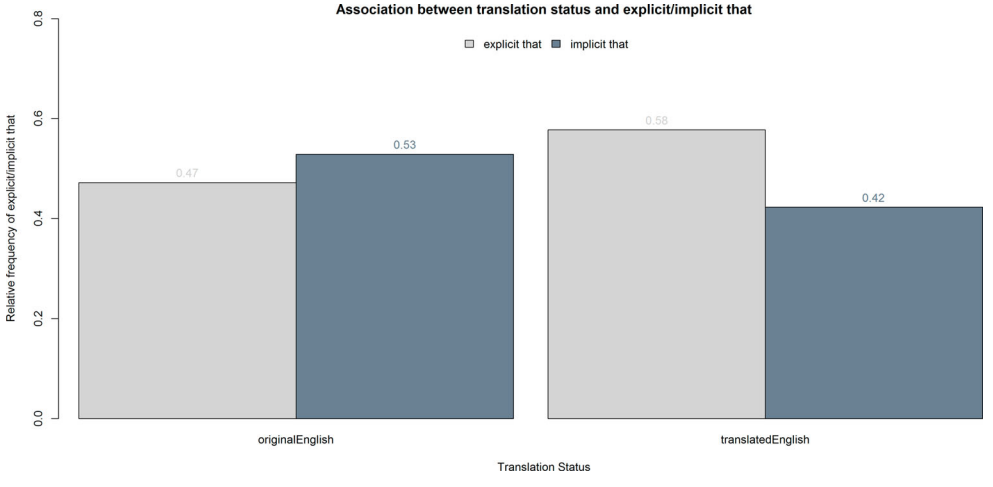

Figure 2. Association of translation status and implicit and explicit *that* in English complement clauses introduced by *say* and *tell* (data: Dutch Parallel Corpus).

status itself does not *directly* influence the alternation, but changes the effect that other explanatory variables have (either decreasing or increasing this effect). In this scenario, *translation status* only has an effect in specific contexts. Third, translation status does not have any effect at all: the alternation is governed by other variables only.

To find out which scenario is the most realistic, we fitted a glmm with *that* alternation as response variable and all the above-mentioned explanatory variables as fixed effects; we also included the text-id and verb lemma as a random effect, so as to accommodate text-specific and lemma-specific variation. We proceeded incrementally, starting out from a model with only one explanatory variable, and then adding other explanatory variables one by one. We used the Aikake Information Criterion to assess whether a newly added variable contributed considerably to the model, and if this was not the case, it was excluded again. In case our main explanatory variable, translation status, did not turn out to be significant as main effect, we tried to include it as a two-way interaction

effect with the other explanatory variables. In order not to over-complicate the model, we did not attempt to include other or more-way interaction effects. The resulting model is presented in the appendix: LengthMCVerbCCSubject, LengthComplement, LengthMCSubject have significant main effects, and there are significant interaction effects between translation status and register on the one hand and translation status and LengthComplementSubject on the other. In Figure 3, we focus on the effects of the significant predictors. The model was checked for multicollinearity (vif < 7) and overdispersion, but none of these appeared to be problematic. The model reaches high performance scores, with an 87% accuracy score and a high prediction score (c = 0.92).

The plots in Figure 3 present the nature and size of the effects of all significant predictors. It is immediately apparent that *translation status* itself has no significant main effect, unlike *LengthMCVerbCCSubject*, *LengthComplement* and *LengthMCSubject*. The first two significant predictors affect the choice between explicit and implicit *that* as follows: if the distance between the matrix-clause verb and the complement-clause subject is small, the probability of implicit *that* is around 50%, but this probability decreases linearly as the distance grows, approaching 0% when the distance is at its largest. Likewise, we see that the length of the complement clause correlates negatively with the use of implicit *that*: when the complement clause is short, implicit *that* is the preferred option, but with increasing length of the complement clause, the probability of implicit *that* decreases. In the top-right plot, we can also see that the probability of implicit *that* increases as the length of the matrix-clause subject increases. This finding is quite intriguing, as it goes against the general complexity trends outlined so far. A closer examination of the dataset seems to indicate that the omission of *that* occurs in a few cases where the complement-clause subject is a personal pronoun whose antecedent is the matrix-clause subject (see examples 3–5; we have marked the personal pronoun in bold and the matrix-clause subject in italics).

(3) When, in November, Unesco demanded the book be removed, *the head of the library* said **he** didn't understand why its inclusion was considered anti-semitic or offensive. (dpc-sta-002480)

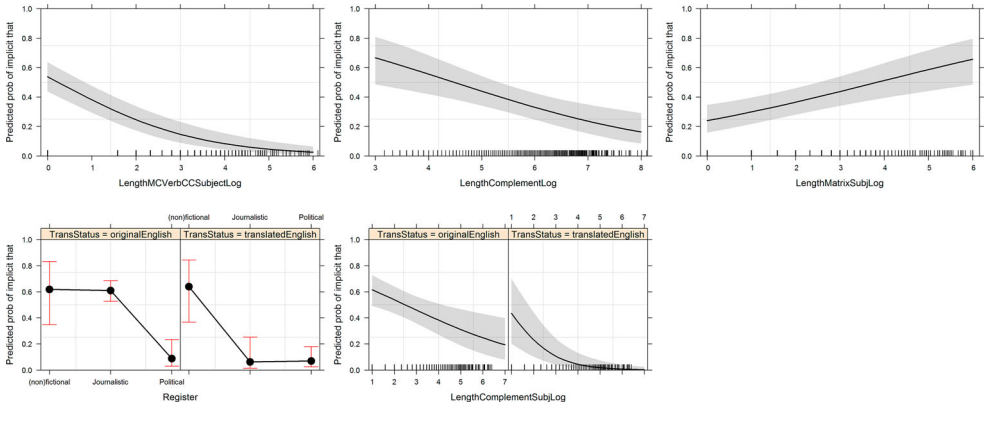

Figure 3. Effect plots of a generalized linear mixed effects model with explicit/implicit *that* as response variable, translation status, register, lengthCIM, lengthComplement, lengthMatrixSubject, lengthComplementSubject, lengthMCSubjectMCVerb, lengthMCVerbCCSubject, lengthCCRemainder as fixed effects and verb lemma and text-id as random effects (data: Dutch Parallel Corpus, $n = 813$).

(4) *Mr Cooper's mother, Patricia, who suffers from a heart condition*, said **she** could see 'no reason' for the attack. (dpc-ind-001728)
(5) *A fair share of the respondents* said **they** had received subsidies from the Flemish Agricultural Investment Fund, mainly via the Rural Development Programme for Flanders. (dpc-vla-001922)

We might hypothesize (very tentatively) that in such cases, *that*-omission leaves the cohesive chain uninterrupted. However, it must be acknowledged that at this stage we have not been able to detect other recurring patterns in our dataset. Should this trend be confirmed in follow-up studies, it would definitely deserve closer attention.

The two bottom plots show interaction effects with translation status: in the first plot we see that in translated journalistic texts, complement clauses with explicit *that* are far more frequent than in non-translated, original journalistic texts. The other register categories behave similarly in both translation status conditions: a preference for implicit *that* in (non-)fictional texts and a strong preference for explicit *that* in political discourse. The second interaction plot shows a steeper fall for translated English than for original English with respect to the effect of the length of the complement-clause subject: in both varieties, longer subjects are associated with lower implicit *that* scores, but this effect is much stronger in translated English than in original English (subjects in translated English with length ≥ 4 have a probability of ca. 0 of being preceded by an implicit *that*, whereas in original English this probability is never reached, even when the subject is at its longest).

In sum, our analysis reveals that the choice between implicit and explicit *that* is mainly governed by complexity-related variables and by register, whereas translation status itself only has an effect in particular contexts: in journalistic texts and in sentences with a medium-long or long complement-clause subject, translators opt for the more explicit structure compared to writers of original English texts. This seems to show that translators turn more rapidly to the clearer, safer, explicit option in at least some more syntactically complex environments (a tendency that also emerged in Kruger & De Sutter, 2018). Why translated journalistic texts have a significant preference for explicit *that* is not quite clear yet, but it is not impossible that this is due to genre or domain: many non-translated texts included in the DPC were published in newspapers such as *The Guardian* and *The Independent*, whereas the translated texts appeared in business magazines. In any case, we can conclude that translators in general make similar linguistic choices to non-translators, and that they basically decide between explicit and implicit *that* on the basis of the complexity of the syntactic environment and on the basis of register. These findings tie in with Kruger (in press) and Kruger and De Sutter (2018). Baker & Olohan's claim that the explicit use of *that* is illustrative of some kind of translation-specific or translation-inherent subconscious process can thus be refuted, at least for the time being.

3.3. Case study 2: optional that in native and non-native (L2) student writing

The second case study explores the usefulness of an interdisciplinary approach to empirical translation studies. In particular, we want to find out to what extent the tendencies found in translated text also apply to a different type of bilingualism-influenced, constrained language use, namely L2 writing. In doing so, we wish to contribute to the

emerging paradigm of constrained communication put forward by Kruger (2018) and inspired by Lanstyák and Heltai (2012). The term 'constrained communication' refers to 'communication taking place under conditions where one or several of the potential limiting factors play a greater than average role' (Lanstyák and Heltai, 2012, p. 100), i.e., language use characterized by prominent or increased constraints, such as editing, translation or code-switching. In their paper, Lanstyák & Heltai focus on two constraint dimensions, namely language activation (whether monolingual or bilingual) and text production (whether the text is mediated – e.g., editing and translation – or not). Kruger (2018) has elaborated extensively on this proposal and puts forward three additional dimensions that all constrain communication: modality and register (spoken, written, multimodal), proficiency (native/proficient vs. non-native/learner) and task expertise (expert vs. non-expert). Importantly, as noted by Kruger (2018), constraint dimensions should be viewed as continua rather than binaries.

In this second case study, the optional *that* choices in L2 student writing are compared with choices in native student writing, using an identical extraction methodology, coding system and multifactorial procedure. A comparison of the findings of the two case studies will also enable us to find out how professional writing relates to novice writing, and translation to L2 writing.

3.3.1. Hypotheses

Based on Wulff et al. (2018) and the findings reported in Section 3.2 (case study 1), we hypothesize that nativeness (i.e., native vs. non-native) has only a marginal effect on *that* variation. More particularly, it is expected that nativeness will only appear as part of interaction effects, not as main effects; hence, in some contexts, Dutch learners of English will show higher rates of explicit *that* compared to native students. Additionally, we expect that student (i.e., novice) writing in general will show higher rates of explicit *that* compared to professional writing (both translators and non-translators), as explicit *that* is the safer option – it is less likely to cause grammaticality or comprehensibility problems (cf. Wulff et al., 2018, p. 118) – and, given their lower level of writing expertise, it can be hypothesized that students have not yet internalized the subtleties of the professionals' probabilistic grammar (Bresnan, 2007).

3.3.2. Data

The case study relies on comparable corpora of native and non-native student writing, made up of 274,000+ and 164,000+ tokens respectively. We make use of 198 argumentative essays from the Dutch component of the *International Corpus of Learner English* (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009) and 217 American university students' argumentative essays taken from the *Louvain Corpus of Native English Essays* (LOCNESS; Granger, 1996) and from a similar corpus made available to us by Randi Reppen. All essays were untimed, with the use of reference tools (such as dictionaries) allowed. L2 essays typically deal with topics such as 'the usefulness of a university degree' and 'adverse effects of feminism'. Essays by native students deal with similar topics (e.g., the legalization of marijuana, abortion and gun control). Table 1 below gives an overview of the constrained language varieties that will be compared in this case study (L1 vs. L2 student writing), as well as across the two case studies (professional writing vs. L1 student writing and professional translation vs. L2 student writing).

As can be seen in Table 1, professional translation and L2 student writing differ in three main respects: text production (translation being mediated, while L2 writing is not), proficiency (native vs. non-native) and task expertise (expert vs. non-expert). In fact, translation is doubly mediated, as the translations included in the DPC are all published translations, which have most probably undergone an editing phase. Editing (or revision) itself is also a form of mediation (in the sense that the revision is 'dependent' on another text, i.e., the text to be revised). Also, it should be noted again that even though both settings are bilingual, directionality differs: while the translators represented in the DPC translate into their L1, Dutch-speaking students write in their L2. These two additional differences between translation and L2 writing will need to be borne in mind when comparing the results of the two case studies.

Table 1. Constrained language varieties under scrutiny in relation to Kruger's (2018) constraint dimensions.

	Professional writing	Professional translation	L1 student writing	L2 student writing
1. Language activation	Monolingual	**Bilingual (L1 production)**	Monolingual	**Bilingual (L2 production)**
2. Modality & register	Written	Written	Written	Written
3. Text production	**Mediated**	**Doubly mediated**	Unmediated	Unmediated
4. Proficiency	Native	Native	Native	**Non-native**
5. Task expertise	Expert	Expert	**Non-expert**	**Non-expert**

In addition, it is important to stress that different registers are represented in the DPC and the ICLE/LOCNESS corpora used in the two case studies: while the DPC data included in case study 1 cover three registers, ICLE and LOCNESS are made up of a single genre, namely argumentative essays (a genre that is typical of student writing).

The student writing dataset analyzed in case study 2 contains 363 relevant instances (*say*: $n = 304$, *tell*: $n = 59$). It was coded for the response variable (*implicitTHAT*, *explicitTHAT*) and the explanatory variables nativeness (*native*, *non-native*), verb lemma (*say*, *tell*), verb token (*say, says, said, saying, tell, tells, told, telling*) and the same complexity-related explanatory variables as in case study 1. Again, it turned out in preliminary tests that the distribution of all the complexity-related variables was highly skewed, so that we logarithmically transformed these variables. These variables were then included in a generalized linear mixed-effects model (glmm), carried out in exactly the same incremental way as in case study 1.

3.3.3. Results and discussion

Figures 4 and 5 present, respectively, the general distribution of explicit and implicit *that* in native and non-native student writing, and the association between nativeness and explicit/implicit *that*.

As hypothesized, there is a high preference among students in general for explicitly signaling the complement clause boundary (74%, $n = 363$), and this preference is even stronger among Dutch learners of English (81%, $n = 117$) as compared to native students (68%, $n = 150$). This difference between native students and learners of English is statistically significant ($X^2(1) = 6.63$, $p = 0.01$). When compared with the other explanatory variables in a generalized linear mixed-effects model, it turns out that the main effect of nativeness

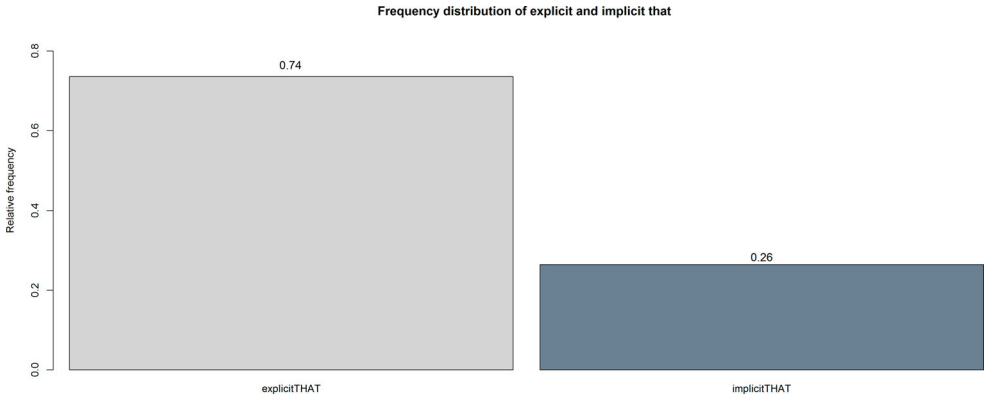

Figure 4. General distribution of implicit and explicit *that* in English complement clauses introduced by *say* and *tell* (data: LOCNESS and ICLE-DU).

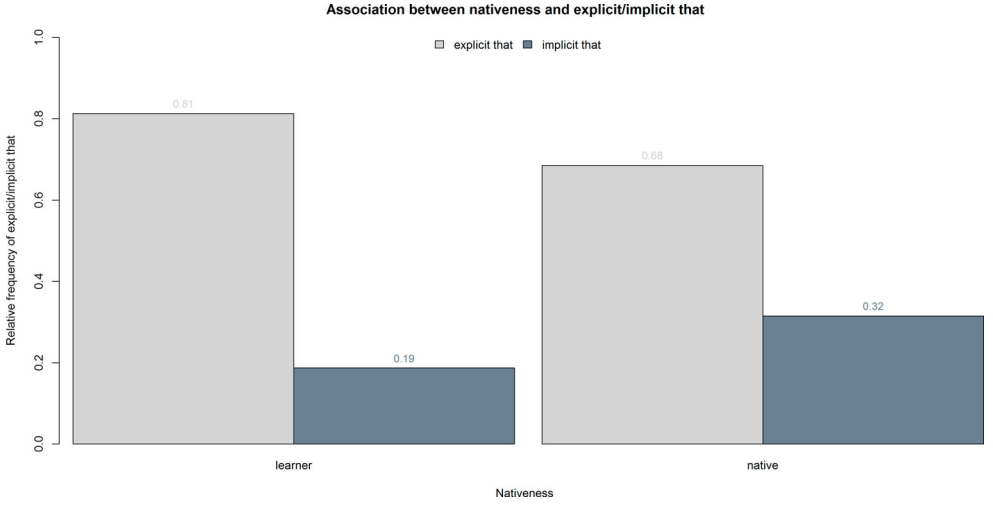

Figure 5. Association of nativeness and implicit and explicit *that* in English complement clauses introduced by *say* and *tell* (data: LOCNESS and ICLE-DU).

remains significant, contrary to our expectation. The full statistical model can be found in the appendix: nativeness, LengthMCVerbCCSubject, LengthMCSubjMCVerb and LengthMCSubject have significant main effects. The visualizations of the significant predictors' effects are shown in Figure 6. Again, there are no multicollinearity or overdispersion issues. The final model reaches an excellent accuracy score of 89% and is able to predict unseen data to a large extent (c = 0.95).

In the top-left plot, we see that native students have a higher inclination to use implicit *that* than L2 learners, although the latter also have a strong preference for explicit *that*. This means that, contrary to the effect of translation status in case study 1, nativeness directly influences the choice between explicit and implicit *that*. There are no significant interaction effects with nativeness and any of the other explanatory variables. However, there are four more main effects. The distance between the matrix-clause verb and the

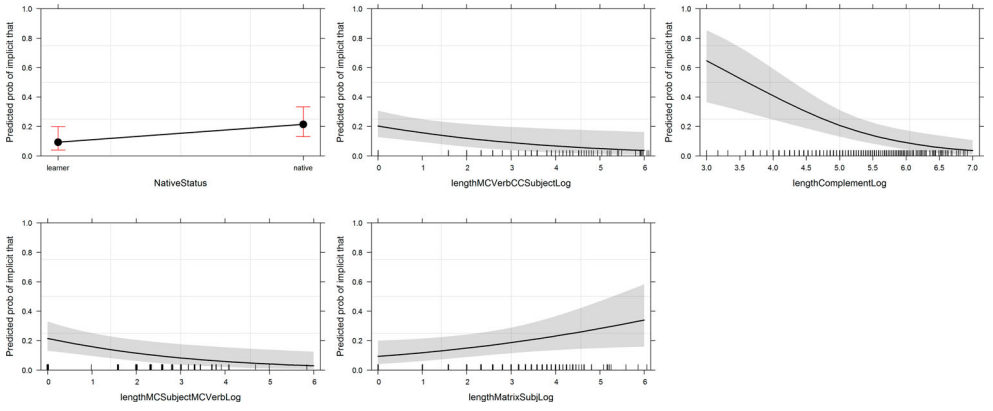

Figure 6. Effect plots of a generalized linear mixed effects model with explicit/implicit *that* as response variable, nativeness, register, lengthCIM, lengthComplement, lengthMatrixSubject, lengthComplementSubject, lengthMCSubjectMCVerb, lengthMCVerbCCSubject, lengthCCRemainder as fixed effects and verb lemma and text-id as random effects (data: LOCNESS and ICLE-DU, $n = 363$).

complement-clause subject negatively influences the use of implicit *that*: the greater the distance, the fewer occurrences of implicit *that*. The effect of complement-clause length goes in the same direction, but is stronger: short complement clauses trigger structures without an overt complement clause marker, whereas long complement clauses display a higher probability of including an overt clause boundary signal. Finally, the length of the matrix-clause subject has a positive effect on the use of implicit *that*, whereas the distance between the matrix-clause subject and the matrix-clause verb has a negative effect on implicit *that*. Except for the latter predictor, which did not prove significant in case study 1, all significant predictors in this case study have the same type of effect as in case study 1. Interestingly, we have also found a few cases of *that*-omission when the pronominal complement clause subject refers back to the (rather long) matrix-clause subject, hinting at the possible role played by cohesion (especially anaphora) in shaping the use vs. omission of the *that*-complementizer (see examples 6 and 7).

(6) *Many of the supporters who were later put to trail for the crimes they committed,* said **they** just followed the orders. (ICLE-DU)
(7) *Most of the undecided voters* said **they** would support capital punishment, if they had to vote on it immediately. (native student writing)

As already pointed out in Section 3.2, this interpretation remains very tentative at this stage (there are still many occurrences that do not fit into this pattern and that we are currently unable to account for).

In conclusion, our analysis shows that students choose between explicit and implicit *that* on the basis of a similar set of complexity-related variables to that of the professional writers in the previous case study, signifying that the novice writers' internalized probabilistic grammar is very similar to the professionals' (since the dataset of the second case study only contained argumentative essays, we could not investigate a register effect). Contrary to what was hypothesized and contrary to what we found in the previous case study

on translated and non-translated English, the difference between native and non-native student writing significantly influences the choice between explicit and implicit *that*, suggesting that Dutch learners of English resort more frequently to explicitly marked syntactic structures, which they might consider 'the "safe" strategy of realizing the complementizer as this choice is never, strictly speaking, ungrammatical, if only, at time, non-idiomatic' (Wulff et al., 2018, p. 118). As Dutch complement clauses are always introduced by overt *dat*, a transfer effect could also offer an additional explanation for the observed pattern.

When comparing the findings in case studies 1 and 2, we see a clear continuum in the use of explicit *that*: the group that uses explicit *that* most frequently, even in contexts where a zero marker would not have caused any comprehensibility problems, are the Dutch learners of English, who have the least experience in English and have little writing experience, followed by the native students, who are obviously more fluent in English, but also (presumably) have little writing experience in comparison with the last two categories, translators and non-translators, who use explicit *that* less often. As observed above, the two groups of professionals hardly differ, although in some very specific contexts translators use explicit *that* somewhat more often than non-translators.

All in all, then, our findings suggest that writing expertise has a major influence on *that* realization (separating students from professionals), whereas a varying degree of English language proficiency is of secondary importance (distinguishing the two student categories). Follow-up research could look into how writing expertise and language proficiency relate to the internalized probabilistic grammar of native and non-native students: do the students' internalized grammars need finetuning through experience and proficiency gains (a cognitive procedure), or does the lack of writing experience cause students to ignore their internalized grammar to some extent while opting for the safest option (a social procedure)? Also, it is important to bear in mind that other, confounding, variables might very well be in play here, such as directionality (different types of dual language activation), editing/revision (mediation) and register.

4. General conclusion and outlook

In this paper, we have set out to propose a new multifactorial, multi-methodological and interdisciplinary research agenda for empirical translation studies. Our basic, starting-point assumption is that translation products and processes are multifaceted and multidimensional. They should therefore be approached as such, rather than monofactorially. The methodological consequences of our agenda is that multi-methodological designs and advanced statistical modeling are essential tools, and that understanding translation inevitably entails an interdisciplinary approach to translation, building on theoretical frameworks and findings from neighboring disciplines, including, but not restricted to, variational corpus linguistics, bilingualism studies and (cognitive) sociolinguistics.

In our illustrative case studies on *that* complementation, we have shown the added value of two aspects of this new research agenda – multifactoriality and interdisciplinarity. We are very well aware, however, that there is still room for improvement or further refinement of the approach. First, other factors should be integrated into the statistical model, such as, for instance, overall syntactic complexity, text length, writing experience/translation expertise, use of computer-aided tools, editorial intervention/revision

and L2 proficiency of the learners of English. Second, our study should be extended to include a close comparison of source and target texts in the DPC, and source languages and mother tongue backgrounds other than Dutch. It also remains to be seen which feature(s) of constrained communication the complementizer *that* actually indexes – risk avoidance, conventionalization, shining through, or a combination of these (see also Kruger, in press).

We believe that the time is ripe for a new era in empirical translation studies. More generally, looking over the disciplinary fence should lead us, translation scholars, to broaden our perspective. This can be done in different ways: by moving beyond the handful of 'teddy-bear' operationalizations of explicitness (e.g., *that* alternation, cohesive devices, full vs. contracted forms), simplification (e.g., lexical density, average sentence length) and normalization (e.g., lexical bundles) we have held on to for more than twenty years, thereby exploring new, more sophisticated linguistic indicators, but also by investigating other phenomena that can potentially help us to characterize translated text, starting with linguistic features that have been said to typify other forms of constrained communication, such as non-native language varieties, editing and student writing.

Acknowledgements

The authors would like to thank Laura Penha-Marion for her help with the coding of the *tell* data and the anonymous reviewers for helpful comments on an earlier version of this article. The first author is grateful to the Faculty of Arts and Philosophy at Ghent University for awarding a sabbatical leave from February to September 2018.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Faculty Research Fund of The Faculty of Arts and Philosophy at Ghent University.

Notes on contributors

Gert De Sutter is Associate Professor in translation studies and Dutch linguistics in the Department of Translation, Interpreting and Communication at Ghent University. He also heads the research unit *EQTIS* (*Empirical and Quantitative Translation and Interpreting Studies*) at the same institution. His research can be broadly characterized as empirical research of linguistic variation in translated and non-translated texts, mainly using corpus data and advanced multifactorial statistical techniques. The overall aim of his research is to get a better insight into the complex interaction of social norms and bilingual cognition in translators.

Marie-Aude Lefer is Associate Professor of translation studies at the Louvain School of Translation and Interpreting, where she teaches translation studies, corpus-based translation studies and English-to-French translation. Her research interests include empirical translation studies, corpus linguistics, contrastive linguistics and lexicology. She has taken part in several corpus collection initiatives at UCLouvain and UBologna. She is the co-director of the *Multilingual Student Translation* (MUST) project, which aims at collecting a large multilingual corpus of learner translations.

ORCID

Gert De Sutter http://orcid.org/0000-0002-1998-6151

References

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). Philadelphia, PA/Amsterdam: Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, *7*(2), 223–243.

Baker, M. (1999). The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics*, *4*(2), 281–298.

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, *9*(2), 167–193.

Becher, V. (2010). Abandoning the notion of 'translation-inherent' explicitation: Against a dogma of translation studies. *Across Languages and Cultures*, *11*(1), 1–28.

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, *28*(1), 58–83.

Bisiada, M. (2014). Lösen Sie Schachtelsätze möglichst auf': The impact of editorial guidelines on sentence splitting in German business article translations. *Applied Linguistics*, *37*(3), 354–376.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherstone, & W. Sternefeld (Eds.), *Roots: Linguistics in search of its evidential base* (pp. 75–96). Berlin/New York: Mouton de Gruyter.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Delaere, I., & De Sutter, G. (2017). Variability of English loanword use in Belgian Dutch translations: Measuring the effect of source language and register. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 81–112). Berlin/Boston, MA: Mouton De Gruyter.

Divjak, D., & Gries, S. T. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory*, *2*(1), 23–60.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61–78.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts: A multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 47–80). Berlin, Boston: De Gruyter.

Gaspari, F., & Bernardini, S. (2010). Comparing non-native and translated language: Monolingual comparable corpora with a twist. In R. Xiao (Ed.), *Using corpora in contrastive and translation studies* (pp. 215–234). Newcastle, UK: Cambridge Scholars.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International corpus of learner English v2*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge, UK: Cambridge University Press.

Gries, S. T. (2018). On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies*, *1*(2), 276–308.
Halverson, S. (2017). Gravitational pull in translation: Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 9–46). Berlin/Boston, MA: Mouton De Gruyter.
Heilmann, A., Serbina, T., & Neumann, S. (2018). Processing of grammatical metaphor: Insights from controlled translation and reading experiments. *Translation, Cognition & Behavior*, *1*(2), 195–220.
Ji, M. (2016). *Empirical translation studies. Interdisciplinary methodologies explored*. Sheffield, UK: Equinox.
Kajzer-Wietrzny, M., Whyatt, B., & Stachowiak, K. (2016). Simplification in inter- and intralingual translation: Combining corpus linguistics, key logging and eye-tracking. *Poznan Studies in Contemporary Linguistics*, *52*(2), 235–268.
Kirkpatrick, A. (Ed.). (2010). *The Routledge handbook of World Englishes*. London, UK & New York, NY: Routledge.
Kruger, H. (2017). The effects of editorial intervention: Implications for studies of the features of translated language. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 113–156). Berlin/Boston, MA: Mouton De Gruyter.
Kruger, H. (2018). Expanding the third code: Corpus-based studies of constrained communication and language mediation. In S. Granger, M.-A. Lefer, & L. Penha-Marion (Eds.), *Book of abstracts. Using corpora in contrastive and translation studies conference (5th edition)*. CECL papers 1 (pp. 9–12). Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain.
Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties. Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, *1*(2), 251–290.
Kruger, H. (in press). *That* again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*.
Kruger, H., & Van Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures*, *13*(1), 33–65.
Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and non-native indigenised varieties of English. *English World-Wide*, *37*(1), 26–57.
Kruger, H., & Van Rooy, B. (2018). Register variation in written contact varieties of English. *English World-Wide*, *39*(2), 214–242.
Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121.
Lapshinova-Koltunski, E. (2017). Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies: New methodological and theoretical traditions* (pp. 207–234). Berlin/Boston, MA: Mouton De Gruyter.
Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, *43*(4), 557–570.
Laviosa, S. (2000). TEC: A resource for studying what is 'in' and 'of' translational English. *Across Languages and Cultures*, *1*(2), 159–177.
Laviosa, S. (2002). *Corpus-based translation studies. Theory, findings, applications*. Amsterdam/New York, NY: Rodopi.
Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, *56*(2), 374–390.
Malamatidou, S. (2018). *Corpus triangulation: Combining data and methods in corpus-based translation studies*. London: Routledge.

Mauranen, A. (2000). Strange strings in translated language: A study on corpora. In M. Olohan (Ed.), *Intercultural faultlines. Research models in translation studies 1: Textual and cognitive aspects* (pp. 119–141). Manchester, UK: St. Jerome.

Oakes, M. P., & Ji, M. (Eds.). (2012). *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research*. Philadelphia, PA/Amsterdam: Benjamins.

Olohan, M. (2004). *Introducing corpora in translation studies*. London/New York, NY: Routledge.

Olohan, M., & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141–158.

Pijpops, D., & Speelman, D. (2017). Alternating argument constructions of Dutch psychological verbs: A theory-driven corpus investigation. *Folia Linguistica*, *51*(1), 207–251.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Redelinghuys, K., & Kruger, H. (2015). Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics*, *20*(3), 293–325.

Rohdenburg, G. (1996). Cognitive complexity and increased Grammatical explicitness in English. *Cognitive Linguistics*, *17*(2), 149–182.

Schönefeld, D. (2011). *Converging evidence: Methodological and theoretical issues for linguistic research*. Philadelphia, PA/Amsterdam: Benjamins.

Seidelhofer, B. (2013). *Understanding English as a Lingua Franca*. Oxford: OUP.

Szmrecsanyi, B., Grafmiller, J., Heller, B., & Röthlisberger, M. (2016). Around the world in three alternations. *English World-Wide*, *37*(2), 109–137.

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135–178.

Wulff, S., Gries, S. T., & Lester, N. (2018). Optional *that* in complementation by German and Spanish learners. In A. Tyler, L. Huan, & H. Jan (Eds.), *What does applied cognitive linguistics look like? Answers from the L2 classroom and SLA studies* (pp. 99–120). Berlin/Boston: Mouton De Gruyter.

Zanettin, F., Saldanha, G., & Harding, S. A. (2015). Sketching landscapes in translation studies: A bibliographic study. *Perspectives*, *23*(2), 161–182.

Appendix

Generalized linear mixed effects model in case study 1

Random effects:

Groups	Name	Variance	Std.Dev.
File	(Intercept)	1.55	1.246e
VerbLemma	(Intercept)	4.254e-17	6.522e-09

Number of obs:813, groups: File, 306; VerbLemma, 2

Fixed effects:

	Estimate	S.E.	z value	
(Intercept)	3.85281	0.88491	4.354	1.34e-05***
TransStatustranslatedEnglish	1.69908	0.95746	1.775	0.075969
RegisterJournalistic	-0.04070	0.59841	-0.068	0.945777
RegisterPolitical	-2.81585	0.81940	-3.436	0.000589
LengthMCVerbCCSubjectLog	-0.63899	0.08854	-7.217	5.33e-13
LengthComplementSubjLog	-0.31744	0.10545	-3.010	0.002610
LengthComplementLog	-0.46657	0.13035	-3.579	0.000344
LengthMCSubjLog	0.30073	0.07909	3.802	0.000143
TransStatustranslatedEnglish: RegisterJournalistic	-3.23474	1.15910	-2.791	0.005259
TransStatustranslatedEnglish: RegisterPolitical	-0.35806	1.10545	-0.324	0.746009
TransStatustranslatedEnglish: LengthComplementSubjLog	-0.62148	0.23637	-2.629	0.008558

Generalized linear mixed effects model in case study 2

Random effects:

Groups	Name	Variance	Std.Dev.
File	(Intercept)	2.56	1.6
VerbLemma	(Intercept)	0.00	0.0

Number of obs: 363, groups: File, 186; VerbLemma, 2

Fixed effects:

	Estimate	S.E.	z value	
(Intercept)	3.0423	1.3395	2.271	0.02314
NativeStatusnative	0.9800	0.4601	2.130	0.03316
lengthMCVerbCCSubjectLog	-0.3154	0.1370	-2.302	0.02132
lengthComplementLog	-0.9747	0.2621	-3.719	0.00020
lengthMCSubjectMCVerbLog	-0.3713	0.1386	-2.680	0.00737
lengthMCSubjLog	0.2691	0.1283	2.097	0.03602