Language power relations and linguistic patterns in translation: A multilingual, corpus-based investigation

Matthew Riemland, M.Phil., B.A.

Thesis submitted for the degree of Doctor of Philosophy
School of Applied Language and Intercultural Studies

Dublin City University
July 2024

Supervised by:
Professor Dorothy Kenny (Dublin City University)
Dr. James Hadley (Trinity College Dublin)

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Matthew Riemland

ID No.: 20214807
Date: 8 July, 2024

Acknowledgements

It is only fitting for me to start by thanking my parents, Susan and James, who may never understand how vital their love and support have been to the production of this thesis, as well as everything that I did before and will do after it. The only thing I can do to repay them, if only partially, is to excuse them from reading it. I would also like to thank my brothers, Michael and Kevin. Along with my parents, they have helped create the comforts of home that I have often needed to return to, and without which I could not have completed this thesis. My extended family has also shown tremendous support and enthusiasm for my academic endeavors, and they deserve much credit.

I am profoundly fortunate in the supervision I have received over the course of my doctoral research. I would like to thank Professor Dorothy Kenny, whose generosity has somehow outshined her brilliance. I would also like to thank Dr. James Hadley, who encouraged and convinced me to pursue a PhD in the first place – an idea that would have never crossed my mind. Apart from my two supervisors, I would like to extend my sincerest gratitude to the rest of my colleagues in SALIS; I could not have asked for a better community.

I have been lucky enough to receive guidance and support from many outside my department, as well. I would like to thank Professor Marie-Aude Lefer, Dr. Antonio Toral, Dr. Lauren Cassidy, Dr. Mícheál John Ó Meachair, Professor Kevin Scannell, Anna Furtado, Professor Beatriz Sánchez Cárdenas, Professor Alice Delorme Benites, Dr. Juan Antonio Pérez Ortiz, and Dr. Chris Mellinger, among many others. I would also like to thank Dr. Andrew Mills, my earliest mentor, who helped jump-start my academic career when I was an undergraduate at the University of Michigan, then shipped me off to Europe as soon as he got the chance.

I am forever indebted to Dr. James Normington, whose knowledge of and enthusiasm for statistics have been invaluable to this thesis. I am indebted to a lesser degree to my friends Joseph Oswald and David Schulz – I will not elaborate. And I would like to thank Karen Reyes for being a thoughtful and supportive friend whose perspective and sincerity I admire. Of course, I am grateful toward many others who have played any part in this journey, whether in Ireland, back home in the United States, or anywhere else.

Lastly, I would like to express my sincerest gratitude to every person who has contributed to the work of the Postgraduate Workers Organisation within our local DCU branch and across Ireland. I dedicate this thesis to all of you.

Contents

List of Tables

List of Figures

Abstract

Language power relations and linguistic patterns in translation:
A multilingual, corpus-based investigation

Matthew Riemland

Formative works in descriptive translation studies assert that language power relations – asymmetries between the "status" or "prestige" of source languages (SLs) and target languages (TLs) – broadly determine translations' linguistic features (Baker 1996, 183; Toury 2012, 314). To date, these claims have not been tested in any systematic, empirical investigation involving a variety of languages and linguistic features. The central research question addressed by this doctoral thesis is thus whether translations from comparatively higher-status SLs tend to exhibit higher levels of SL influence, conceptualized as interference and foreignization.

The project applies comparable corpus methodology. It constructs a corpus of literary prose from the late 19th and early 20th century, where texts are either translated into or originally composed in English, French, German, Italian, Swedish, Croatian, or Irish. Using a novel method of assessing language status developed from Lewis and Simons' (2010) EGIDS model, the relative status for each selected language is expressed ordinally and synchronically. The thesis subsequently conducts corpus-based studies measuring the potential association between SL status and SL influence on the lexical, syntactic, and paratextual features of translations. Lexical interference is operationalized as the relative frequency (RF) of loanwords originating in the SL and attributable to the translator. Syntactic interference is operationalized using a novel metric called the syntactic interference/normalization coefficient (SINC), which measures the extent to which a translation's RF distribution of part-of-speech (POS) n-grams resembles those of comparable SL and TL texts. Paratextual foreignization is operationalized as the RF of translator-attributed footnotes and endnotes. The studies test for the hypothesized positive association between SL status and each of the aforementioned response variables using the Kendall rank correlation coefficient. Finally, the results of the three studies are synthesized to determine whether there is a positive association between SL status and SL influence on translations' linguistic properties.

1. Introduction

1.1. Context and motivation

The history of language is fraught with conflict. Inevitably, the geographical trajectories of diverse language communities have collided, often leading to volatile confrontations between their respective tongues. Such instances of language contact, whether gradually sustained or transpiring abruptly, have precipitated the many transmutations, ascents, and demises of the world's various languages, of which there are currently some seven thousand, with only a few dozen spoken by the majority of the world population (Ostler 2005, 527-528). Considered from a distance, it is undeniable that the languages used on a mass scale at one or another time in history have generally risen to prominence via "conquest, commerce, and conversion" (De Swaan 2001, 7), coinciding with primary sites of human struggles for social influence and power.

A frequent byproduct of language contact scenarios is cross-linguistic influence (CLI), where properties of one language are imprinted on the other (Kotze 2021, 115). If power, in the most intuitive sense, constitutes some general capacity to exert influence, then it is naturally expected that CLI is largely governed by power relations between languages. Empirically-based work in contact linguistics has repeatedly confirmed this intuition with respect to gradual language change (see Hoffer 2002; Rollason 2005; Haspelmath 2009, 35; Kotze 2021, 125). As a specific form of language contact, translation has also been projected to reflect CLI in proportion to language power dynamics, yet this area has inspired far less empirical research despite the prominent theme of power in the relevant scholarship.

As Marais (2014, 187) writes, translation studies "has focused, similar to literary studies, postcolonial studies, and even history, on power struggles." Nevertheless, this interrogation has largely taken place on theoretical or anecdotal grounds, even after the discipline's markedly empirical turn in the 1990s (Snell-Hornby 2006, 115-116). Toury's (2012) highly influential effort to solidify a more descriptive and thus empirical arm of translation studies culminates in an assertion of the allegedly universal tendency for translations' features to reflect power relations between their source language (SLs) and target languages (TLs). He hypothesizes that translators working in a given TL tend to be more tolerant of "interference" – i.e., the cumulative effects of linguistic and cultural peculiarities of the SL being reproduced in target texts despite contrasting with TL conventions – when "carried out from a 'major' of highly prestigious language/culture" into a comparatively "minor" or "weak" language and culture (Toury 2012, 314). Similarly, Baker (1996, 183) tentatively hypothesizes that the likelihood of translations exhibiting "normalisation" – i.e., the "tendency to exaggerate features of the target language and to conform to its typical patterns" – could decrease "the higher the status of the source text and language." Supposing "prestige" and "status" to constitute intertwined expressions of languages' power (as will be thoroughly discussed in Chapter 2), Toury's and Baker's hypotheses represent the CLI expected to result from language contact scenarios: more powerful SLs will tend to induce more interference in translation, given comparatively weaker TLs' susceptibility to their influence. In the early days of descriptive translation studies, these preliminary assertions of the expected correlation between language power relations and interference in translation were explicitly intended to be subjected to rigorous empirical testing.

In fact, it was Baker's (1993, 1996) introduction of corpus methodology to translation studies which had endowed the discipline with a new descriptive potential on a scale which seemed to match Toury's aspirations of developing translation theories with real explanatory power. However, while corpus-based translation research has flourished, there have been few such empirical investigations of the effects of language power dynamics on interference in translation, all confined to highly limited contexts (see Mauranen 2004; Becher et al. 2009; Evert and Neumann 2017). Many other subdisciplines in translation studies have since come to fruition and criticized the descriptive tradition for neglecting intergroup power dynamics, despite its foundational emphasis on power asymmetries between languages and cultures (Assis Rosa 2023, 202-203). Even so, the more qualitative and prescriptive approaches to translation and language power dynamics often adopt a similar view to Toury and Baker. Nowhere is this tendency more apparent than in Venuti's highly impactful *The Translator's Invisibility* (1995), in which he asserts that the global dominance of English leads anglophone translators to formulate strategies which "domesticate" (i.e., normalize) target texts according to English-language and Anglo-American cultural conventions, and consequently proclaims translators' moral imperative to counteract this phenomenon by applying "foreignizing" strategies to translation. However, as observed by Assis Rosa (2023, 205), the sum of "[c]ommitted research intent on not only describing but changing power relations" in translation studies has only been possible thanks to the "contextually informed descriptive approach to translation as a social activity, constrained by *prestige and power relations* [emphasis added]" between languages and cultures under their specific historical conditions. The necessity of the descriptive approach in laying the foundation for more prescriptive calls gives all the more reason to return to corpus methodology – the branch's preeminent methodological tool – as the basis for empirically-grounded accounts of language power relations in translation.

De Sutter and Lefer (2020) have reflected on the role of corpus methodology in translation research thus far and appraised both the achievements as well as the shortcomings of this empirical framework, ultimately formulating an updated research agenda with a renewed focus on translation's inherent interdisciplinarity and multidimensionality. They explicitly call for corpus-based translation research to draw from the work of related fields such as sociolinguistics and dedicate more efforts toward the investigation of hitherto neglected factors such as "source-language prestige" in affecting translations' observable linguistic features (ibid., 5-6). While the notable shortage of research on "prestige" in translation reflects the continued neglect of Toury's original hypothesis, it is undeniable that corpus methods remain vital in translation research today. Asscher (2022) advocates for Toury's theoretical approach as the optimal basis for describing the linguistic features of modern, corpus-driven machine translation (MT), which has become an enormously popular application and research area in natural language processing (NLP) over the past several decades. But despite its breadth of scholarly interest, MT research has neglected to investigate the possible manner in which the "asymmetry born of unequal power relations" between languages and cultures influences MT outputs (Asscher 2023, 8).

Evidently, the question of power has been central to translation studies and its numerous developments since the discipline adopted Toury's descriptive agenda and acknowledged translation's indispensable cultural aspects. Yet despite these developments, Toury and Baker's original hypotheses on the effects of language power dynamics on translations' linguistic features have not been tested in a systematic fashion. This project aims to address this research gap.

1.2. Aims of the research project

The overarching aim of this thesis is to determine whether there is empirical evidence for a general (i.e., observable across a diverse range of language pairs) correlation between language power differentials – namely, the comparative power of SLs relative to TLs – and the manifestation of SL linguistic conventions in translation. This goal necessarily entails some crucial terminological distinctions. The project takes "prestige" and "status" (as briefly encountered in the preceding section) as fundamentally distinct yet complementary expressions of language power. These terms have often been conflated in both translation studies (see Baker 1996, 183; Toury 2012, 314) and sociolinguistics (Mackey 1989, 4 cited in Edwards 1996, 703). The literature review conducted in the following chapter parses the core concepts underlying these terms in sociolinguistics and evaluates their amenability to the current project's overarching goal; it ultimately determines language status to be the ideal expression of language power for a project of this nature.

Similarly, the diametric conceptualization of a continuum of SL- and TL-oriented strategies that is recurrent in translation studies has undergone numerous processes of renaming despite only minor conceptual adjustments, including not only interference vs. normalization but also semantic vs. communicative translation, overt vs. covert translation, adequate vs. acceptable translation, foreignizing vs. domesticating translation, and so forth (Blumczynski and Hassani 2019, 8). As articulated later on, this thesis adopts the interference vs. normalization continuum to describe SL-oriented and TL-oriented translation strategies on (purely) linguistic levels (e.g., translations' lexical and syntactic features), and frames translations' paratextual features in terms of Venuti's foreignization vs. domestication. The term "interference" is preferred above the alternative "shining-through" given its direct lineage to Toury and the apparent predilection for it among other researchers. The term "SL influence" is introduced in order to refer jointly to interference and foreignization and align these concepts terminologically and conceptually with “cross-linguistic influence” as described in contact linguistics. Thus, (relative) SL status constitutes this project’s explanatory variable while SL influence constitutes its response variable.

To test for correlation between these two variables, it is necessary to devise a systematic classification system for language status. A secondary aim of this thesis is to provide corpus-based translation studies with a replicable, systematic classification model for language status, in order to enable the empirical study of this variable as a potentially influencing factor on translation products and processes more generally. Translation studies currently lacks a classification system of this kind. The language status assessment model must also allow for a valid comparison of language power dynamics as they manifest in various contexts and domains – across time periods, text types, geopolitical realms, and so on. That is, a translation-focused method for categorizing language status must be workable for all observable language pairs and, more broadly, the greater geopolitical and historical contexts in which these language pairs are situated.

Another secondary aim of this thesis is to devise language-agnostic operationalizations of different forms of SL influence so as to enable cross-lingual comparisons. Such cross-lingual comparisons are indispensable, as the project necessarily involves a diverse range of languages and language pairs. Furthermore, these replicable operationalizations are intended to ensure the comparability of these project’s results with any future studies involving different language pairs in different contexts. Given the assumed universal application of Toury’s and Baker’s complementary hypotheses, the intrinsic comparability of results and methods with future work is essential.

Finally, this project also aims to reaffirm the overlapping interests of translation studies with contact linguistics, sociolinguistics, and NLP research by emphasizing translation as a specific form of language contact and demonstrating the potential for corpus-based translation research to both benefit from and enhance each of these related fields.

1.3. Research question and hypothesis

This project's central research question is formulated as follows:

RQ1: Is SL status positively associated with SL influence in translation?

In order to answer this overarching research question, it is necessary to formulate three lower-level research questions:

RQ2: Is SL status positively associated with lexical interference in translation?

RQ3: Is SL status positively associated with syntactic interference in translation?

RQ4: Is SL status positively associated with paratextual foreignization in translation?

Each of these lower-level research questions form the respective bases of the project's constituent studies. This thesis adopts as its central hypothesis the assumption that SL status is positively associated with SL influence in translation, where SL influence assumes the various forms indicated in RQ2, RQ3, and RQ4. Confirmation of this central hypothesis necessitates affirmative results in all three of the project's studies.

1.4. Synopsis of research design

For each of the constituent studies, this project deploys a bivariate research design involving SL status as its explanatory variable and SL influence as its response variable. A language status assessment model is devised and applied to a range of selected languages. In order to measure the hypothesized association between SL status and SL influence in translation empirically, the project constructs a multilingual comparable corpus composed of both non-translated and translated literary fiction (prose) both in and between the selected languages. Comparable corpus methodology is used to measure the level of SL influence in all translations on the lexical, syntactic, and paratextual levels. These various forms of SL influence are assessed in relation to SL status.

As will be articulated in Chapters 3 and 4, language status is operationalized as an ordinal variable, while all forms of SL influence are operationalized as continuous variables. Therefore, the project tests for the *association* between the variables using Kendall's rank correlation as its primary statistical test. Translations are grouped into subcorpora according to their common TLs and SLs, where the hypothesized associations are subsequently tested for statistical significance. In order to provide more granular analyses, the project also uses a variety of secondary tests based on lists of the data points ranked according to alternative groupings. This sequence of primary and secondary data analyses is conducted for each of the project's three studies. Collectively, their results are synthesized to answer the central research question.

1.5. Structure of the thesis

Chapter 2 conducts a review of the relevant literature across the interrelated disciplines of sociology, sociolinguistics, translation studies, and natural language processing. It starts by summarizing Pierre Bourdieu's work on linguistic capital, then illustrates its commonalities with the concept of language status in sociolinguistics. The chapter then traces the underlying theme of language power relations in translation studies and, subsequently, corpus-driven machine translation, demonstrating language status to be the missing variable in the core research agenda of descriptive, empirical translation research.

Chapter 3 reviews previous attempts to systematize assessments of language status (or closely related concepts), and ultimately develops a novel language status assessment model on the basis of core concepts from sociolinguistics and language vitality. It selects a range of languages to be used in this project, then uses the newly devised language status assessment model to rank them based on their status relative to one another.

Chapter 4 details and justifies the methodology used in this thesis. It first discusses the principles used to design and construct the comparable corpus, then articulates the rationale for the project's bivariate design as well as the nature of its primary and secondary data analyses.

Chapter 5 tests for a positive association between SL status and lexical interference in translation. It operationalizes lexical interference as the relative frequency of translator-attributed loanwords. Chapter 6 tests for a positive association between SL status and syntactic interference in translation. It operationalizes syntactic interference and its converse, syntactic normalization, using a novel method based on comparisons between the frequency distributions of part-of-speech sequences in translations and comparable texts in their respective SLs and TLs. Chapter 7 tests for a positive association between SL status and paratextual foreignization in translation. It operationalizes paratextual foreignization as the relative frequency of translator-attributed footnotes.

Chapter 8 summarizes the project's theoretical foundation and methodology as well as its results and contributions. It then synthesizes the constituent studies' findings into an overarching discussion of the project's outcome in relation to other works in translation studies and adjacent disciplines. The chapter then critically reflects on the broader limitations of its research design and approach before offering a detailed discussion of the possibilities for future research conveyed by its findings. Finally, the project provides some brief remarks on the utility of corpus methodology in investigating language power relations in translation studies and related fields.

2. Literature review: translation and language power relations

2.1. Chapter introduction

This chapter conducts a wide-ranging review of the literature related to language power dynamics and translation, while the individual studies in Chapters 5-7 contain more targeted literature reviews related to SL influence on their respective linguistic levels. This literature review first outlines and adopts Bourdieu's concept of linguistic capital as the project's basis for conceptualizing language power. It then discusses sociolinguistic accounts of language status and prestige in relation to Bourdieu's work, identifying language status as an ideal foundational concept for operationalizing language power in the context of the thesis' aims and highlighting its key determinants. The chapter subsequently establishes power as a central theme in translation studies, charting the rise of descriptive translation studies and emphasizing its central tenets. It then demonstrates that the foundational hypotheses of descriptive and corpus-based translation studies – as put forth by Toury and Baker – remain unexplored in systematic terms. Finally, the chapter examines relevant research on machine translation (MT) and natural language processing (NLP) more broadly to illustrate the substantial research opportunity that these hypotheses (and their corresponding methodological and theoretical approaches) offer researchers in these areas.

2.2. The basis of language power

There are seemingly countless ways to characterize and approach the intersection of language and power. The simple matter of where to begin in attempting to theorize and articulate this relation in operational terms is by no means straightforward, as any intuitive account is almost certain to prove simplistic. In very general terms, the link between language and power might logically begin with reference to languages' historical development and spread. The world's most prominent and widely-used languages have tended to be the languages of “commerce, conquest, and conversion” (De Swaan 2001). Still, as painstakingly depicted by Ostler (2005, 556) in his sprawling volume *Empires of the Word*, which charts the rise and fall of prominent languages throughout world history, economic and military power may be highly influential factors, yet they ultimately offer insufficient explanations of language spread, as “world languages are not exclusively the creatures of world powers.” To describe and compare the power of languages thus necessitates a wide-reaching theoretical framework of the myriad ways in which power manifests in complex social relations.

Widely considered the “most influential sociologist in Europe” (Phillipson 2008, 26), Pierre Bourdieu (1930–2002) constructed a comprehensive framework for understanding the nature of power dynamics in society at large as well as its manifold subsets. Many characteristics of his thought underscore the methods and aims of this thesis. Mirroring the initial motivation for descriptive translation studies (explored later on), Bourdieu (1991, 36) regards the early tradition of linguistics as “the intellectualist philosophy which treats language as an object of contemplation rather than as an instrument of action and power.” Though any straightforwardly operational account of power’s elusive and multifaceted nature should merit criticism, Bourdieu offers what is perhaps a semi-satisfactory compromise between the strained simplicity of structuralism and the stubbornly deconstructive tendencies of post-structuralism. His work exhibits “a firm commitment to the value of empirical investigation” and “makes no apologies for his use (at times extensive) of statistical and quantitative methods,” all while retaining a “sharp critical edge” (Thompson 1991, 31). The intellectual attractiveness of this balance is perhaps why Bourdieu’s thought continues to be a major influence in translation studies (see Heilbron and Sapiro 2007; Wolf 2007; Assis Rosa 2010, 95; Hermans 2019, 146) as well as sociolinguistics (see Blommaert 2015; Pennycook 2022, 15).

The Frenchman’s body of work is premised on the notion that capital assumes additional forms beyond merely capital in Marx’s classic economic sense, materializing also in cultural, social, and even symbolic forms. These other forms of capital are still primarily derived from economic capital, though only via some manner of conversion or transformation that is not always possible (Bourdieu 1986, 253). Broadly speaking, capital may be understood as “accumulated labor (in its materialized form or its ‘incorporated,’ embodied form) which, when appropriated on a private, i.e., exclusive, basis by agents or groups of agents, enables them to appropriate social energy in the form of reified or living labor” (Bourdieu 1986, 241). As with the accumulation of economic capital, Bourdieu’s other forms of capital both embody and provide the means to expand upon previously successful efforts to attain social status via the social practices and behaviors into which they are encoded. It is through the widespread acceptance – conscious or unconscious – of this embedded capital that the relative status of social agents is determined and legitimized. In this sense, capital, regardless of its form, endows agents with the ability to *influence*, serving as a lever of power by which social agents reproduce their positioning within a hierarchical structure. Thus, according to this framework, capital is synonymous with power (ibid., 242). Social agents are generally oriented toward the aim of accruing capital (i.e., power), whether economic or symbolic: the former reflects the pursuit of profit while the latter reflects the pursuit of prestige (Thompson 1991, 15).

These concepts are applied to the development of languages in *Language and Symbolic Power* (1991), a collection of Bourdieu’s essays translated into English by Gino Raymond and Matthew Adamson and edited by John B. Thompson. Language communities, to the extent they may be treated as discernable entities, encounter one another in ways that reflect “relations of symbolic power in which the power relations between speakers or their respective groups are actualized” (Bourdieu 1991, 36). In his introduction to the translated essay collection, Thompson (1991) paraphrases the underlying idea driving Bourdieu’s groundbreaking theory:

> Through a complex historical process, sometimes involving extensive conflict (especially in colonial contexts), a particular language or set of linguistic practices has emerged as the dominant and legitimate language, and other languages or dialects have been eliminated or subordinated to it. (Thompson 1991, 6)

Dominant and legitimate linguistic practices – including languages themselves – accumulate *linguistic capital*, whose distribution both between and within languages is inextricable from the “distribution of other forms of capital (economic capital, cultural capital, etc.)" (Thompson 1991, 18). Despite its embeddedness in these other forms of power, linguistic capital may be isolated and distinguished according to languages' observable competitive (dis)advantages in language contact scenarios. Bourdieu's (1991, 69) chief example depicts a case in which a mayor in the French region of Béarn deigns to address the audience of a ceremony in Béarnese, the regional patois, despite the implicit expectation that the formal setting would call for French. The perceived goodwill behind this gesture is rooted in the broad acceptance of the hierarchical relation between French and Béarnese that creates an expectation of the former's appropriateness in formal settings. This implicit hierarchy reflects the higher linguistic capital of French. By observing and articulating the outcomes of languages competing in multilingual social contexts as in the example above, it is possible to formulate "a system of *specifically linguistic relations of power* [emphasis added] based on the unequal distribution of linguistic capital" (ibid., 58). Language power – or linguistic capital – may thus be abstracted from power in other forms, despite its inherent reliance on them.

The competitive pursuit of linguistic capital also occurs among intralingual variants. Consistently high levels of linguistic capital often provide the illusion of ("official") languages and language communities having stable and indisputable demarcations, where their historical formations and subsequent transformations are somehow incidental to their position within broader power struggles over economic, social, cultural, and symbolic capital. In reality, the boundaries of languages and their (imagined) communities are unfixed, constantly changing in response to speakers' practices and perceptions. Any widespread "recognition of the legitimacy of [an] official language" is built upon its gradual historical accumulation of linguistic capital at the expense of its competitors; the language's (or particular variant's) legitimacy is embedded in and perpetuated by the typical social behaviors of speakers, who maintain "a form of complicity which is neither passive submission to external constraint nor a free adherence to values" (Bourdieu 1991, 51). Despite the social expectations or pressures imposed by the linguistic capital of languages within their social orbit, speakers retain some degree of agency (albeit constrained) to resist these external forces in their linguistic exchanges according to Bourdieu's framework. In this way, Bourdieu's dual-faceted notion of linguistic capital coincides with more practical attempts by sociolinguists to distinguish and examine these different elements of language power, particularly with respect to their effects on the intersecting trajectories of competing languages.

2.3. Sociolinguistic accounts of language power

As described previously, the ever-changing distribution of language power or linguistic capital is predicated on the interplay between individual agents and society at large. Phillipson (2008, 29) paraphrases Bourdieu's view of linguistic capital by noting that it is "some combination of internal motivation and external pressure, push-and-pull factors" that governs its accumulation, much like the accumulation of economic capital in market societies. This dichotomy joins the irreducible autonomy of individual agents ("internal motivation") on the one hand with the broader sociocultural conditions to which they are subjected ("external pressure") on the other. Thus, language power may be defined in a dialectical relationship, where shifting and value-laden language attitudes of individuals or groups are nonetheless formulated in relation to a tacit consensus of languages' discernable social prominence and historical legacies. As will be demonstrated in this section, these two polarities roughly correspond to sociolinguistic concepts of *language prestige* and *language status*, respectively, the latter of which may be assessed and operationalized in a reasonably systematic fashion.

Ammon (1992, 421) comments on the historically haphazard use of terminology related to status, even by those for whom it is a primary focus. The terms "status" and "prestige" have often been conflated in both sociolinguistics (Mackey 1989, 4 cited in Edwards 1996, 703) and translation studies (Baker 1996, 183; Toury 2012, 314). Ammon (1989) offers straightforward definitions for these two concepts. According to him, a language's prestige is not a matter of its "social distribution" but rather of *attitudes* toward it, which may be either positive or negative (ibid., 69). The subjective, individualized character of language prestige is made evident in the Latin American context: Howard et al. (2018, 25) observe that, in Peru, "the stigma often attached to indigenous languages can lead to speakers being reluctant to avail themselves of the provision to which they are entitled, for fear of being considered inferior." This negative self-perception amounts to a negative form of language prestige (or "language stigma"), wherein speakers' attitudes toward their language contrast with its official state recognition and with the actual social opportunities it affords.

Language status denotes a language's "position (in the respective [language] system)" or its "rank (in a hierarchy or in a rank order)" (Ammon 1989, 26). Edwards (1996, 703) affirms this definition, discerning that "the status of a language is its position vis-à-vis others" – a definition in line with the etymological origin of *status*. It is perhaps tempting for the sake of convenience to conflate any given language's status with its recognition as an *official* language (most frequently at the national level), as the term is often informally applied. This narrow conceptualization of language status as a purely legal concept is generally not supported by sociolinguistics (Ammon 1992, 421). In fact, the manner(s) and extent to which a language is actually used across social contexts regularly conflict with the formal legal or state recognition it is granted (Ammon 1989, 26). This critical distinction may also be illustrated by turning to the example of (Hispanic) Latin America, where a wave of state reforms beginning in the 1980s granted official status to indigenous languages, yet in practice these changes often did little to correct the lack of indigenous-language translation and interpreting for public services (Howard et al. 2018). Language status also necessarily encapsulates the degree to which languages capture various social roles – or *domains* – outside of their formal political recognition, as further explored later on. This view is highly similar to the manner in which the market of linguistic capital reacts according to "a whole set of specific institutions and mechanisms" of which state intervention and planning (i.e., language policy) "form only the most superficial aspect" (Bourdieu 1991, 51). Thompson (1991, 9) explains that Bourdieu's use of "institution" is more closely aligned with the French *institution*, reflecting a much more dynamic concept than the English word typically indicates; Bourdieu uses "institution" to refer to "any relatively durable set of social relations which endows individuals with power, status and resources of various kinds." Conceptually speaking, language status is thus closely aligned with Bourdieu's portrayal of the macro-level indicators of linguistic capital.

As an expression of language power, language status is far more amenable to empirical inquiries than language prestige: the former's prerequisite arrangement of languages into a ranked hierarchy more readily inscribes power asymmetries into a quantitative form than the latter's inherently subjective nature. The thesis therefore focuses on language status as a specific and operational form of the overarching concept of language power. However, as will be outlined in the following chapter, previously formulated models for assessing language status prove unworkable and/or theoretically flawed, making it necessary to develop an operational language status assessment model for the purposes of this thesis. In order to accomplish this task, it is necessary to extract from the sociolinguistic literature a primary set of criteria needed to assess language status – that is, to rank or position languages in relation to one another.

In *Status Change of Languages*, Ammon and Hellinger (1992, viii) compile numerous case studies of shifts in language status, providing an extensive yet non-exhaustive list of the numerous contributing factors that may be "specified and operationalized in various ways – which shows the enormous complexity of the concept." The aim of operationalizing language status as an explanatory variable in empirical research is surely complicated by this "enormous complexity". Still, these various dimensions are at least somewhat interrelated, making it possible to identify some primary indicators of language status. Elsewhere in the same volume, Ammon (1992, 421) asserts that the status of a language principally varies according to "its speakers, i.e. their number and/or their social attributes, or in respect to the domains in which it is used." For Fishman (1991, 81), too, a decline in language status manifests primarily as "the shrinking number of *users* that a language has or… the meager importance of the *uses* with which it is commonly associated in its speech and/or writing community." Throughout sociolinguistic literature, these two criteria emerge as the primary components of language status: a language's domains of use (rather, the prominence of its demonstrable societal functions in the aggregate) and its degree or scale of use (frequently characterized by its number of users).

The emphasis on a language's associated domains as an indicator of its power in relation to other competing languages is recurrent throughout sociolinguistic literature. Stewart (1968, 540-541) categorizes possible sociolinguistic functions of languages, which may concern their geopolitical scope (e.g., regional, national lingua franca, or international) or their confined domains (e.g., literary or religious). According to him, situations are "stable" when languages are "geographically, socially, and functionally non-competitive" (ibid., 541). This deeply intertwined and competitive system is frequently characterized by sociolinguists as a metaphorical ecology. Edwards (1994, 142-143) notes that the term “ecology of language” was popularized by Haugen (1972), for whom “[language] status signifies the power, prestige and influence the language possesses through the social categorization of its speakers.” (Again, note the terminological overlap of status and prestige.) The notion of speakers’ social categorization as used here coincides with a language’s demonstrable domains of use, as the more prominent domains will naturally be dominated by more powerful individuals and, consequently, their languages. This phenomenon works to the detriment of low-status languages, as “the uses to which these languages are commonly put are not only few, but, additionally, they are typically unrelated to higher social status (prestige, power) even within their own ethnocultural community, this being a reflection of the relative powerlessness of the bulk of their users” (Fishman 1991, 81). It is evident that scholars in sociolinguistics have long considered a language’s domains of use to play a central role in determining its status. While the number of speakers of a language has also played a factor in this determination, it is widely considered to be less consequential than a language’s uses.

In his descriptive framework of common national multilingual scenarios, Stewart (1968, 542) posits a language’s degree of use (expressed as a language’s percentage of all possible speakers within the national boundaries) as a prominent factor of competition among languages, though stresses that this criterion “cannot be taken by itself as an index of relative sociolinguistic importance.” Of course, the influential reach of languages is not always confined or even relatable to national boundaries. The observation that languages spoken widely within a state’s borders do not necessarily gain prominence on a larger, transnational scale “leads directly to the issues of power, prestige and dominance which are often more important than mere numbers in determining majority or minority status” (Edwards 1994, 139). The linking of a language’s “power, prestige and dominance” with its geopolitical transmission and only secondarily with its number of speakers is an important aspect of the assessment of language status that will be developed in the subsequent chapter. Edwards (ibid., 139) illustrates this point by drawing attention to the South African language system, where he intuits the internationally used English as being higher in status than languages such as Xhosa or Zulu, despite these languages having greater numbers of speakers in the country.

Claims of the preeminence of a language's domains over the size of its user base warrant a historical example. Crystal (2003, 7) recounts the rise of Latin to its historical dominance as an "international language throughout the Roman Empire," an accomplishment owing less to the number of Latin speakers than to Roman military strength and, later, to the far-reaching power of the Catholic Church. He goes on to note that, throughout the Roman Empire, Latin was the widely used administrative language of the ruling elites, whereas conquered peoples were unlikely to speak it (ibid., 11-12). This historical case depicts the need to distinguish between the domains in which a language is used and the size of its pool of speakers, where the latter is clearly subservient to the former in determining language status.

Latin gradually declined in use, to the extent that it is now scarcely used outside of historical research and religious traditions. The totalizing decline in language status, actualized in the loss of both domains and speakers, is commonly referred to as language extinction, which represents the extreme end of a gradient continuum of language shifts. Matters of ethnolinguistic self-perception notwithstanding, such cases of language shift entail speakers flocking to a higher-status language, or a "language of greater power and opportunity" while abandoning their own low-status language (Fishman 1991, 16). In this manner, the speakers "collaborate in the destruction of their instruments of expression" (Bourdieu 1991, 50). These parallels between language status and what is referred to as language *vitality* – which depicts a spectrum expressing languages' endangerment or risk of extinction (see Fishman 1991) – are key to the language status assessment model developed in Chapter 3.

According to Fishman (1991, 59), language shifts are driven by power imbalances between language communities in terms of their size as well as their collective or associated political, economic, and cultural strength. Naturally, there are various stages in the language shift process, and language vitality is therefore necessarily gradient. Generally speaking, endangered (i.e., very low-status) languages are relegated to more peripheral domains of use. In Fishman's (ibid., 44) view, domains are "interactions that are rather unambiguously related (topically and situationally) to one or another of the major institutions of society: e.g. the family, the work sphere, education, religion, entertainment and the mass media, the political party, the government, etc." Less prominent domains ("lower levels") comprise "face-to-face, small-scale social life" whereas the most prominent domains ("higher levels") are associated with top-down social forces that are "more complex, more encompassing, [and] more power-related" (Fishman 1991, 4). The latter are those encompassing the "highest educational, occupational, governmental and media activities" (ibid., 107). The hierarchical relations between domains are central to the Graded Intergenerational Disruption Scale (GIDS), which is Fishman's (ibid., 87-111) systematic method of assessing language vitality.

As a theoretical and practical framework for describing and projecting shifts in the vitality of languages, GIDS bears close resemblance to Bourdieu's linguistic capital and reinforces the common sociolinguistic definition of language status outlined earlier. The fundamental interconnection between various forms of Bourdieusian capitals is reflected in Fishman's (1991, 59) observation that speakers of low-status languages tend to be "less educationally and economically fortunate" than members of competing language communities. In agreement with his peers, he identifies the two primary indicators of a language's decreasing status as its loss of speakers and relegation to marginalized social situations (ibid., 81). Although the GIDS framework may be described first and foremost as gauging language vitality or endangerment, Fishman (ibid., 87) seems to use vitality and status synonymously throughout his work, calling his model a "graded typology of threatened statuses". For these reasons, the GIDS framework (examined more closely in the following chapter) forms the core of the novel language status assessment model presented in this thesis.

So far, the geographical dimension of language status has not been properly addressed, apart from being an implicit aspect of domains. Naturally, competition among languages implies the existence of common territories where language communities encounter and influence one another. This point was already acknowledged by early sociolinguists: Stewart (1968, 541) describes multilingual situations as "stable" when languages are "geographically, socially, and functionally non-competitive" and Haugen (1972) famously describes languages inhabiting the same area as comprising an intertwined, competitive ecology. Sociolinguists initially concentrated on language ecologies within national boundaries. By the end of the 20th century, it was impossible to deny that the forces of globalization had in many ways unified the world's languages into a single ecology, with English emerging as the "hypercentral language that holds the entire world language system together" (De Swaan 2001, 17). Even so, competition between languages could still be observed on smaller scales of sustained interlingual interactions, pressurized by communities' geopolitical proximities.

According to De Swaan (2001), cohesive geopolitical units containing competing languages may be conceptualized as language constellations, which can align with multilingual nation-states (e.g., India, Indonesia, South Africa), supranational organizations (e.g., the European Union), and contiguous geographical regions (e.g., Sub-Saharan Africa). Dominant languages capture prominent domains of use whose scopes coincide with the boundaries of language constellations, such as region-wide commerce and mass media (see Lewis and Simons 2010). The nature of these domains also illustrates the interdependencies of different capitals. As noted by Phillipson (2008, 29), Bourdieu's notion of linguistic capital is best exemplified in the continued and growing domination of English in continental Europe, as the language has overtaken "key societal domains" such as commerce, scientific research, and higher education at the expense of local languages like Swedish and Danish. The substantive qualities of these domains reinforce the notion that the ascent of English has been driven by the far-reaching "economic, technological, and cultural power" of Anglophone nations and their imperial legacies (Crystal 2003, 7). This process of domain capture entails the disruption of local language ecologies, meaning that "weaker [languages] become physically and demographically dislocated" (Fishman 1991, 59).

The historical forces of colonialism and globalization have been paramount in determining the scopes and shapes of modern language constellations, as illustrated by the linguistic situation in Africa in the middle of the 20th century. Even as many colonies won independence from their European oppressors, colonial languages remained firmly implanted in higher social domains, while indigenous African languages were "for the most part confined to informal domains of use and had less overtly recognized "prestige" even where occurring as regional lingua francas among larger populations" (Simpson 2008, 3). In West Africa, for example, languages such as Hausa and Pulaar are commonly used as regionally specific lingua francas, though they are never considered as "international languages" or "languages of wider communication" – even in Africa (Kanana Erastus 2013, 59). Instead, these designations are reserved solely for English and French, the colonial holdovers in West Africa and much of the rest of the continent (ibid., 59). Independent African nations have often viewed these languages as indispensable links to the scientific and technological advancements of other countries (Adegbija 1994, 97). The African linguistic context makes evident that languages' global status positionings transfer to localized language constellations, overriding the sizes of local languages' speaking populations – even those of regional lingua francas. It also attests to the tendency for sociolinguistic domains to supersede national boundaries, an observation which prompted Lewis and Simons (2010) to expand Fishman's GIDS scale to include a category for *internationally* prominent languages, as will be explored in the following chapter. In Africa as in other parts of the world, it has been these few privileged international languages which have frequently overpowered their local competitors.

What forms do the competition between languages take? Languages' struggles for influence across domains and geographic areas are rarely one-dimensional, as there are numerous different ways in which languages come into contact and compete with one another. The field of contact linguistics aims to typify these various interactions, identifying their common elements and effects. Cross-linguistic influence (CLI) – referring to the emergence of "particular linguistic features as a consequence of the co-activation of two languages" – is a regular byproduct of language contact scenarios (Kotze 2021, 113). It has long been recognized that "borrowed lexemes, borrowed morphosyntactic features, and borrowed phonemes" are among the most apparent of these linguistic features, which are often imposed by a "prestige language" onto a comparatively disadvantaged language (Kahane 1986, 503). The preeminence of language power relations in determining CLI has also been adopted in translation studies.

As a type of language contact, translation is frequently conceptualized within translation studies (particularly its empirical branch) as a negotiation between the polarities of SL-oriented translation strategies (often called "interference" or "shining-through") and TL-oriented translation strategies (typically "normalization") (Kotze 2021, 119). As the following section will demonstrate, it has long been asserted that power relations between source and target languages dictate the degrees of interference and normalization evident in translated language, although this basic relationship has not been investigated in systematic and quantitative terms.

2.4. The undercurrent of language power in translation theory

Theories of translation preceded the widespread recognition of translation studies as a standalone discipline. Writing from the vantage point of German Romanticism, Friedrich Schleiermacher (1768–1834) offered an essential dichotomous framework for translation that, in many ways, still endures today. He contends that translation strategies may be sorted into two distinct categories: the translation either moves the writer toward the reader or the reader toward the writer (Schleiermacher 1816/2012). In the first option, the translator renders the translation in a manner that conforms to the linguistic and cultural conventions of the TL, omitting or downplaying the elements of the source text which may seem linguistically or culturally peculiar or foreign to the target readership (moving the writer toward the reader). In the second strategy, the translator reproduces SL peculiarities in the target text at the expense of its perceived naturalness within the TL context (moving the reader toward the writer). This framing constitutes what is perhaps the most essential of the many persistent binaries that have characterized translation thought over the years (Álvaro Marín García 2023, 13-16). Schleiermacher's dichotomy initiated the basic dichotomy of SL- and TL-oriented translation strategies that has been recast at various points as interference and normalization or foreignization and domestication, as explored later on. The role of (language) power in dictating translators' inclinations toward each of the polarities did not come into view until much later.

It was not until the decades following World War II that translation studies emerged as a discipline in its own right, with scholars such as Vinay and Darbelnet, Catford, and Nida initially approaching translation as a series of strictly linguistic operations intended to achieve incontestable semantic equivalences across languages (Malmkjær 2023). These early, linguistically-oriented approaches to translation are now commonly recognized as *prescriptive* in nature; like Schleiermacher before them, early translation theorists envisioned the formation of a set of rules to govern translation proper, failing to consider the irreducibly elusive nature of translation and its cultural dynamism.

Translation studies' prescriptive orientation persisted until the rise of descriptive translation studies, conceived in the 1970s and arguably reaching its apex in the 1990s (Assis Rosa 2010). This new branch was a pillar of translation studies' so-called cultural turn, which expanded the field far beyond its previous linguistic confinements, examining the influences of culture and, perhaps more importantly, power on translation (Snell-Hornby 2006, 47). It is here where the field's intersection with Bourdieu's work becomes apparent. Bourdieu (1991, 36) regarded the early linguistics tradition as an "intellectualist philosophy which treats language as an object of contemplation rather than as an instrument of action and power." The impact of the sociologist's thought was undeniable in descriptive translation studies, as the branch's most prominent figures – Even-Zohar, Toury, and Hermans – were eventually swayed to revise their approaches according to Bourdieu's work, having initially been "accused of overlooking questions concerning power relationships between social groups or polities" (Córdoba Serrano 2010, 251).

Starting in the 1970s, Even-Zohar's polysystem theory provided the foundation for a research agenda dedicated to the *description* of translation – translation not as it should be, but as it is actually practiced. Polysystem theory stipulates that all "sign-governed human patterns of communication" – e.g., language and literature – are governed by highly complex, interconnected systems of influence, collectively termed the polysystem (Even-Zohar 1990, 10). Within the polysystem there are "center-and-periphery relations" with these hierarchies of power embodying a "permanent struggle" to move toward the center of influence (Even-Zohar 1979, 293). Polysystem theory was envisioned as a framework for articulating and unraveling the complex interrelations between literature and other domains such as the political economy (ibid., 300). Even-Zohar (1990, 46) is very explicit in his view that it is these power struggles which principally govern the linguistic features (or linguistic "norms") of translated literature.

The principal byproduct of these power struggles is interference, which he posits as "a relation(ship) between literatures, whereby a certain literature A (a source literature) may become a source of direct or indirect loans for another literature B (a target literature)" (Even-Zohar 1990, 54). He insists that interference occurs in all literary systems, and cannot be abstracted from specific historical circumstances (Even-Zohar 1990, 54). The most prominent of the historical circumstances governing interference are power dynamics between languages and cultures. Naturally, contact between languages is necessary for interference to occur; translation is merely one form of contact, albeit a (potentially) highly influential one (ibid., 57). Interference may take many forms, such as Hebrew's lexical borrowings from Yiddish (ibid., 124). In situating interference within the polysystem framework, Even-Zohar also highlighted translation as a form of language contact capable of inducing CLI, as Kotze (2021) would later emphasize. The cultural and social linkages that polysystem theory offered translation studies laid the groundwork for subsequent theorists to formalize a hypothesis for the relationship between language power dynamics and interference.

A protégé of Even-Zohar, Gideon Toury has one of the most enduring legacies in translation studies. Toury's volume *Descriptive Translation Studies – and beyond* (originally published in 1995 and revised in 2012) presented a groundbreaking approach to the study of translation. This new outlook brought translation studies into the realm of scientific inquiry with its orientation toward the production of theories with real predictive and explanatory power. It eschewed previous scholars' heavily prescriptive notions of what constituted proper translation, embracing the pursuit of explanation and description as continuously refined through hypothesis testing. He envisioned that the empirical findings produced under this descriptive branch would lead to a "series of coherent *laws* which would state the inherent relations between all variables" acting on a translation; nevertheless, such laws would be "anything but absolute" and merely convey the "*likelihood* that a certain kind of behaviour, or surface realization, would occur under a particular set of conditions" (Toury 2012, 9-10).

This new perspective reflected a fundamental shift in the assumptions of the primary constraints under which translations were produced. The discipline had previously conceptualized translations as primarily source-oriented products, embodying a set of mechanical linguistic operations necessary to achieve a notion of absolute equivalence to the source text. Toury (2012, 23) reconceptualized translations as "facts of target cultures". As a socially-situated and culturally-bound activity, translation is shaped by norms, or socially enforced instructions for behavior that are grounded in target-culture values; deviation from cultural norms typically elicits negative social consequences for translators (Toury 2012, 63). Translation norms thus provide potential explanations for observable aspects of translation products and processes (ibid., 65). The primary manner in which norms differ between target cultures is in their amenability to interference, which is presented as an intrinsic byproduct of translation (ibid., 310-313).

In Toury's (2012, 310) definition, interference refers to "phenomena pertaining to the make-up of the source text… [being] transferred to the target text." Interference may be further divided into the subcategories of negative and positive transfer, where the former represents "deviations from normal, codified practices of the target system" and the latter represents the adjustment of frequencies of already-existing target-system features toward their corresponding frequencies in the source system (ibid., 311). He reasons that the diversity of norms governing translation practices across cultures will lead to radically different cultural tolerances of translation interference, and therefore the "socio-cultural factors" – as opposed to the purely cognitive or linguistic factors – of translation should be considered among the most important variables in the formulation of a descriptive hypothesis of interference in translation (ibid., 311).

Near the conclusion of his volume, Toury (2012, 314) presents his "law of interference" as the culmination of his theory, asserting cultures' propensity toward interference in translation to be directly determined by "power relations" between SLs and TLs:

> tolerance of interference – and hence the endurance of its manifestations – tends to increase when translation is carried out from a 'major' or highly prestigious language/culture, especially if the target language/culture is 'minor', or 'weak' in some other sense. (Toury 2012, 314)

Toury's use of "prestige" in this instance exemplifies the term's casual usage in translation studies: it does not appear to be aligned with the concept of language prestige as carefully constructed in sociolinguistics. As used in his law of interference, "prestige" coincides with the (perceived) strength of languages and cultures taken as singular, cohesive entities. Toury's use of the term must also be placed in the context of his foundational assertion of translations as facts of target cultures, which ostensibly formulate fairly cohesive sets of translation norms. The law of interference thus entails the interplay between cohesive sets of target-culture norms and the hierarchical relationships between languages and cultures as singular entities. For these reasons, Toury's use of "prestige" in fact aligns more closely with the established concept of language status in sociolinguistics.

What made Toury's law of interference a fitting foundation for empirical translation research was that it expressed, for the first time, an intuitive and straightforward relationship between the relative "weakness" of target languages/cultures and the level of interference that their translations were expected to exhibit. This proposition served as the most logical culmination of the field's dual embrace of 1) scientific inquiry (i.e., the pursuit of translation theories holding real predictive power), and 2) the ramifications of (systematically discernable) power imbalances between languages and cultures. As such, Toury's law of interference is assumed as the central hypothesis of this thesis.

In the same vein as Toury's *Descriptive Translation Studies*, Hermans' 1999 volume *Translation in Systems* builds on Even-Zohar's polysystem theory, supporting its central tenets of language and culture as sites of power struggles while explicitly incorporating principles from Bourdieu's work. He describes translation as being "deployed in the context of existing social structures" and thus subject to configurations of power in various forms, i.e., Bourdieu's various capitals (Hermans 1999, 80). Translations between English and Irish, for example, are produced within the context of the "massively unequal power relations between both languages" (ibid., 40). These underlying power relations are perpetuated by norms governing – among other aspects of the translation process – the linguistic composition of translations, in Hermans' framing as in other central texts in the descriptive translation studies tradition.

The view of translation as a norm-governed activity is essential to descriptive translation studies and its enduring legacy (Assis Rosa 2023, 194). As Hermans (1999, 83) points out, the role of norms in influencing translator behavior coincides with Bourdieu's habitus. In Bourdieu's work, habitus reflects "a set of dispositions which incline agents to act and react in certain ways" (Thompson 1991, 12). Neither translation norms nor Bourdieu's habitus are proposed as strictly deterministic rule sets governing agents' actions; rather, they are broad, socially-enforced pressures (implicit or explicit) encouraging certain behaviors. While translation norms are inextricably bound to the "hierarchical power structures" of languages and cultures, translators retain their agency to embrace or contradict these pre-existing norms (Hermans 1999, 82). Recalling the distinction outlined in the previous section, the relation or contrast between language status and language prestige resembles a necessary choice in terms of translators' adherence to norms: the indisputable social dominance and functional dynamism of high-status languages may be met with favorable or unfavorable attitudes (prestige). Language status is thus grounded in relatively observable social hierarchies, as opposed to the ideological inclinations of individuals and groups. It therefore more naturally aligns with the allegedly transferable nature of translation norms among various agents in a given target culture, viewed by descriptive theorists as a unified entity with a coherent and shared set of norms.

As summarized by Hermans (1999, 159), descriptive translation studies reflected a "reorientation which brought first culture and then politics and power" within the discipline's orbit. The early cultural turn's original preoccupation with the role of power in translation gained enough of its own momentum to inspire a distinct "power turn" in which translation was posited as a driver of cultural and social change (Tymoczko 2014, 44). This development created further opportunities for translation scholars to take their cues from Bourdieu, whose imprint on descriptive translation studies was already undeniable. The field soon began positioning translation as an explicit object of sociological inquiry, drawing mainly on the work of Bourdieu (Heilbron and Sapiro 2007; Wolf 2007; Assis Rosa 2010, 95; Hermans 2019, 146). His sociological approach allowed for the positioning of translators as "social and cultural agents and as active participants in both the production and reproduction of social and discursive practices" (Inghilleri 2023, 241).

Despite the usefulness of his ideas with respect to the relation between translation and power, Bourdieu scarcely discussed translation and translators, although his protégée, Pascale Casanova, later applied and reformulated his model in the context of the international literature market (Córdoba Serrano 2010, 252). Casanova (2002) sought to model the flows of imports and exports of translated literature as indicators of power relations between literary systems. As explored in depth in the following chapter, her method proves unworkable, and it is more theoretically and practically justified to conceptualize language status as a language's collective prominence across *all* domains, not only in literature. Despite descriptive translation studies' original fixation on literary translation (Assis Rosa 2010), Toury (2012, 205) intends his target-oriented approach as extending to other forms of translation beyond the literary realm.

Still, in comparison with other forms of translation, literary translation provides a favorable domain in which to examine Toury's proposed law of interference, as many scholars have asserted that it strongly reflects power differentials between languages and cultures. As discussed later in this section, Venuti (1995) based his highly influential theory on literary translation, asserting the dominance of Anglophone literary and linguistic norms on the manner in which non-English works are translated into English. While Cronin (1998, 155) advocates for an expanded perspective of the multitude of dimensions in translation, he suggests literary translation to be the domain most visibly dictated by the "more powerful Other" of dominant languages. Likewise, in his highly impactful *Culture and Imperialism*, Said (1994, xii) contends that literary fiction was "immensely important in the formation of imperial attitudes, references, and experiences" for 19th- and 20th-century European powers. Said's view points toward the diversity of approaches that, alongside the newly formed descriptive branch, characterized translation studies' cultural turn, which also incorporated ideas from feminist and postcolonial studies (Snell-Hornby 2006, 164). In fact, despite descriptive translation studies' joint emphasis on power relations and scientific methods, many other works belonging to the cultural turn and its legacy eschewed structuralist approaches and embraced poststructuralist approaches to the relation between translation and power (Gentzler and Tymoczko 2002, xiv).

Writing from a postcolonial perspective, Niranjana (1992, 60) argues that the descriptive branch's empirical inclinations require the "repression of the asymmetrical relations of power" between languages, regardless of scholars' early intentions. This critique has been recurrent in postcolonial perspectives on translation, which broadly centered the "imbrication of translation in processes of subjugation, exploitation, inequality and resistance" in response to the "descriptive blindness to questions of politics and power differentials" (Hermans 2019, 146). Nonetheless, postcolonial theories of translation arguably have many commonalities with the descriptive branch, not least of which is the emphasis on language hierarchies (Merrill 2019, 429). Tymoczko (1999/2014) demonstrates this compatibility by examining English translation strategies of early Irish literature through a postcolonial lens, drawing from Bourdieu as well as Even-Zohar and Toury. Whether or not they explicitly aligned with the descriptive theorists, postcolonial and feminist translation scholars also corroborated the essential link between the linguistic features of translated language and dichotomous SL/TL power asymmetries, where languages possessed a form of power that was separable – if ultimately derived – from their associated social groups.

Asad (1986, 160) describes an "inequality in the power of languages" and asserts the implications of this inequality for translation in anthropological settings. He asserts that more powerful languages "reshape" less powerful languages via translation, a process driven by the political and economic asymmetries between countries as well as the increased demand for knowledge disseminated in dominant languages (ibid., 157-158). The example Asad (ibid., 158) cites is that of 19th-century Arabic "undergo[ing] a transformation (lexical, grammatical, semantic)" that rendered it more similar to those structures of English and French, from which many texts were being translated. Said differently, Asad supposes translators' proclivity toward interference to stem from language power imbalances, noting that they may be so widespread as to fundamentally change the structure of Arabic.

Rafael (1988) explores the power dynamics governing the translational activity between Spanish colonists and the subjugated Filipino population. He explains: "For Spanish missionaries, translation thus presupposed the existence of a hierarchy of languages" (ibid., 27). Latin was situated atop this hierarchy and believed to be the most divine language from the perspective of the Catholic Church, while Spanish was viewed as a necessary step between Tagalog, the local vernacular, and Latin (ibid., 28). Translating into Tagalog, Spanish missionaries opted to import numerous Latin and Spanish terms, believing the target language to be inadequate and signaling their "belief in the intrinsic superiority of some languages… over others" (ibid., 29). Rafael's example demonstrates a power imbalance in which translations into a comparatively lower-status language are rendered using features directly imported from the higher-status source languages, aligning with Toury's law of interference.

Also during this period, feminist translation scholars began linking translation thought to gender studies, a discipline formed on the basis of “asymmetrical power relationships, as caused by patriarchal hegemony” (Snell-Hornby 2006, 100). Spivak (1993) implores translators to reflect on language power dynamics in both selecting texts to translate as well as formulating actual translation strategies. She refers to the inclination of “metropolitan feminist[s]” to approach minor-language works with the intent to produce “a too quickly shared feminist notion of accessibility” that unilaterally shapes translations according to the experiences of readerships whose knowledge is restricted to more hegemonic languages and cultures (ibid., 322).

This notion of target-side accessibility prioritized by those translating into high-status languages resembles the famous dichotomy at the center of Venuti’s work. The global political economy at large underwent major changes in the 1990s that would heavily influence perspectives in translation studies, as with many other academic disciplines. This period of globalization reconceptualized the world map as a unified market, bringing otherwise distant languages and cultures into contact with one another like never before. Amidst the global dominance of Anglophone economic powers, the nearly ubiquitous adoption of English by “people and institutions in various parts of the globe for economic or political survival (or profit)” has brought “deep-seated consequences for translation” (Snell-Hornby 2006, 140).

Venuti (1995) presented his culturally-oriented theory of translation against the backdrop of the unfolding global domination of English. Invoking Schleiermacher’s prescriptive orientation, he reformulated the classic dichotomy of translation strategies as foreignization (SL-oriented translation, akin to interference) and domestication (TL-oriented translation, akin to normalization), arguing passionately for the widespread adoption of foreignizing strategies as a conscious resistance to the domesticating tendencies of Anglophone literary translation. In his seminal work *The Translator’s Invisibility* (1995, 17), Venuti asserts the widespread practice of Anglocentric translation strategies, which he interprets as a consequence of the “global domination of Anglo-American culture”. In his view, the global hegemony of English is closely tied to the observable features of texts translated into English, which tend to neglect the linguistic or cultural peculiarities of their respective source texts. Primary aspects of this “domesticating” translation strategy are the avoidance of “foreign words” and SL syntactic structures that differ from typical TL syntax (Venuti 1995, 5).

However, it was not always evident how these translation strategies would take shape, practically speaking. As depicted by Snell-Hornby (2006, 146), Venuti’s ideal foreignization strategy would prioritize, for instance, “archaic terms or idiosyncratic word-order” as a means of amplifying a translation’s markedly foreign elements, but it is not clear whether these decisions would amount to the intended effects on readers’ perceptions. Venuti’s work also drew criticism for offering a curated historical account of translation theory and practice and, in doing so, mistakenly universalizing the dynamics supposedly reflected by the English-speaking literary translation market (ibid., 147).

The focus on translation into English as a general proxy for translation writ large was not entirely far-fetched, however. Depicting a “core-periphery structure” reminiscent of polysystem theory, Heilbron (1999, 435) observes the tendency for central languages to mediate translation between two peripheral languages, with English occupying a “hyper-central” position in the present age. This process has come to be known as pivot or indirect translation (i.e., the process and products of translating translations), and has received increasing attention in recent years, particularly with respect to the larger implications of dominant languages serving as intermediaries (see Whyatt and Pavlović 2021). Drawing from polysystem theory, Pięta (2016) portrays the history of direct and indirect translations between Polish and Portuguese – two “(semi-) peripheral languages” – as indicative of the underlying power relations between core and periphery systems. Hadley (2017, 195) calls for further research on “the role of language power and prestige” in influencing translators’ strategies throughout indirect translation chains. The establishment of indirect translation as a subfield in translation studies constitutes a natural step in descriptive translation studies’ evolution, with its overt inspiration from polysystem theory and aspirations toward a comprehensive framework for assessing the combined or compounded effects of SL-TL power relations.

As this section has demonstrated, translation scholars have regularly theorized and examined the impacts of power relations between source and target languages on translated texts in qualitative terms. Most of these accounts lacked empirical foundations, as exemplified in the questionable descriptive basis on which Venuti’s prescriptive view depended. Postcolonial perspectives on literary and cultural translation also remained separate from the practices of more empirically-minded translation researchers (Venuti 2012, 190).

Toury's laws were perhaps initially envisioned as the joint introduction of power dynamics as well as the scientific method to translation studies. The law of interference, which Chesterman (2016, 71) calls "[p]erhaps the most pervasive of all translation laws," is the culmination of this vision, explicitly naming interlingual and intercultural power asymmetries as key predictors of translation strategies – namely, of interference. Despite their scientific framing, Toury's laws are put forth on theoretical rather than empirical grounds. They are therefore more aptly characterized as scientific hypotheses, given that they are intended to hold predictive power but have yet to be confirmed or rejected via empirical testing (ibid., 71). As the scientific aspirations for translation studies took shape over the subsequent decades, empirical research scarcely incorporated SL/TL power relations as an explanatory variable in the manner suggested by Toury, with the exception of several small-scale studies outlined in the following section.

Concurrent to the emergence of these new approaches near the close of the 20th century, translation studies was undergoing another major overhaul. The original empirical aspirations of descriptive translation scholars required a complementary methodological innovation – one that eschewed the discipline's traditional reliance on close readings and isolated textual examples in favor of more systematic and quantitatively holistic methods. The discipline's methodological breakthrough came in the form of its adoption of methods in corpus linguistics, made widely available during the 1990s thanks to substantial increases in computational power.

2.5. Digital corpora and their descriptive potential

Newly stored in digital format, corpora provided the means for researchers to perform dynamic queries of linguistic features and patterns across swaths of texts with great efficiency. With translation texts being excluded from early corpus-based research due to their ambiguous position in traditional linguistics research, Baker (1993) was the first to propose corpus methodology as the primary means of conducting empirical translation research. As she boldly predicted, corpus methodology would allow researchers to uncover, “on a larger scale than was ever possible before, the principles that govern translational behaviour and the constraints under which it operates” (Baker 1993, 235). Citing the influence of Even-Zohar’s polysystem theory and Toury’s concept of norms, she presented corpus-based translation research as descriptive translation studies’ missing methodological foundation (ibid., 237-241).

For Baker (1993, 243), corpus-based translation studies’ foremost aim would be to identify universal features of translation, or those typical features that distinguish translations from non-translations – aside from those attributable to “interference from specific linguistic systems”. Such universal features were assumed to be consistent across cultures and languages, and could be revealed via extensive corpus research in various contexts. One universal feature of translation posited by Baker (1996, 183) was normalization – a phenomenon describing translations’ “tendency to exaggerate features of the target language and to conform to its typical patterns.” The strength of translations’ normalizing tendencies is suggested to be sensitive to language power relations:

> This tendency [of normalization] is quite possibly influenced by the status of the source text and language, so that the higher the status of the source text and language, the less the tendency to normalise. (Baker 1996, 183)

This assertion reflects the inverse yet complementary correlation to Toury’s hypothesized law of interference: translations from relatively higher-status SLs tend to exhibit less normalization than translations from lower-status SLs. Logically speaking, Baker’s assertion may also be reformulated as follows: translations into relatively higher-status TLs tend to exhibit more normalization than translations into lower-status TLs. Formulated in this manner, Baker’s supposition of the effects of language status asymmetries on translated language mirrors Venuti’s own claim about the strongly domesticating tendencies of translators in the Anglophone literary market. Translators working into English are expected to render texts using a strongly TL-oriented strategy (i.e., normalization or domestication) given the language’s high status.

Like Toury's law of interference, Baker's assertion of normalization as a universal feature of translation constituted a preliminary hypothesis to test using the empirical foundation that corpus-based methodologies offered translation researchers.

Nevertheless, the empirical merits of corpus methods relied on subjective assumptions that were, practically speaking, irresolvable. The most glaring challenge of corpus design is that of representativeness – "the extent to which a sample includes the full range of variability in a population" (Biber 1993, 243). A corpus is a collection of texts deliberately selected according to some organizing principle(s) and intended to serve as a representative sample of some larger population which is the ultimate object of inquiry. In practice, the text selection process is heavily constrained by text availability, and the aim of representing a larger population is restricted to researchers' limitations in knowing what the boundaries and variability of a given population actually are. These difficulties were already anticipated by Toury (2012, 71), who recognized actually-existing corpora to be "more or less arbitrary selection(s)" that are "not representative of anything but [themselves]." This fundamental limitation is one of the foremost reasons that the empirical basis of corpus-based translation studies has been called into question.

The issue of representativeness in corpus design may also invoke a subtle distinction in Bourdieu's work, which theorizes not only the power dynamics between languages but also *within* them. For Bourdieu (1991, 59-60), writers of prominent literature serve as authorities of a language's use, i.e., consequential individual holders of linguistic capital. Their practices and preferences are much more consequential to a language's ongoing development than, for instance, L2 speakers with lower linguistic competence and social or cultural influence. Applying Bourdieu's framing, it may be intuited that a language's most acclaimed authors reflect its most promising sources of linguistic capital, thereby offering a reasonably representative sample of a language's literature. (This point will be expanded on in Chapter 4, which presents the thesis' methodology.) Corpus linguists have adopted similar lines of reasoning in justifying text selections to be sufficiently representative, even as the ideal of absolute representativeness remained perennially out of reach. The inherent challenges of sampling in corpus linguistics have not dissuaded translation researchers from pressing forward with this empirical methodology.

The dichotomy of interference and normalization has provided a common descriptive framing for corpus-based translation research, as demonstrated in Lefer and Vogeleer's (2013) collected volume of corpus-based studies measuring levels of translation interference and normalization in a variety of forms. Although interference and normalization are conceptually linked, this thesis first and foremost focuses on interference, which is generally perceived to be the more apparent feature of translated language (ibid., 16). In one of the most well-known studies, Teich (2003) conducts a thorough investigation of "shining through" (i.e., interference) and normalization in translations between English and German based on comparisons of diverse lexico-grammatical features' distributions across translations, their source texts, and comparable TL corpora. This comparative framework in fact serves as a widely-accepted means of characterizing interference and normalization in translation, given the intrinsic comparability of linguistic features' relative frequencies across various texts and corpora (Kotze 2021, 119). Studies of this nature are generally restricted to a single language pair (see De Sutter and Van de Velde 2008; Bernardini and Ferraresi 2011; Hansen-Schirra 2011; Delaere and De Sutter 2017). Just as with CLI in contact linguistics more broadly, linguistic features examined as potential indicators of interference or normalization in translation tend to be broadly sorted into overarching categories of lexical borrowing and structural (or grammatical) borrowing (Kotze 2021, 118). Paratextual features of translations have similarly been examined within the SL-/TL-oriented dichotomy of translation strategies, yet these elements are typically framed within Venuti's framework of translator visibility (Batchelor 2018, 32-33). While the interference/normalization dichotomy remains a persistent theme in corpus-based translation research, the relationship between translation interference/normalization and SL-TL power dynamics has received far less attention.

Corpus-based research on the effects of language power relations on translated language has been scattered, focusing exclusively on isolated linguistic features and individual language pairs – typically involving English and another European language. Researchers widely favor the term "prestige" over "status" when referencing the (relative) power of languages, therefore matching Toury's terminology, but they do not define or describe the term in much detail. In one of the most illustrative recent examples, Evert and Neumann (2017) find strong evidence of a "prestige effect" in a bidirectional corpus-based study, whereby translations from German into the comparatively higher-status English exhibit more lexico-grammatical normalization than translations in the opposite direction. As they point out, however, similar studies have yielded mixed conclusions about the so-called prestige effect. Mauranen (2004) tests for this same alleged phenomenon and finds that Finnish translations of Russian texts do not exhibit more lexical normalization than Finnish translations of English texts, an unexpected result given the (perceived) higher prestige of English compared to Russian. Becher et al. (2009) hypothesize that the prestige of English leads to more normalization (which they call "convergence") in English-to-German translations than in translations in the opposite direction. Their study yields mixed results among the various normalization markers (ibid., 147). Van Poucke (2011) investigates Russian loanwords in a corpus of 20 Dutch translations of Russian novels, finding that the average number of loanwords decreased from the 1980s to the 1990s. He attributes this change to the decline in the prestige of Russian over this same period, during which the Soviet Union's disbandment and subsequent integration into the global capitalist order rendered the language less consequential on the world stage (ibid., 118). In order to detect evidence of interference and normalization, Van Oost et al. (2016) compare frequencies of prepositional phrase placement in translations and original texts in Dutch and the more prestigious German; they find a clear prestige effect, therefore confirming Toury's hypothesis.

Apart from these limited studies, the effects of asymmetries of power ("prestige") between languages remain largely unexamined in corpus-based translation research (De Sutter and Lefer 2020, 4). The fundamental relationship between SL/TL power relations and translations' linguistic features, as initially postulated by Toury and Baker, has yet to be investigated systematically across a range of language pairs and linguistic features. Undoubtedly, one of the factors contributing to this research gap is the terminological and conceptual stability of language prestige and/or language status in translation scholarship. Interpretations of the results of the studies described previously are obviously contingent on the manner in which a language's relative power ("prestige") is assessed, with the researchers relying on intuitive judgments to make such determinations, most often concentrating on English.

The dominance of English in the world language system is commonly portrayed as a gravitational force that inspires translators in other target cultures to adopt English-centric translation norms, thereby inducing language change more broadly. House (2011) conducts a critical investigation of this supposed phenomenon, combining qualitative and quantitative (i.e., corpus-based) methods. Constructing a parallel corpus composed of non-fiction texts in several genres, her study primarily investigates diachronic changes in the frequencies of selected linguistic features for the German-English language pair in both original and translated texts, finding no clear evidence of the invasiveness of English-language norms (ibid., 204). Nonetheless, the alleged "omnipresence of Anglo-American linguistic-cultural norms" still perhaps constitutes a default example in discussions of the effects of language power relations in translation and language change (ibid., 189). While there may exist a broad consensus regarding the dominance of English in the modern age, the relative positionings of the rest of the world's languages are undoubtedly less clear, especially when examined in their various possible historical contexts.

A corpus-based research agenda centering on the "principles that govern translational behaviour and the constraints under which it operates" (Baker 1993, 235) cannot draw meaningful conclusions about language power relations without also investigating these dynamics across diverse, non-English language pairs. Moreover, the comparability and replicability of results across these diverse research contexts requires careful consideration (Chesterman 2004, 46; De Sutter et al. 2012, 138). It is therefore imperative to develop a systematic assessment model for codifying language power relations, such that they may be operationalized consistently across studies involving diverse language pairs. For reasons outlined earlier, this thesis assesses language power in the form of language status; a novel method for assessing language status as a quantifiable variable for corpus-based translation research is developed and applied in Chapter 3. The language status assessment model presented in this thesis may then be used to articulate language power asymmetries within translation corpora, thereby representing language power as an explanatory variable in corpus-based research.

This missing element has rendered descriptive and corpus-based translation studies' foundational hypotheses untestable except in very limited contexts. Nonetheless, corpus-based translation research has in many ways proliferated, and the practical applications of digitally-stored translation corpora have led to a renaissance in translation technologies, where language power has also been a recurrent, if frequently mischaracterized, theme.

2.6. Automatic translation and language power relations

The advent of statistical MT (SMT) in the 1990s reflected a fundamental shift in the way that automatic translation had previously been conceptualized and practiced. Unlike its predecessor, rule-based MT, SMT represented a data-driven – or, more specifically, a *corpus-driven* – approach to MT, where instead of following explicit translation rules as articulated and manually encoded by linguists, systems were developed on the basis of training data, in the form of (ideally) parallel corpora assembled from human-produced translations. SMT constituted a major improvement upon the rule-based approach, no doubt thanks to its reliance on samples of pre-existing translated segments instead of rigid interlingual operations. The MT landscape was transformed yet again with the introduction of neural MT (NMT) in the mid-2010s. Like SMT, NMT is a data-driven architecture, though it requires vastly larger quantities of training data for optimal or even adequate performance, and the availability of training data is highly uneven among the world's languages.

Stark differences in performance between high- and low-resource languages for data-driven MT and other natural language processing (NLP) tasks have thus been a major concern among researchers since the inception of these technologies (see Koehn and Knowles 2017). Preliminary research on the multilingual and translation capabilities of large language models (LLMs) also demonstrates the correlation between the amounts of language-specific training data and models' performance in these tasks (Robinson et al. 2023). As this section will demonstrate, inequalities between languages in terms of their availability of training data and the performance of their NLP tools have served as a proxy for language power dynamics since these technologies gained widespread prominence.

Crucially, this perspective oversimplifies more comprehensive sociolinguistic accounts of language power and, in doing so, also undermines the potential to investigate the potential impacts on MT output. As this section will demonstrate, the analytical frames of language status and linguistic norms have been largely absent from empirical research on the properties of automatic translation output, despite their great explanatory and descriptive potentials. While the evidently hierarchical relationships between high- and low-resource languages are undeniably crucial to language equality within NLP research, this framing does not fully encapsulate language power relations writ large, as the availability of digital resources does not correlate straightforwardly with the language status criteria often proposed by sociolinguists – e.g., domain capture and number of speakers. While there are several noteworthy attempts by NLP researchers to elucidate these technological inequalities by reference to external social factors, there has been little engagement with sociolinguistic literature.

Joshi et al. (2021) survey the landscape of digital language technology and introduce an empirically-grounded, six-category typology for defining the level of digital support (i.e., available technology and data) for each of the world's languages, spanning those which are wholly excluded from the digital sphere to those with ample resources and support to maximally benefit from cutting-edge NLP developments. As noted by the researchers, despite Dutch and Somali having comparable speaker population sizes, the former has vastly better technological support than the latter (ibid.). Gaspari et al. (2022) likewise introduce a quantitative classification system for languages' levels of technological support as part of the European Language Equality project. Behind English, German, Spanish, and French, Finnish registers as having a stronger digital infrastructure than Italian (ibid., 5-6). As with Dutch and Somali, this case demonstrates the manner in which languages' level of digital support often diverges from the more fundamental notions of language status as articulated in sociolinguistics: although there are many fewer Finnish speakers than Italian speakers, Finnish outranks Italian in terms of its digital language tools and resources. Gaspari et al. (ibid., 7) further observe that unofficial EU languages like Catalan, Galician, and Welsh are disproportionately technologically-supported, more so than some official EU languages. Evidently, familiar classifications of high- and low-resource languages often roughly align with intuitive notions of high- and low-status languages, but they do not consistently represent the power conferred to languages in broader social contexts.

Other studies have posited a link between languages' digital support and the economic vitality of associated national economies. Faisal et al. (2022) evaluate language communities' (geographical) representativeness among NLP tools by connecting language data sets to their associated countries, illustrating a crucial lack of geographical diversity in the field. It is also made apparent in their work that the geographical distribution of data sets corresponds to countries with high economic outputs as measured by gross domestic product (GDP) (ibid., 6-7). Blasi et al. (2021, 7) also find that, more so than speaker population sizes or number of relevant academic publications, "it is the economic prowess of the users of a language (rather than the sheer demographic demand)" which determines languages' digital vitality. They formulate a singular indicator of languages' economic power by first aggregating the national GDP for each country in which the language community has a presence, then proportionally allocating the community's share of the GDP according to the percentage of speakers within that country (ibid., 13). While workable and perhaps somewhat intuitive, this method embodies the long-standing criticism of applying GDP as a measure of economic strength, as it assumes the even distribution of benefits of economic production throughout (national) populations. In the example provided, the authors attribute 1.3% of Mexico's national GDP to Nahuatl speakers in accordance with the size of their speaker population, when in reality the economic strength of the disadvantaged indigenous community may be much lower. As such, this superficial, economically-oriented assessment falls short of capturing linguistic capital in a meaningful capacity.

Discussing the theme of language power in relation to translator and interpreter training, Whyatt and Pavlović (2021, 144) distinguish between low-resource languages and "languages of low diffusion" (LLDs), despite the "asymmetry in power relations" that defines both and the tendency for the two categories to overlap. In addition to being endangered or demographically vulnerable, LLDs are "usually but not necessarily small in the number of native speakers" and "rarely learned by non-native speakers" (ibid., 102). For instance, the disproportionate level of digital support for Czech contrasts with the language's standing as an LLD, whereas Hausa has scarce technological support, despite being a widely spoken lingua franca for tens of millions in West and Central Africa (ibid., 144). Like language status, the LLD category does not necessarily coincide with that of low-resource languages, yet its focus on numbers of native and non-native speakers crucially excludes the role of social domains in conferring power to languages.

While several NLP works have rightly observed the potential discrepancies between languages' level of digital support and their perceived power in society at large, these studies offer highly limited accounts of language inequalities, as they are primarily concerned with the performance of NLP systems and the availability of data for different languages. It appears that sociolinguistic concepts of language status or prestige have not yet been linked to the realm of digital language technologies. Also seemingly absent is the discussion of translation norms and their potentially compounded effects in large-scale MT systems. With MT systems trained on the corpora comprising human-produced sample translations, the underlying translation norms embedded in training data are necessarily reproduced – and perhaps even exaggerated – in systems' output, as the de facto translation strategies for future unseen inputs are derived from these aligned source-target training segments and encoded in massively complex mathematical representations (see Blodgett et al. 2020; Schneider 2022; Navigli et al. 2023). Toury's law of interference presents the relative power differentials between languages as the preeminent factor in determining translation norms regarding the acceptability of SL influence in any given target culture. As such, it is possible that the process of creating MT training datasets and benchmarks for quality evaluation for low-resource languages may be perpetuating radically different translation norms compared to more high-resource language pairs, given that low-resource languages are commonly also low-status languages.

A particularly noteworthy initiative exemplifying this potential risk is Facebook's creation of the FLORES-101 dataset, which constitutes an evaluation benchmark for MT performance for a wide range of low-resource languages (Goyal et al. 2022). Once again, it is noted that "[many] languages are spoken by millions, despite being considered low-resource in the research community" (ibid., 528). What makes this example particularly crucial is that the dataset contains translations from the same set of sentences in English – undoubtedly the most globally dominant language – into various (severely) low-resource languages such as Cebuano and Māori. This design is intended to facilitate MT quality evaluation for all possible language pairs among the selected languages, as any aligned segment for an English/non-English language pair may ostensibly be easily converted into an aligned segment for a non-English/non-English language pair, given their shared source sentences. However, this supposed convertibility is contingent on a simplistic view of translation equivalence that disregards the prospect of divergent translation norms among diverse target cultures.

The project leaders outsourced the translations and the subsequent quality evaluation tasks to various language service providers, where the initial translators' work is reviewed by evaluators using a standardized quality scoresheet provided by the Facebook team (Goyal et al. 2022, 528). The lone indication of the team's overt attempt to shape translation strategies beyond the narrow equivalence-based framing is their instruction to "translate abbreviations and idiomatic expressions to their best knowledge of *how these terms and phrases usually appear in the target language* [emphasis added], finding equivalents rather than literal word-for-word translations" (ibid., 526). Beyond this glancing reference, the researchers' analysis of the resulting evaluation data reveals their commitment to the equivalency-based approach. They note that "mistranslations" were the most commonly observed error across all languages, describing the phenomenon as "a broad category that generally notes that the source text was not translated *faithfully* [emphasis added]" (ibid., 528).

The workflow employed in creating this evaluation benchmark mirrors translation studies' early conceptualization of translation as a purely linguistic process, and the FLORES-101 initiative is a perfect distillate of this embedded and ongoing trend in MT research and development. MT research and development has overwhelmingly prioritized this one-dimensional notion of "quality" above all else (Way 2018). Seemingly unacknowledged in MT research is the notion that massively multilingual datasets, particularly those including low-resource languages such as FLORES-101, are composed of translations into a plethora of TLs that are situated in highly diverse cultures with (likely) radically different notions of what constitutes translation proper. It is here where Toury's work in deconstructing the concept of equivalence and asserting the existence of diverse target-culture norms holds great potential for advancing the current state of MT research. Comparisons between MT output for different language pairs – particularly those involving low-resource languages, which more often than not are also low-status – might not only take place on the basis of perceived quality but also on the basis of purely descriptive empirical accounts of their linguistic features.

As demonstrated previously, MT research has thus far constituted a continual pursuit of improving system performance in relation to static benchmarks underscored by reference translations, conceptualizing translation in prescriptive terms. It has come at the cost of approaching automatic translation on purely descriptive grounds, particularly with the seemingly opaque inner workings of the modern era's enormously complex NMT and LLM systems. There are a few noteworthy exceptions to this trend in the literature, however. The most prominent example is Toral's (2019) highly influential study detecting consistent and substantial differences between the linguistic features of human-produced, machine-translated, and MT post-edited translations. Examining three datasets and five different translation directions, he demonstrates post-edited MT output to be simpler (in terms of lexical variety and lexical density) and more normalized (in terms of sentence length), and reflective of a higher degree of SL interference (discussed in Chapter 6) than human translation (ibid.). Volkart and Bouillon (2023) offer conflicting evidence, showing that the effects of post-editing on linguistic features are heavily dependent on the language pair and MT system under examination. Other research on the linguistic features common to post-edited MT output also produces mixed results (see Daems et al. 2017; Castilho and Resende 2022). These studies overtly affirm the inclinations of Toury and Baker to reveal linguistic features common to translated language (translation universals), regardless of language pair, extending this research strand to MT output.

The field's recent interest in the various forms of bias exhibited in generative AI outputs is conceptually related to the investigation of translation and language models' strictly descriptive features as untethered from rote performance metrics. There is a growing body of research in this area (see Navigli et al. 2023). Blodgett et al. (2020) contend that methodological rigor and interdisciplinary perspectives have been crucially absent from research on bias in NLP. The researchers argue that future work in this area should seek to understand "how existing social hierarchies and language ideologies drive the development and deployment of NLP systems, and how these systems therefore reproduce these hierarchies and ideologies" (ibid., 5459). Bias, as thus far examined in NLP literature, has been framed as a linguistic representation of the hierarchical relationships between various social groups, as drawn along the lines of race, class, and gender, for example. Bias between hierarchically-ordered languages, however, has been scarcely examined. Choudhury and Deshpande (2021) posit languages as discernable entities for whom the fairness of MT models and LLMs may be assessed, yet their approach also prioritizes one-dimensional quality metrics, conceptualizing fairness in terms of the relative distribution of (losses in) quality among various language pairs owing to their consolidation into a single multilingual translation model. While research on bias has incorporated the notion of norms as linguistic conventions "implicitly assumed to be standard, ordinary, correct, or appropriate" (Blodgett et al. 2020, 5459), the implications of the existence of translation norms as posited by Toury remains unexplored.

The manner in which underlying translation norms reflected in MT training data are potentially perpetuated by systems is undoubtedly rendered even less predictable by the opacity of modern, state-of-the-art NMT and LLMs and the sophisticated computational mechanisms that facilitate low-resource MT. The primary mechanism enabling low-resource MT in modern (multilingual) NMT systems is cross-lingual transfer learning, in which "a high-resource *transfer* language is used to improve the accuracy of a low-resource *task* language" through sharing previously trained weights (Lin et al. 2019, 1). This setup, with the aid of machine learning strategies, allows for monolingual training data and parallel training data for linguistically similar languages to enhance data-driven MT systems' performance for low-resource languages (see Haddow et al. 2022). The switch to "massively multilingual" NMT models, in which transfer learning unified all languages and language pairs in a single model, has enabled developers to drastically expand the coverage of languages (see Bapna et al. 2022; NLLB Team 2022). The proliferation of transfer learning techniques has substantially lessened the necessity of parallel data in MT as well as other multilingual NLP tasks, as monolingual data often suffice (Joshi et al. 2021, 2).

The multilingual approach proved highly effective in advancing not only NMT but also in endowing LLMs with multilingual capabilities, despite the latter being trained on overwhelmingly monolingual English data (Kew et al. 2023). It was not until the more recent LLM iterations, which are significantly larger than their predecessors, that models began showing potential as a tool for translation and translation-related tasks such as quality assessment (see Kocmi and Federmann 2023). The highly effective transfer learning technique is now firmly established as best practice in NMT and (multilingual) LLM training, which now reflects the amalgamation of copious parallel and monolingual data for a staggering number and variety of languages. Because state-of-the-art automatic translation technologies aggregate training data from diverse sources then process them in a highly complicated manner, it is unclear how the linguistic norms underpinning training data might be propagated or distorted by subsequent processing.

This computational complexity has rendered the inner mechanics of these systems almost entirely opaque, heightening the need for "explainable AI" – research methods designed to explain the relationship between system inputs and outputs (Kenny 2022, 42-43). One attractive feature of the primitive, rule-based approach to MT, despite its generally dismal performance, was its absolute transparency: developers could easily create, observe, and modify as needed systems' operations for converting input into output. The massive performance benefits of NMT have come at the cost of this technical legibility. Previous attempts to achieve some degree of NMT explainability involve strategically and systematically manipulating MT input in order to track subsequent changes to output (Stahlberg et al. 2018; He et al. 2019). Small-scale approaches may prove useful in describing specific linguistic phenomena, but it is difficult to envision their results leading to scalable generalizations – a task that is perhaps better-suited for translation studies' already well-established empirical tradition. Asscher (2022) argues that the theoretical framework articulated by Toury in the early aughts of translation studies' descriptive branch presents the best option for characterizing the typical features of NMT output. In fact, Baker (2004, 184) considers transparency to be one of primary strengths of corpus-based translation research, as its methodology allows for other scholars "not only to check the validity of the basic claims being made but also to offer different interpretations of the same data."

Descriptive translation studies' central preoccupations perhaps offer a crucial change of perspective to the NLP field's instrumentalist view of language and language data. The fixation on unidimensional, performance-focused metrics in NMT research and development reflects early accounts of translation as a purely linguistic phenomenon, neglecting the "asymmetry born of unequal power relations" which leads to the diversity of views about what constitutes translation proper (Asscher 2023, 8). The NMT training process entails the continuous adaptation of the system's internal mathematical representation to recreate or approximate the presupposed source-target translation equivalencies reflected in its parallel training data; in this manner, NMT systems inherit and reproduce whatever translation norms are inscribed in their training data (Asscher 2022, 10-11). It is therefore not only the "uneven performance of NMT in different language pairs and directions" that "replicat[es] unequal geo-political and cultural power relations," but also the norms that are embedded in NMT systems and potentially perpetuate the effects of language power dynamics (Asscher 2023, 9).

If Toury's law of interference has not yet been adequately investigated in empirical translation research, it is even further removed from the MT research agenda. Given that human-produced training data provide the foundation of NMT, investigations of the potential relationship between language power relations and the linguistic patterns of NMT output would naturally follow from more systematic empirical research regarding this phenomenon in human translation.

2.7. Conclusions of literature review

This chapter has demonstrated the merits of Bourdieu's concept of linguistic capital – used synonymously with language power in this project – in articulating the ways in which the status and prestige of languages have been conceptualized in sociolinguistics and deployed in translation studies. It has defined language status and language prestige on this theoretical basis: prestige refers to the conscious attitudes of a language's advocates or detractors, whereas status reflects the broad contours of its social functions, as well as the demographics and geographic expansiveness of its user base. The chapter has presented the core elements of operational sociolinguistic accounts of each concept, arguing for the adoption of language status as the preferred concept for operationalizing language power dynamics in the complementary hypotheses of Toury and Baker. Key factors contributing to language status have been derived from a brief review of sociolinguistic literature, enabling the following chapter's development of a language status assessment model for corpus-based translation research.

It has also been shown that the SL-TL dichotomy has been fundamental to translation theory since its inception, and that the advent of descriptive translation studies explicitly proposed the effects of language power relations on translations within this dichotomous framing. Much as in Bourdieu's framing, the hypotheses put forth by Toury and Baker predicted that the norms of target cultures entail translation strategies that prioritize the conventions of whichever language (SL or TL) is higher in status (i.e., linguistic capital). As in much of the discipline's history, descriptive translation studies and its immediate corollaries focused mainly on literary translation, and it was in this domain that language power dynamics were assumed to be the most impactful. Some scattered attempts to detect the effects of SL-TL power relations in translation have been made using corpus methodology, though these efforts have been limited to single language pairs and isolated linguistic features. NLP and MT research has also inadequately accounted for language power relations, as efforts to characterize these dynamics have solely conceptualized power in terms of digital resources and superficial socioeconomic factors. As such, the absence of a systematic investigation into a quantifiable relationship between language power relations and SL influence in translation constitutes a glaring gap in the current research.

3. Developing a language status assessment model

3.1. Chapter introduction

Following translation studies' cultural turn and its subsequent sociologically-motivated inquiries, scholars became more concerned with uncovering the manner in which "social constraints and dynamics" govern the translation process and are reflected in translations' compositions, which "call[ed] first and foremost for the fostering and the refinement of the methodologies" necessary for drawing any such conclusions (Wolf 2007, 141). The preceding chapter illustrated this need in relation to the incorporation of language status as a systematized variable in empirical translation research.

This chapter explores several previous attempts to devise systematic assessment models for language status and its related concepts. It then combines the strengths of each approach with the insights gleaned from sociolinguistic literature in Chapter 2 to develop a novel language status assessment model tailored to corpus-based translation research. Finally, it selects a range of languages to be used in the current project, applying the model to assess the relative status positionings of each language.

3.2. Previous attempts to assess language status

This section highlights several previous attempts to systematize social characteristics of languages that are conceptually linked to their status: De Swaan's (2001) Q-value, the evaluation of the international exchange of literary translations put forth by Heilbron (1999) and Casanova (2002), and finally Fishman's (1991) GIDS framework and its expanded version by Lewis and Simon (2010). It outlines the general strengths and weaknesses of each approach in consideration of their workability in the context of the current project and future corpus-based translation research.

3.2.1. De Swaan’s Q-value

Abram de Swaan (2001, 32) proposes a model for assessing languages’ social value or communicative potential, beginning with the premise that language constitutes a hypercollective good – a good whose value increases the more that people use it. According to this view, the fundamental utility or value of a language lies in its communication potential – the extent of communication that it facilitates or enables between its users – and individuals make rational choices about which new languages to acquire based on this perceived value (ibid., 26). As a language gains new users, its communication potential simultaneously increases for all users (ibid., 27). This process gives rise to a competitive linguistic ecosystem in which widely used languages tend to gain new speakers and increase their value, while less popular languages may stagnate or lose speakers over time. Like Bourdieu and other sociolinguists, De Swaan (ibid., 7) acknowledges that languages expand and diminish according to an interdependent system in which contact between different language communities results in the capitulation of more peripheral languages to more central languages, the latter of which constitute the dominant languages of “conquest, conversion, and commerce” (ibid., 7). Encounters between languages of unequal communication potentials are pressurized by their competition within shared political economies, where languages provide different levels of access to social and economic opportunities (ibid., 18).

Recognizing that competitive tensions between languages are heightened by their proximity, De Swaan’s model emphasizes the geopolitical contexts in which languages are situated. As such, he does not purport to measure the value of a language in global or absolute terms, but instead asserts that a language’s communication potential may only be expressed *relatively* within a defined language constellation (De Swaan 2001, 34). These constellations align most naturally with broad geopolitical units, whether multilingual nation-states (India, Indonesia, South Africa), supranational organizations (the European Union), or geographical regions (Sub-Saharan Africa) (ibid., 21-22). For a specific language constellation, De Swaan (ibid., 34) calculates the communication potential of each language as a “Q-value”. A language’s Q-value is not calculated simply according to its number of speakers, but rather by multiplying two proportional values – prevalence and centrality. Prevalence denotes the proportion of speakers of a given language out of the total number of speakers in the constellation, while centrality refers to the proportion of *multilingual* speakers in the constellation who also speak that language, serving as an "indication of its connectedness to other languages" (De Swaan 2001, 33). Calculating Q-values in this manner, a language's communication potential is represented as a continuous variable, making it an attractive operationalization for measuring a potential correlation between SL status and SL influence in translation. However, De Swaan's approach has drawn substantial criticism.

According to Phillipson (2004, 74) De Swaan's model mischaracterizes language shifts as merely the aggregate of individual language preferences, ignoring for instance the effects of language policies and other political forces that deliberately promote hegemonic languages and suppress minority languages. The reliance on individual language learners' supposed economic rationalism to characterize competition among languages mistakenly divorces language from "issues of identity and power" (Phillipson 2008, 9). Ives (2006, 130) echoes this criticism, pointing out that De Swaan's model falsely portrays language shifts as occurring independently from "systemic issues of economic and political power or cultural prestige and identity". Moreover, when putting his theory into practice, De Swaan seems to undermine his own model by suggesting that the global prominence of English could override the higher Q-value of Hindi for language learners' preferences in the Indian language constellation (Phillipson 2004, 75). Hjorth-Andersen (2006, 17) acknowledges the utility of De Swaan's prevalence and centrality concepts, but questions the conceptual validity of simply multiplying these two values together – an operation based on the unsubstantiated assumption that the two values should be weighted equally. Of course, performing these calculations necessitates sufficient data for the number of speakers of each language in a given constellation as well as the overlaps in speakers' multilingual repertoires. While these figures may be feasible in certain modern, data-rich contexts, it is exceedingly difficult to obtain reliable and precise data – or even approximations – for earlier time periods. Data would also be much more difficult to obtain for less prominent languages or language constellations. For these reasons, De Swaan's model does not provide a practical or theoretically sound framework for assessing language status. Still, his concept of language constellations is useful, as the status of a language is prone to change according to specific geopolitical contexts. The variability in the status of French between the European and North American contexts illustrates this point. In Europe, French serves as one of the major working languages of supranational governance, and is widely spoken by both native and L2 speakers alike, whereas its presence in North America is highly geographically concentrated, despite enjoying official status in Canada. The language status assessment model presented later in this chapter therefore incorporates the language constellation concept as a means of controlling for this variability.

3.2.2. Sociological models of literary translation flows

Heilbron (1999) models the "sociology of translation" based on international flows of published translations, which are highly asymmetrical across the world's languages. His view applies polysystem theory in its fundamental claim that the "international [translation] system is, first and foremost, a hierarchical structure, with central, semi-peripheral and peripheral languages" (ibid., 433). He quantifies languages' *centrality* – a term used in a manner similar to language status – using annual data provided by UNESCO on the number of translations *from* each language (ibid., 438). In this way, language status is conceptualized as the demand for texts in a given language, measurable in terms of the quantity of translations produced and published from a given SL. Crucially, this model accounts for changes in language status over time, a phenomenon which Heilbron discusses at length (ibid., 434-435). This proposed method is sounder in theory than in practice, however; Heilbron himself concedes that statistics on published translations are highly unreliable, given the apparently erratic fluctuations from year to year (likely attributable to inconsistencies in reporting) as well as varying cultural interpretations of what counts as a published book (ibid., 432-433).

Bourdieu's mentee, Pascale Casanova, offers a highly similar perspective. Casanova (2002) proposes literary capital as an additional form of capital, predicated on power struggles between literary systems. Literary capital may accumulate (somewhat) separately from linguistic capital, but it may also be more accurately referred to as linguistic-literary capital – built on a literary system's prestige (i.e., attitudes toward it) as well as its prior successes and established tradition, the number of translations it inspires, its historical endurance, and so forth (Casanova 2002, 8). Supplanting polysystem theory's center/periphery spatial metaphor, she classifies languages as *dominant* or *dominé* – dominant or dominated – in order to emphasize the power relations actively governing translational activity, which she explicitly purports to be centered on the accumulation of literary capital (ibid., 8). Casanova (ibid., 9) concurs with De Swaan's assertion of the importance of centrality, adapting this concept by positing that literary capital is similarly contingent on the number of translators who facilitate the dissemination of the literature in question to other languages and cultures.

Literary capital also informs translation strategies through its expression of the power dynamics between literary systems. Borrowing Schleiermacher's (1816/2012) example, Casanova (2002, 10-11) observes how the dominant status of French in the Romantic era enticed its translators to render texts according to target-side norms, whereas German's subordinate position in the literary field caused its translators to adopt the opposite strategy. As with many works in translation studies' cultural turn, TL status is presented here as the driving force of normalizing (i.e., TL-oriented) translation strategies, formulated according to perceived differences in the literary capitals of SLs and TLs.

The workability of Casanova's framing proves challenging despite the attractiveness of its theoretical underpinnings. Her approach encounters the same pragmatic barrier as Heilbron's, as the number of literary translators or any potential quantifications of translational activity for a given language are not feasible data to obtain. Moreover, the other factors she lists as contributing to literary capital are highly subjective and thus resistant to straightforward operationalizations. The sociologically-oriented perspectives provided by Heilbron and Casanova to characterize the international system of literary exchanges do not offer adequate grounds for a systematic language status assessment model. Their common theoretical foundation does however urge consideration of a critical question in the context of this thesis: should language status be assessed with respect to language in all possible forms or literature in particular?

Although systematic research on Toury's law of interference should ideally begin with literary translation, as it is commonly asserted to be the form of translation most forcefully impacted by language power relations (see Section 2.4.), conceptualizing the status of a language in terms of its collective (i.e., maximally inclusive) standing is both more fitting and more feasible for the project's aims. Toury (2012, 205) stresses his explicit desire to avoid presenting his target-oriented approach as being restricted to literary translation alone. His formulation of the law of interference reflects this desire in its reference to the power relations between languages or cultures instead of between literary systems specifically (ibid., 314). Other translation scholars echo this sentiment, arguing that marginalized languages must be "as concerned about their technical, commercial and scientific translators as they are about their literary translators" and calling upon scholars to "see translation in all its dimensions as cultural" (Cronin 1998, 155). The socially-cohesive nature of language power is thus an essential component of Toury's law of interference and, consequently, this project's approach.

As illustrated in the previous chapter, language status reflects a collective linguistic capital spanning all domains, with literature merely constituting one possible domain. It is more accurate to frame literary capital in terms of prestige, as agents' subjective attitudes dominate its list of determining factors, among which "prestige" is listed first and foremost (Casanova 2002, 8). The numbers of literary translators and literary translations working into or from a given language are alluded to as quantitative factors (ibid., 8-9), yet these prove unworkable, as conceded by Heilbron (1999, 432-433). It is possible that language power relations have substantially different impacts on literary translation compared to translation in other domains.

Bourdieu's insistence on the interrelatedness of various capitals – implicitly acknowledged in Casanova's brief discussion of the relation between linguistic capital and (linguistic-)literary capital – also merits consideration here, particularly his central claim that economic capital prevails over other downstream forms of capital, e.g., linguistic capital. Thompson (1991, 16) summarizes this key aspect of Bourdieu's thought by stating that "understand[ing] the interests at stake in literary or artistic production" requires the contextualization of these processes with respect to the political economy. Though Bourdieu (1991) perhaps misguidedly circumscribed much of his discussion of linguistic capital in *Language and Symbolic Power* to literary production, Casanova's distinction between linguistic and literary capital illuminates the relationship between the two. The same hierarchical relation between economic capital and other forms of capital (e.g., linguistic capital) also exists between linguistic capital and literary capital, despite their possible divergence.

As highlighted in Section 2.6., the breadth of future research contexts envisioned for this model includes investigations of generic NMT systems and LLMs, whose training corpora are compiled from massive data sets composed of diverse domains. Training data from highly varied domains are ultimately combined into the same translation model via complex machine learning techniques, making it impossible to distinguish the manner in which domain-specific translation norms may influence systems' output. For these reasons, it is advantageous to concentrate on power dynamics at the higher level of language systems rather than literary systems more narrowly. The operationalization of language status based on its linguistic capital – assessed collectively across domains – fulfills the project's aim of providing a language status assessment model for facilitating corpus-based research on the effects of language power dynamics across diverse contexts.

3.2.3. The (E)GIDS framework

Briefly introduced in Section 2.3., Fishman's (1991) theoretical and methodological framework for categorizing language vitality is well-suited to characterize language status, given the relatively stable and observable language hierarchy it conveys. Its distinction from language prestige is made explicit, as Fishman (ibid., 96) contends that prestige is more aptly conceptualized as individuals' or groups' subjective attitudes toward a language, regardless – or even deliberately in defiance – of its status. He also uses status and vitality as near synonyms, generally preferring the latter given his focus on language preservation.

The Graded Intergenerational Disruption Scale (GIDS) provides eight categories of language vitality, and Fishman (1991, 88-109) offers lengthy descriptors for each. The higher numerical levels of the scale apply to more endangered languages, and tend to emphasize the language's lack of intergenerational transmission. The lower numerical levels of the scale apply to more institutionalized languages used consistently in prominent social domains such as government, mass media, and education. The model's emphasis on intergenerational transmission demonstrates the necessity of framing language status temporally: a language's strength or stability hinges on its ability to maintain (or increase) its domains of use and thus its user base over time. Moreover, he highlights the competition between languages varying in status, where "weak" languages tend to concede users and uses to "strong" languages (Fishman 1991, 81). Extinct languages constitute the lowest possible status: they have neither users nor uses (domains of use). Theoretically, languages with the highest possible status would be used in all domains and by all speakers.

According to *Ethnologue* researchers Lewis and Simons (2010, 104), Fishman's GIDS model "remains the foundational conceptual model for assessing the status of language vitality." *Ethnologue* began as an effort to map remote languages for Bible translation, and is managed by SIL International – a Christian nonprofit supporting endangered and under-resourced languages (SIL International 2024a). It now serves as an annually updated database of all languages currently identified in the world (ibid.). As part of their mission, the *Ethnologue* team also aims to assess language status systematically. To this end, Lewis and Simons (2010) synthesize the approaches of Fishman with an earlier model from UNESCO and the previous *Ethnologue* efforts, creating the Expanded Graded Intergenerational Disruption Scale (EGIDS).

Their principal justification for expanding the GIDS model is that it does not encompass the fullest possible scope of language status, as Fishman's highest possible status (Level 1) denotes only a national status, despite the fact that some languages have clearly achieved an international status (Lewis and Simons 2010, 106). Lewis and Simons (ibid., 107) also expand the lower end of the scale to provide a more nuanced understanding of language loss and revitalization, and provide the corresponding UNESCO classifications for each level. The EGIDS model is provided in Table 1 (ibid., 110):

*Table 1: Expanded Graded Intergenerational Disruption Scale**

Level	Label	Description
0	International	The language is used internationally for a broad range of functions.
1	National	The language is used in education, work, mass media, and government at the nationwide level.
2	Regional	The language is used for local and regional mass media and governmental services.
3	Trade	The language is used for local and regional work by both insiders and outsiders.
4	Educational	Literacy in the language is being transmitted through a system of public education.
5	Written	The language is used orally by all generations and is effectively used in written form in parts of the community.
6a	Vigorous	The language is used orally by all generations and is being learned by children as their first language.
6b	Threatened	The language is used orally by all generations but only some of the child-bearing generation are transmitting it to their children.
7	Shifting	The child-bearing generation knows the language well enough to use it among themselves but none are transmitting it to their children.
8a	Moribund	The only remaining active speakers of the language are members of the grandparent generation.
8b	Nearly Extinct	The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency.
10	Extinct	No one retains a sense of ethnic identity associated with the language, even for symbolic purposes.

*From Lewis and Simons (2010, 110).

Whereas previous sociolinguistic approaches merely typify domains, the EGIDS clearly hierarchizes the social domains in which languages are used. This distinguishing feature is apparent in the Level 0 (International) classification, which applies to languages "used internationally for a broad range of functions." Levels 1 (National) and 2 (Regional) also cover a broad range of functions (e.g. education, mass media, and government) on smaller geopolitical scales. The lower levels cover progressively fewer and more private domains, and also become more focused on the language's prospects of intergenerational transmission. Traditionally, translation research – corpus-based or otherwise – tends to focus on languages near the top of the hierarchy, as marginalized or endangered languages tend to have weaker and less standardized writing traditions. For high-status languages, the EGIDS emphasizes a language's level of institutional support as the decisive measure of its status (Lewis and Simons 2010, 107). This institutional factor, in combination with the higher levels' focus on national boundaries (i.e. international, national, and regional), demonstrates an important geopolitical dimension to language status.

In order to sort languages into each of these categories, Lewis and Simons (2010, 113) lay out five guiding questions, the first two of which are most pertinent to corpus-based translation research. The first Key Question asks: *What is the current identity function of the language?* The possible responses are historical, heritage, home, and vehicular. The historical and heritage functions lead to Level 10 (Extinct) and Level 9 (Dormant) classifications. The home function leads to the third Key Question, from which languages are eventually classified between Level 4 (Educational) and Level 8b (Nearly Extinct). Together, these three functions lead to language status classifications that imply a weak writing tradition, and thus are less relevant to corpus-based translation studies. The vehicular function may convey that a language is "used to facilitate communication among those who speak different first languages" or refer to a language used by the overwhelming majority of a nation-state (ibid., 115).

Selecting the vehicular function leads to the second Key Question: *What is the level of official use?* The possible levels of official use are listed with their corresponding descriptors:

- International – The language is used internationally as a language of business, education, and other activities of wider communication. This corresponds to EGIDS Level 0 (International).

- National – The language has official or de facto recognition at the level of the nation-state and is used for government, educational, business, and for other communicative needs. This corresponds to EGIDS Level 1 (National).

- Regional – The language is officially recognized at the sub-national level for government, education, business, and other functions. This corresponds to EGIDS Level 2 (Regional).

- Not Official – The language is not officially recognized but is used beyond the local community for intergroup interactions. These may include business (trade), social or other communicative functions. This corresponds to EGIDS Level 3 (Trade).

Based on the response to this Key Question, languages are sorted between Levels 0 and 3. Throughout the history of the discipline, translation studies has primarily focused on languages falling in this range. Future research may answer scholars' calls to incorporate more marginalized and thus lower-level EGIDS languages into translation studies. Given the EGIDS' ability to classify all possible languages systematically, and its arrangement of domains into a clear hierarchy, this model serves as the ideal foundation for a language status assessment model tailored to corpus-based translation research. The typology's gradient structure naturally lends itself to the operationalization of language status as an ordinal variable, as will be further discussed in the following chapter.

3.3. Presenting a novel language status assessment model

This project modifies the EGIDS framework (Lewis and Simons 2010) to assess the differences in SL and TL status that potentially influence the linguistic composition of translated texts. In this model, language status is determined according to two ordered criteria: the EGIDS scale serves as the first-order criterion, and the approximate number of language users – within the selected constellation in the selected time period – serves as the second-order criterion. Takeaways from the review of sociolinguistic literature in Section 2.6. are combined with the highlighted strengths of the approaches discussed in the preceding section.

Given the contextual variability of language status, it is first necessary to define the specific geopolitical context – or *language constellation*, to borrow De Swaan's term – in which the corpus-based study takes place. Language constellations align most naturally with geopolitical boundaries, such as nations, continents, or supranational bodies like the European Union. Once the language constellation is specified, it is then necessary to delineate a specific time period in order to determine synchronic language status assessments as well as the bounds of the corpora.

Because language status is framed relatively within a localized context, it must be expressed as an ordinal ranking rather than an absolute value. That is, the model may assess the relative positionings of the languages under examination, but it does not stipulate the degree of distance between consecutive or nonconsecutive positionings. As detailed in the next section, the EGIDS provides a ready-made gradation in its initial sorting of levels of use, as does the scale of the approximate number of language users, and languages in the selected constellation may be grouped accordingly. Once a constellation and time period have been defined for the language selection, these also determine the scope of the corpus. That is, the corpus ideally should not include texts from outside the language constellation or time period. As part of a systematic, translation-focused model for assessing language status, then, it is first necessary to establish a set of axioms based on the points raised previously:

1. **Language status is a ranking relative to other languages.** It is not expressed in absolute terms, and the ranking does not permit calculations in absolute terms of the distance between positions on the language status scale. In practice, language power dynamics are dependent on specific, localized contexts, and it is not possible to calculate or compare absolute (i.e., non-localized) measurements for two completely independent language contexts.

2. **This ranking takes place within a specific language constellation.** The language constellation dictates which languages can be included in the study, i.e., each language under examination must be present in the designated constellation. Ideally, all possible language pairs in the constellation should have a history of translation. The language constellation also dictates which texts can be used in the corpora. It provides the population boundaries for the second-order criterion: the approximate number of language users within that specific language constellation may be used to further sort language status rankings beyond the EGIDS levels.

3. **Language status is assessed within a specific time period, as it is prone to gradual change**. It is necessary to delineate a specific time period in which to make this determination, in order to determine the approximate number of users for a language in a defined constellation. This is ideally accomplished by using state-produced statistics, scholarly estimations, or other available indicators of the approximate number of language users at a specific point in time. The median year of publication for the texts comprising translation corpora may serve as the reference point for the aforementioned materials.

With these axioms in place, language status is assessed using a two-tiered process:

I. The language's EGIDS level constitutes the first-order sorting criterion.

II. The approximate number of the language's users constitutes the second-order sorting criterion. The approximate number of language users is calculated within the selected constellation and as close as possible to the corpus texts' median year of publication.

It is important to emphasize that the designated language constellation only influences the second-order criterion. The ordering of these criteria demonstrates a crucial logic underlying the assessment model: the EGIDS level necessarily disregards the designated language constellation. This feature is necessary because the wider recognition of a language's status – especially for status classifications at the higher end of the scale – may be expected to transfer to specific, localized contexts; however, the number of language users in such localized contexts (the second-order criterion), by definition, cannot transcend its locale. In this language status assessment framework, then, the number of language users in the constellation cannot override or negate the initial ordering provided by the EGIDS. It is not always straightforward to quantify, or even necessarily define, language users or speakers. Sociolinguists are typically interested in further distinguishing between first-language (L1) and second-language (L2) speakers or scrutinizing criteria of language competency in order to separate language users/speakers from language learners. These nuances are beyond the scope of the current project. A rough idea of the number – or simply the scale – of language users suffices as the second-tier modifier for this language assessment model.

3.4. Selecting and ranking languages according to status

The current project provides an opportunity to demonstrate the language status assessment process using this new model. This task requires a considerable amount of background research. First, it is necessary to establish the language constellation, as well as the designated time period, in question. The current project focuses on late 19th- and early 20th-century European literature, as this designation is favorable to text availability (see Chapter 4). "European" is retroactively defined by the borders of the modern European Union, including the United Kingdom. The languages involved in the project are presupposed to represent a range of languages varying in status: Croatian, English, French, German, Irish, Italian, and Swedish.

It may be observed that the first-order sorting criterion – the EGIDS – presents a methodological complication given the higher end of the scale's focus on languages' relation to national borders. During the designated time period, European nation-states were not clearly separated, as the continent was home to several sprawling empires, such as the Austro-Hungarian Empire. This arrangement leads to a rather complicated question for the first-order criterion in the language status assessment: should the lingua francas of these empires be considered as "international" or "national" languages? In turn, this dilemma hinges on another question: does a historical empire constitute what the modern EGIDS conceptualizes as a unified nation, or is it better conceptualized as an aggregate of disparate nations, making the historical empire inherently "international"? The answer perhaps depends on the scale and composition of the empire itself. Breuilly (2017, 12) defines an empire as "a state consisting of a core and one or more peripheries" and draws a distinction between pre-modern and modern empires: the former lack a decidedly "national core" and the latter are built around one. This distinction is further explored in the specific assessments of English, French, German, and Italian.

Furthermore, what exactly is meant by "international"? *Ethnologue* indicates the category's exclusivity: "EGIDS 0 (International) is a category reserved for those few languages that are used as the means of communication in many countries for the purposes of diplomacy and international commerce" (SIL Institute 2024b). As such, only six languages are currently classified as EGIDS 0 (International): Arabic, Chinese, English, French, Russian, and Spanish. It is no coincidence that the languages classified as EGIDS 0 are the six official languages of the United Nations, the world's most powerful supranational institution. Clearly, the EGIDS 0 classification is rather exclusive: it does not simply refer to those languages spoken in multiple countries, but rather to those that are used on a truly global scale. National (Level 1) languages are described as being primarily on the level and scale of the *nation-state* (ibid.).

In order to categorize language status for translation research in historical contexts, then, it is necessary to scrutinize the descriptors offered in the predetermined EGIDS Key Question responses so that a historical language selection may be retroactively fitted with approximate categorizations in this model. Of course, this process requires a considerable amount of subjective interpretation on the researcher's part, but these descriptors offer a reasonably solid foundation for this undertaking. Considering the corpus texts' temporal positioning in the advanced stages of European colonialism, it is necessary to transpose the exclusivity of the EGIDS 0 category onto the nation-states and empires affiliated with each of these languages by assessing the geopolitical reach of these political entities. In order for a language to register as an EGIDS 0 language for this period, its imperial expansion must have achieved a truly global scope that would roughly align with the dominance of the current United Nations official languages. As such, a language that is merely spoken in multiple countries within a confined region does not qualify as "International", and instead is categorized as "National". The project's language selection warrants that the principal distinction for the language status assessment model's first-order sorting criterion is between these two EGIDS levels: "International" (EGIDS 0) and "National" (EGIDS 1). Following the establishment of these primary categories, language status is further distinguished according to speaker populations.

As with corpus design, it is crucial to be as transparent and intentional as possible in justifying each step of the language status assessment framework. For the second-order criterion, emphasis must be placed on the *approximate* number of speakers, as exact data are expected in many cases to be difficult or impossible to find. Moreover, easing this process is the stipulation that the emphasis on *rankings* of language status only requires the determination of which language has more speakers and not *how many* more. In order to make this determination, it is ideal to obtain quantitative figures like census data or other scholarly estimations of language community sizes for the median year[1] of publication for all texts in the corpus. For all texts in the corpus constructed for and used in this project (see Chapter 4), the median year of publication is 1909.

Admittedly, it is often challenging to obtain reliable data on the number of language users in historical contexts. In such cases, it may be necessary to infer the approximate number of language users based on the population sizes of countries where the language is primarily spoken. This method is, of course, an imprecise metric for the number of language users, since it necessarily excludes language users in countries outside of those where the language is primarily spoken. (Alternatively, the EGIDS captures the scale of a language's use by speakers outside of its primary geopolitical frame by designating languages as international, national, or regional.) The cumulative population sizes of primary language-speaking countries may provide a reasonable idea for the general size of language communities. At minimum, they are sufficient to determine differences in the *scales* of language communities and, therefore, to rank language status in relative, ordinal terms.

The main outcome of the language status assessment model is therefore a ranked hierarchy of the selected languages. Still, superordinate groupings of the ranked languages may prove useful to the project's data analysis, as they may reveal consistent patterns among languages that are similar in their status positionings. Analyses on the basis of these groupings may also at least partially mitigate the limitations of relatively small sample sizes of translations in individual language pairs as described in the following chapter. Naturally, the initial sorting mechanism for EGIDS 0 and EGIDS 1 languages provides readymade, hierarchized groupings of languages, where EGIDS 0 languages may be designated as relatively "high-status". According to the model, "National" languages are then further sorted by their approximate number of speakers, such that languages with comparable speaker population sizes may be further grouped together as, e.g., "medium-status" and "low-status" languages, in the event that the population sizes of these EGIDS 1 languages substantially differ.

With these sorting criteria in place, the current project applies the two-tiered language status assessment model to the selected languages in the subsequent section.

[1] If data are not available for the precise median year, data for the nearest year are provided.

3.4.1. English

It is necessary to start with the first Key Question in the EGIDS decision tree: *What is the current identity function of the language?* (Here, it is important to note that "current" should be interpreted as the relevant time period, or more precisely, in the median year of publication for the texts comprising the corpora under examination.) In the late 19th and early 20th century, English was a vehicular language, since it was undeniably "used to facilitate communication among those who speak different first languages". English had long been established as a lingua franca across Europe: Crystal (2003, 75) refers to an 1829 writer who described already how extensively English was taught in educational systems around the world. The sprawling British Empire cast the language across most of the world map, mandating its use as a vehicle of "political unity" (ibid., 79).

Given this response to the first Key Question, it is necessary to move to the second: *What is the level of official use?* Undoubtedly, English registers as an "international" language during the relevant time period, as it stood firmly as "the dominant language of global politics and economy" upon the dawn of the twentieth century (Crystal 2003, 85). English would further solidify its international prominence shortly following the project's reference year (1909): alongside French, it became widely used as a language of international diplomacy in the early 20th century (Ammon 1992, 426). Established with the Treaty of Versailles in 1920, the League of Nations (a precursor to the United Nations) had two official languages: English and French (Crystal 2003, 86-87). Clearly, English had already established its status on the global scale. It thus registers as an EGIDS 0 (international) language.

With English established as an EGIDS 0 language, it is now necessary to turn to its approximate number of users. Again, it must be emphasized that the number of users is assessed only within the constellation (i.e., Europe). The project uses the cumulative population of English-speaking countries as a proxy for the general number of English speakers. As reported in the census, the population of the contemporary United Kingdom (including England, Wales, Scotland, and what is now Northern Ireland) in 1911 (the nearest available year) was roughly 42,082,000 (Macrory 2010, 29). The entire population of Ireland (excluding the six counties of Northern Ireland, which were already included in the UK census) at this time was approximately 3,139,688 (Hindley 1990, 23). In total, then, the population of the primary English-speaking countries was roughly just over 45 million in 1911. In reality, this figure was likely much higher, considering the language's emergence as a lingua franca in Europe as described earlier. The EGIDS 0 classification combines with this high total to establish English as a high-status language in the context of this current project.

3.4.2. French

For the designated time period, French constituted a vehicular language. Throughout the 19th and much of the 20th century, it was the "pre-eminent vehicular language in Europe" – a common language among Europe's upper classes (De Swaan 2001, 16-17). The second Key Question leads to another obvious designation as an EGIDS 0 (international) language, given the role of French alongside English in the League of Nations, and since French – and, to a lesser extent, Belgian – colonial exploits spread the language around the world (Wright 2006, 37).

Turning to the approximate numbers of French users, it is necessary to note that Europe had three primary French-speaking countries: France, Belgium, and Switzerland. In 1909, the population of France (excluding its territories outside of continental Europe) was approximately 39,024,322 (Institut National d'Études Démographiques n.d.). The population of Belgium in 1910 was reported to be 7,423,784, though not all were French speakers (Direction générale Statistique - Statistics Belgium 2017). Given the comparable sizes of the Dutch-speaking and French-speaking populations in Belgium, it is reasonable to take half of the total population (3,711,892) as a very rough estimate of the number of French speakers. The total population of Switzerland in 1910 was listed as 3,753,300, and roughly 21.1% were French speakers, meaning the total French-speaking population in Switzerland at the time was approximately 791,946 (The Swiss Federal Statistical Office 2021a, b). In total, the cumulative size of the French-speaking populations in these three countries is estimated to be around 43,528,160. The actual number of French users is likely higher, given that figures for the number of users among Europe's upper classes are not included in these figures. The EGIDS 0 classification combines with this high total to establish French as a high-status language in the context of this current project.

3.4.3. German

The task of determining German's relative status within the project's language selection illustrates the complexity of retroactively assigning EGIDS categorizations to historical languages. In the late 19th and early 20th centuries, German was the dominant and unifying language of two major imperial powers on the European subcontinent: the German Empire and the Austro-Hungarian Empire. In order to classify as an EGIDS international language, however, it is not enough for a language to be used across multiple states – it must also reach the global scale of the modern UN official languages. With this in mind, it is necessary to examine the global reach of the German language.

While the Austro-Hungarian Empire constituted an amalgam of bordering and semi-autonomous nations (Breuilly 2017, 22), the German Empire managed to transcend Europe and establish a colonial presence beyond the European continent (see Conrad 2011). German colonialism was significant in reach, but it did not endure as long as British and French colonialism (ibid., 1). Austro-Hungarian and German imperial pursuits were both thoroughly dismantled over the course of World War I. Both empires were disbanded into their constituent nations upon the war's end (Kumar 2010, 123), and the Treaty of Versailles rid the German Empire of its colonial territories (Conrad 2011, 6).

It is crucial to define the scope of the language's use on the world stage. De Swaan (2001, 13) asserts that German had nearly achieved the prominence of English and French near the beginning of the 20th century. However, the German language did not achieve the same territorial reach as English or French (Ammon 1992, 433). Darquennes (2006, 63) puts forth that German's prestige in the mid to late 19th century was acquired by virtue of its speakers' scientific advancements rather than their comparatively meager colonial aspirations, and stresses that German was never a true competitor to the global hegemony of English and French at the turn of the century.

With this background, it is evident that German did not transcend the European language constellation in the same manner as English and French. German therefore registers as an EGIDS 1 (national) language.

With German established as an EGIDS 1 language, it is necessary to turn to the approximate number of users. The population of the German Empire in 1910 was approximately 64,925,993 (Sensch 2007). Census data reports the population of Austria in 1909 (excluding all other constituent nations in the Austro-Hungarian Empire) as 6,517,500 (Statistik Austria 2022). Approximately 69.1% of the Swiss population spoke German in 1910, meaning that there were roughly 2,593,530 German speakers in Switzerland (The Swiss Federal Statistical Office 2021a, b). There were also negligible German-speaking minorities in Belgium and Luxembourg. In total, the size of the German-speaking populations in these three countries is estimated to be around 74,037,023. The EGIDS 1 classification combines with this total to establish German as a medium-status language in the context of this current project.

3.4.4. Italian

In the middle of the 19th century, Italian was standardized around the Florentine dialect, which was elevated above Italy's numerous other dialects to serve as the country's unifying language (Berruto 2018, 495). Thus, Italian registers as a vehicular language in response to the first Key Question. Like the German Empire, Italy established a colonial presence in Africa that was not nearly as "established" as those of the British and French (Srivastava 2018, 1). International treaties scarcely used Italian during this time period (Ammon 1992, 428). Since the Italian language never established itself on a global scale, it therefore classifies as an EGIDS 1 language.

The resident population of Italy in 1911 was reported as 35,845,000 (Istituto Nazionale di Statistica 2012, 98). In 1910, roughly 8.1% of the total Swiss population of 3,753,300 spoke Italian, meaning that there were approximately 304,017 Swiss Italian speakers (The Swiss Federal Statistical Office 2021a, b). In total, there were approximately 36,149,017 residents in the Italian-speaking regions of the European language constellation in 1910. The EGIDS 1 classification combines with this total to establish Italian as a medium-status language in the context of this current project.

3.4.5. Swedish

Sweden has long been a predominantly Swedish-speaking country, and the language has been mostly confined within the country's borders. Finland was ruled by Sweden for much of the nation's history, and still has a small Swedish-speaking minority (Östman and Mattfolk 2011, 75). These characteristics make Swedish a vehicular language. Given the language's historical stronghold over its home nation, Swedish classifies as an EGIDS 1 language.

The population of Sweden in 1909 was reported to be 5,476,441 (Statistics Sweden n.d.). The population of Finland in the same year was reported to be 2,914,800 (Statistics Finland n.d.), meaning that the Swedish-speaking minority could not have increased the total number of Swedish speakers to the degree that it would affect the language status classification. The cumulative Swedish-speaking population was therefore likely far below 8.5 million. Clearly, this total is far lower than those of the German- and Italian-speaking populations depicted earlier. The EGIDS 1 classification combines with this estimate to establish Swedish as a low-status language in the context of this current project.

3.4.6. Croatian

Croatian presents another interesting case for this language status assessment model. Croatian is often considered to be (part of) a "pluricentric" language – that is, "a language which serves different populations in different states, taking on different guises as necessary" (Alexander 2006, 425). This pluricentric language is now typically referred to as BCS (Bosnian-Croatian-Serbian) or BCMS (Bosnian-Croatian-Montenegrin-Serbian). It has long been debated whether Bosnian, Croatian, Montenegrin, and Serbian are four separate languages or a single language with national variants (ibid., 379). In addition to the language politics derived from these distinct national identities, significant dialectical differences were and continue to be present in this pluricentric language. While Croatia historically embraced all three dialects (štokavian, čakavian and kajkavian), Serbia insisted upon the supremacy of štokavian (Alexander 2006, 390-391). Nonetheless, the štokavian dialect serves as the "basis for all the literary standards subsumed in BCS" given its unmatched geographical coverage (ibid., 388).

The history of language standardization and unification in the South Slav countries is complex. In 1850, a joint agreement between Croatian and Serbian linguists stipulated that their respective languages were actually one and the same, differing only in their writing system – Croatian used the Latin alphabet, while Serbian used the Cyrillic alphabet (Alexander 2006, 385). From that point in history, intermittent efforts to distinguish and unify the BCMS national variants reflected the region's political turbulence. Croatia was under Austro-Hungarian rule up until the end of World War I, when it became part of the Kingdom of the Serbs, Croats, and Slovenes, whereas Serbia had already been an independent kingdom (ibid., 384-385). However, Croatian linguists have made efforts to distinguish Croatian lexis from other national variants, particularly Serbian, for most of the past century (ibid., 402). Bearing these points in mind, Croatian is treated as a separate language in the context of this project.

Given its undeniable historical importance in establishing a national identity (Alexander 2006, 388), Croatian classifies as an EGIDS 1 language. Even if Croatian is assumed to have the widest possible reach by adopting the geolinguistic boundaries of BCMS for the purposes of this project, it is clear that the language would still not classify as "international" on the same scale as EGIDS 0 languages. The population of Croatia in 1910 was reported to be 3,460,584 (Croatian Bureau of Statistics 2018, 107). As with Swedish, this total is far lower than those of the German- and Italian-speaking populations depicted earlier. The EGIDS 1 classification combines with this total to establish Croatian as a low-status language in the context of this current project.

3.4.7. Irish

Irish presents a unique case, as the roles of translation and literature in its historical demise and subsequent (partial) resurgence have drawn much scholarly attention (see Fishman 1991, 122; Edwards 1994; Tymoczko 1999/2014; Fhrighil et al. 2020). Applying the language status assessment model to Irish leads to two other Key Questions in the

EGIDS model that have not yet been used. In response to the first Key Question, the identity function of Irish does not rise to the level of a vehicular language, as the language was still very much minoritized in relation to English during the relevant period in Ireland (Ó Buachalla 1984; Hindley 1990). Instead, Irish is classified as a home language, as it was "used for daily oral communication in the home domain by at least some" (Lewis and Simons 2010, 113). This response leads to the third Key Question: *Are all parents transmitting the language to their children?* An affirmative response requires that "intergenerational transmission of the language is intact, widespread and ongoing" (ibid., 115). It is highly debatable whether the transmission of Irish from generation to generation was sufficiently "widespread" and consistent, yet it is undeniable that there remained a resilient, if diminishing, subpopulation of L1 Irish speakers around the turn of the century (Ó Laoire 2009, 286). Clearly, not all parents were transmitting Irish competency to their children. However, a negative response to the third Key Question leads to another question focusing on the youngest generation of proficient speakers, which would seem to muddle the historical trajectory of Irish. For the sake of convenience and the purposes of this project, it is therefore assumed that Irish at this particular historical juncture reflects adequate intergenerational transmission, resulting in an affirmative response to the third Key Question.

This outcome leads to the fourth Key Question: *What is the literacy status?* The answer is not straightforward, either. Naturally, literacy is closely connected to education policy, and the Irish language's position in Ireland's education system during this time was rather complicated. Established in 1893, the Gaelic League led a concerted political campaign to promote the general public's competency in Irish by codifying mandatory Irish-language instruction in Ireland's education system; these efforts generated limited yet noteworthy success, as Irish was taught as a non-compulsory subject in roughly one-fifth of the Republic of Ireland's national schools in 1909 (Ó Buachalla 1984, 83-84). The fourth EGIDS Key Question's middle response option between "institutional" and "none" is "incipient", subsequently leading to the language's categorization at the EGIDS 5 level ("Written"), which does not seem to capture the state of Irish in the relevant historical period, either. Therefore, it is once again assumed for the sake of convenience that the literacy status of Irish at this time is characterized as “institutional”. This outcome leads Irish to be classified at the EGIDS 4 level (educational).

The latter half of the 1800s saw a massive shift from Irish to English (Hickey and Amador-Moreno 2020, 11). By the end of the 19th century, the total proportion of Irish speakers had fallen to less than 15% of the Irish population, and a miniscule number of these Irish speakers were monolingual (Hindley 1990, 19). The percentage of Irish speakers in Ireland did not rise significantly in the first few decades of the 20th century (ibid., 23). The decline of Irish language use continued even after Irish independence in 1922 (ibid., 219).

According to Census Reports, there were approximately 582,446 Irish speakers across all of Ireland in 1911 (Hindley 1990, 23). The EGIDS 4 classification combines with this total to establish Irish as a *very* low-status language in the context of this current project, as its level of domain capture demonstrates it to be categorically different from the two low-status languages. While the assumptions made in answering the EGIDS model’s Key Questions are certainly subject to scrutiny, the only possible alternative responses would have led to an even lower-level classification, meaning that Irish would nonetheless register as a very low-status language in this project.

3.5. Status groups

With these language status assessments, the project groups languages according to their comparable positions on the status hierarchy:

Table 2: The status rankings and status groups of the project's selected languages

Language	Ranking	Status group (SG)
English	1	high-status
French	2	high-status
German	3	medium-status
Italian	4	medium-status
Swedish	5	low-status
Croatian	6	low-status
Irish	7	very low-status

As will be demonstrated in the following chapter, these groupings allow translation subcorpora to be formed around status pairs (SPs), e.g., translations from high-status into low-status languages.

3.6. Chapter conclusion

This chapter has examined several approaches to language status in sociolinguistics and related disciplines, highlighting the strengths and weaknesses of each. It has offered a novel, systematic approach to assessing language status for the purposes of empirical translation research. This new approach is based on two ordered criteria: the first-order criterion is Lewis and Simons' (2010) EGIDS model, and the second-order criterion is the approximate number of language users in the designated language constellation. The justification for this two-tiered assessment model was rooted in the extensive sociolinguistic literature on the nature of language status. Using a modified version of EGIDS as the foundation for this translation-focused language status assessment model also creates much-needed links between empirical translation studies and related disciplines, in particular sociolinguistics (De Sutter and Lefer 2020, 19). Instead of creating an isolated system of determining language status from scratch, adopting the EGIDS allows empirical translation researchers to draw upon a widely used global standard for assessing language status and grants access to a wealth of supporting data as well as nuanced explanations of its categories and applications. Moreover, an international team of experts provide the data and classifications for this first-order criterion, and are continuously monitoring and updating these categorizations accordingly (SIL International 2024b). The addition of a second-order classification provides a layer of granularity necessary to distinguish languages classified at the same EGIDS level.

The project has applied this novel language status assessment model to a range of European languages, ranking them as follows (in descending order): English, French, German, Italian, Swedish, Croatian, and Irish. Subsequently, English and French have been categorized as high-status languages, German and Italian have been categorized as medium-status languages, Swedish and Croatian have been categorized as low-status languages, and Irish has been categorized as a very low-status language. Using these classifications of language status, the project will measure SL influence in translation using comparable corpus methodology.

4. Methodology

4.1. Chapter introduction

This chapter outlines the project's methodological approach, both in terms of the design and construction of its corpus and as well as its statistical analysis. It first discusses the core concepts underlying the corpus design, providing the rationale for its genre selection and its process of selecting and retrieving texts. The chapter then offers a detailed account of the practical decisions and operations related to the preparation of texts for corpus processing, as well as summary statistics depicting the composition of the corpus and its various parts. It subsequently describes and justifies its statistical approach, including the necessity of its primary and secondary data analyses.

4.2. Corpus design

The overarching purpose of this corpus is to investigate a possible positive association between SL status and SL influence on translated texts representing diverse language pairs. The alleged ubiquity of this association aligns with the translation universals research agenda, which has underpinned corpus-based translation studies from its inception. Translation universals are those features which "typically occur in translated text rather than original utterances" regardless of the language pair (Baker 1993, 243). As Chesterman (2004, 39) contends, hypothesized universal features of translation may be characterized as S-universals or T-universals; features constituting the former are determined by comparison between the translation and its source text, and features constituting the latter are determined by comparison between the translation and a comparable corpus of original TL texts.

The precise form of corpus methodology appropriate for translation research depends on whether S-universals or T-universals are being examined: parallel corpus methodology typically suits S-universals, whereas comparable corpus methodology typically suits T-universals. Chesterman (2004, 40) categorizes Toury's law of interference as an S-universal and Baker's notion of "conventionalization" (i.e., normalization) as a T-universal. Broadly speaking, these proposed translation phenomena may be conceived as diametrically opposing concepts, although the manner in which they are being investigated should determine whether they may be jointly operationalized, as will be discussed later on in this chapter. Scholars have long advocated for the combination of comparable and parallel corpus methodologies in translation research (Olohan 2004, 43; Laviosa 2008, 309; McEnery and Xiao 2008, 22). The expanded capabilities that this combined approach provides are perhaps most thoroughly demonstrated by Teich (2003), whose research on syntactic interference in translation will be discussed in Chapter 5. The mixed approach combining comparable and parallel corpus methodology may constitute the ideal approach for a project of this nature, but practical limitations preclude its actual implementation.

The construction of "robust and reliable parallel corpora" is widely recognized as "demanding and laborious work" (Zanettin 2013, 30). In praising the dynamism of the bidirectional Norwegian-English Parallel Corpus, Chlumská (2018, 105) also notes the extreme difficulties of constructing such corpora, pointing to the asymmetries in the availability of texts in different language pairs and translation directions, which are caused by "the status of the language in terms of its prominence, general demand for certain text types in a given culture, etc." Given the range of language included in this project, which intentionally includes low-status languages, it was self-evident that the construction of parallel corpora for various possible language pairs would not be feasible. Chesterman (2004, 43) asserts that this shortcoming prevents the investigation of interference in translation, as research on S-universals relies on comparisons between translations and their source *texts*. Contrary to Chesterman's claim, however, Zanettin (2012, 21) contends that Toury's hypothesized laws and the universality of interference in translation may be "tested by comparing texts translated from one or more languages versus a comparable corpus in the target language." The strengths of parallel corpus methodology in identifying and articulating the nature of S-universals such as interference may thus be replicated by comparable corpus methodology, provided that the comparable corpora are "sufficiently large and balanced" with texts translated from a variety of source languages (ibid., 47-48). In fact, Bernardini and Ferraresi (2011, 228) assert that monolingual comparable corpora are "arguably more versatile resources than parallel corpora" as they often prove more amenable to analytical tools and methods by replacing the "painstaking, low-level analysis of parallel concordance lines" with a higher-level focus on the frequencies of linguistic features.

The role of parallel texts in the current project requires clarification. For instance, an additional advantage of *bidirectional* parallel corpora is that they contain built-in comparable subcorpora: in the Norwegian-English Parallel Corpus, Norwegian translations may be compared not only with their English source texts but also with the comparable, original Norwegian texts. The comparable Norwegian and English corpora may be considered incidental in the sense that they are secondary to the (intended) central appeal of the corpus' parallel structure. Similarly, the corpus constructed in this project was envisioned to reflect the inverse phenomenon, wherein the comparable subcorpora for each language could contain *some* source texts of corresponding translations in the translation subcorpora, but this was not a primary aim in designing the corpus.

This thesis requires the construction of a multilingual comparable corpus, including translations for as many different language pairs as possible among the selected languages, and in which each of the selected languages has a subcorpus of comparable texts. The benefit of this design is that any given SL>TL translation may be compared to a comparable SL subcorpus and/or a comparable TL subcorpus. As will be discussed later on, translations may be grouped into many different subcorpora, though generally they are grouped according to their common TLs and SLs. The corpus' primary organizing principle thus entails:

- Subcorpora for all translations into a given TL (termed the "fixed [language] TL subcorpus")
- Subcorpora for all translations from a given SL (termed the "fixed [language] SL subcorpus")

Naturally, subcorpora may also be formed around each language pair, though there are significant differences in the sizes of language pair subcorpora, as illustrated later on. With the basic corpus design and the overarching corpus-building approach in place, it is necessary to provide a more thorough discussion of the project's genre selection.

4.2.1. Genre selection

As described in Chapter 2, literary translation has traditionally been translation studies' primary focus, and assertions of the effects of language power relations on the composition of translated texts have often envisioned this phenomenon in the context of translated literature. It is therefore rational to focus this project's systematic investigation of SL influence on translations in the literary sphere. However, a more in-depth discussion of the project's genre selection is needed in this section.

There have been a multitude of definitions put forth for genre as well as other highly related concepts such as register and text type in linguistics (see Lefer and Vogeleer 2013, 13-15). Lee's (2001, 46) popular framing distinguishes between register and genre merely as "two different points of view covering the same ground." He defines the former as a view of "text as language" with context-specific functions and the latter as the perception of a text's externally-defined membership in a category (ibid., 46). Given the shared emphasis on characterizing interference and normalization in translation using corpus methodology, this project follows the example set by Lefer and Vogeleer (2013) in adhering to this general conceptualization of genres as "culturally recognised artifact[s]" (Lee 2001, 46). This view closely aligns with Toury's (2012, 201) assertion that literature is "first and foremost a kind of cultural institution" expressed as a system of texts generally perceived to be typologically related. In this manner, a target culture's literature provides a "target model of text formation" – Toury's definition of "genre" – for translators of literary source texts (ibid., 202). The usefulness of comparable TL corpora within this theoretical model is evident, as interference may be partly characterized by the deviation of a translation's features from those of comparable texts in the target culture.

A major challenge in delineating genres is the possible variation in their "levels of generality" (Lee 2001, 48). Lee (ibid., 48) posits "literature" as a superordinate category encompassing such "basic-level" genres as novels, poems, and dramas, which contain further "subgenres" such as, e.g., romance and adventure novels. Based on this typology, this thesis designates literary fiction (i.e., prose texts, including novels, novellas, and short stories) as the genre under examination. It is envisioned that this genre designation provides ample flexibility for the expected cross-cultural variation in literary subgenres (see López-Arroyo 2020), yet remains specific enough to ensure a sufficient degree of comparability. This designation also aligns with Zanettin's (2012, 45) glancing reference to "written fiction" as a cohesive genre. The limitations of the project's genre designation along with the potentially significant consequences of disregarding register in constructing the corpus will be revisited in the thesis' conclusion.

The previous chapter highlighted the need to assess language status with respect to a specific time period, also stressing the importance of temporal specificity for constructing this project's corpus. As indicated earlier, the thesis primarily focuses on *historical* literary fiction for practical purposes. The copyrighted protections of recent literary works often render them inaccessible to researchers in absence of special permissions, particularly for the digital versions of texts that would be amenable to corpus processing (Zanettin 2012, 52). In light of this constraint, the project focuses on works with expired copyright protections, as these constitute publicly available texts likely to be available in pre-digitized format. Under standard international copyright laws, works available in the public domain tend to be those which were published in the mid-19th to early-20th century. Historically, texts in major European languages also tend to be more consistently available. The texts under examination in this thesis are therefore European literary translations published in the mid-1800s to early 1900s.

4.2.2. Text selection

Sampling methods for selecting texts to form a corpus are of paramount importance in corpus linguistics, as their proper execution ensures that findings from the corpus may be generalized to the larger population that it is intended to represent. However, it has long been acknowledged that it is exceedingly difficult, both theoretically and practically, to apply "sampling" – a concept originating in statistics – to research on naturally-occurring language, given the often-impossible task of demarcating target populations and the inevitability that they are in some manner inadequately represented in the corpus (see Atkins et al. 1992).

In theory, the textual population to be represented by this corpus-based research project is all translated literary prose between the selected languages published from the mid-19th to early-20th century. Unfortunately, there are no reliable resources that accurately define this population of texts; as perhaps the most ambitious and far-reaching effort to keep record of published translations, the UNESCO's Index Translationum database is highly inconsistent across languages and excludes translations pre-dating the 1930s (Heilbron 1999, 433). This gap precludes the use of any rigidly systematic approach – e.g., random sampling or a stratified approach – to select texts to represent the target population (see Kenny 2001, 107). As such, it was necessary to devise an *ad hoc* sampling frame in view of the project's overarching aims. This improvised sampling frame prioritized achieving a reasonable balance in the language pairs represented in the translation subcorpus, thus approximating a stratified approach in which the slots represent all possible language pairs and translation directions as well as their superordinate arrangements (e.g., status pairs, SPs).

As briefly discussed in the chapter's introduction, the availability of translations for the various possible language pairs was expected and later confirmed to be highly uneven, meaning that the aim of procuring translations to represent rare language pairs in the corpus necessarily superseded the possibility of selection based on other criteria. The construction of this project's corpus was therefore heavily reliant on text availability, such that the processes for text selection and text retrieval were mutually informative and largely concurrent.

Selection criteria for both the translation and comparable subcorpora were formulated in line with Zanettin's (2012) basic principles for the design of translation-focused corpora. In taking "written fiction" as a sample genre, he considers authors, publishers, and publication dates to be the most pertinent sampling criteria in selecting both translated and non-translated texts for an adequately representative corpus (ibid., 45). The text selection process principally sought to capture the most prominent authors for each language within the selected language constellation and time period, while also attempting to avoid overrepresenting any single author, as in Van Poucke's (2011, 107-108) investigation of loanwords in literary translations. (Data on the balance of authors in the various subcorpora are provided later in this chapter.) As indicated in the coarsely stratified approach to sampling based on balancing translations for different language pairs, translation texts were identified and acquired first. The advantage of this method was that it provided a strong indication of which texts and authors might be considered internationally renowned, and therefore also reasonably representative of a given language's literature during the relevant time period. Potential pitfalls of this approach to text selection will be explored at length in the thesis' conclusion. Here, it is worth briefly noting that this presupposition emulates Casanova's (2002, 13) view that the "great heroes of literature" (*les grands héros de la littérature*) are consecrated according to their success within well-established national literary traditions; this esteem may only be accessed by authors of dominated (i.e., peripheral) national literatures via translation into dominating (i.e., central) national literatures, whereas authors of those dominating national literatures are already in high demand for translation.

In this manner, translated texts were identified and confirmed to be available prior to the original texts for the comparable subcorpora, as the original texts by authors whose works appeared in translation were taken to be ideal candidates for the construction of representative comparable subcorpora. The majority of the original-language texts for comparable subcorpora and translations were obtained from online repositories described in the subsequent section. However, these digital repositories primarily hosted texts in the project's high- and medium-status languages, and did not tend to provide translations into or from low-status languages. In order to achieve an adequate balance of translations for all possible translation directions among the seven selected languages, it was therefore necessary to identify and procure available print translations for rare language pairs such as Croatian>Swedish from various international booksellers. For language pairs with few or no digital translations available, print translations were typically identified by searching for popular SL authors in TL versions of Wikipedia, as authors' pages tended to provide information about their translations. Library book aggregators such as WorldCat[2] offered helpful guides for identifying existing translations, while websites such as the Royal Irish Academy's Historical Irish Corpus[3] also provided useful language-specific resources.

[2] https://search.worldcat.org/

[3] http://corpas.ria.ie/

A modest attempt was made to represent different author nationalities for pluricentric languages; for instance, the comparable German subcorpus includes Austrian and Swiss authors. However, this process was also constrained by text availability and which authors tended to appear in translation. Regardless, texts written by authors from outside the European language constellation (e.g., the original or translated works of North American authors writing in English) were entirely excluded from the corpus. The text selection process also attempted to balance as much as possible the gender of authors represented across the comparable and translation subcorpora. Texts were also selected with an eye toward attaining a reasonable balance of publication dates within the time period in question, although once again, the scarcity of translations in certain language pairs – particularly those involving low-status languages – required flexibility in this area. In cases where certain SLs, TLs, and language pairs seem to have few candidate texts available for the selected time period, more recent translations were identified and obtained.

Lastly, it should be noted that the text selection process made a reasonable effort to ensure that each translation was a direct translation of its source text. In many cases, the language pair was explicitly stated in the text's imprint. In several cases, confirming the text as a direct translation required researching the translator, although it was not always possible to confirm outright that they translated directly from the source text. The possibility of covertly indirect translations (see Pięta 2017) appearing in the corpus will be further explored in the conclusion. With this overarching strategy for text selection in mind, the practical details of retrieving selected texts for the translation and comparable subcorpora are explored in the following section.

4.2.3. Text retrieval

Whether translations or original texts, pre-digitized texts were retrieved from online archives wherever possible. Most of the corpus texts, whether original or translated texts, were retrieved from Project Gutenberg[4] – an online repository of digital texts, primarily literary fiction, that are available in the public domain according to US law.

[4] https://www.gutenberg.org/

Other texts were retrieved from the Internet Archive[5]. Language-specific digital repositories also provided a reliable source of texts for the corpus. Litteraturbanken[6] provided two Swedish translations of works by Franz Kafka, and Project Runeberg[7] provided Swedish translations of several English and French novels. Nearly all of the Croatian texts were drawn from the Portala e-lektire[8] – a wide-reaching and publicly funded initiative to provide and enhance literary educational materials for Croatian students.

4.4. Pre-processing texts

4.4.1. Converting texts into plain text files

All texts, regardless of their source, needed to be converted into plain text files (.txt) with UTF-8 character encodings. Digital archives of historical literature typically consist of texts that have been digitized using optical character recognition (OCR), the process of rendering images of letters and words machine-readable. This process necessitates a person or team of people manually scanning printed texts, automatically converting the scanned images into a machine-readable format using OCR software, then proofreading and correcting the output – a very time- and energy-intensive process.

The major benefit of using Project Gutenberg was that its uploaded texts are already processed with OCR software, thoroughly proofread by a team of volunteers, and converted into plain text files (Brooke et al. 2015, 43). Other text files needed to be proofread using language-specific spell-checking tools in Microsoft Word and, if possible, manual comparisons between OCR output and the text's actual print pages. Internet Archive's uploaded texts tended to constitute plain text files converted directly from raw OCR output, meaning that its texts needed to be proofread manually. Language-specific

[5] https://archive.org/

[6] https://litteraturbanken.se/

[7] https://runeberg.org/

[8] https://lektire.skole.hr/

repositories mostly provided texts in the standard e-book (.epub) file format. All .epub files were converted into a machine-readable format using the calibre ebook management software program[9], then subsequently proofread and converted into plain text files.

Print translations required the most effort, as their preparation also involved manually scanning all book pages then processing the scanned images with the OCR software ABBYY FineReader 15. The OCR output was subsequently edited using the software's proofreading tool, which includes features such as multilingual dictionary lookup and the identification of low-confidence characters. The proofread OCR output then underwent another round of proofreading in Microsoft Word. The author's language competencies for the project's selected languages are presented in Table 3 below:

Table 3: Author's competencies in the project's selected languages

Language	Level of competence
English	native fluency
French	basic
German	advanced
Italian	advanced
Swedish	intermediate
Croatian	basic
Irish	none[10]

[9] https://calibre-ebook.com/

[10] While proofreading the OCR output of print translations in Irish, I compensated for my lack of competency in the language by exercising extreme care in comparing the software's output against the print text, relying less on the available spell-checking tool and more on visual verification of Irish orthography.

4.4.2. Removing paratexts

Regardless of their origin, all texts included in the corpus contain a variety of paratexts (see Batchelor 2018) that are ancillary to the body of the work itself, henceforth referred to as the “main text”. For example, all text files retrieved from Project Gutenberg include a lengthy description of the terms of use for texts retrieved from the repository. Clearly, such paratexts are not relevant to this project’s research question. Determining the other kinds of paratextual phenomena to be excluded is decidedly less straightforward, however. Paratextual elements of literary works may include footnotes, tables of contents, author forewards, translator prefaces, publisher imprints, and the like – the nature and appearance of which vary widely across texts in the corpus. The division between a main text and its associated paratexts is heavily debated, particularly for texts in non-traditional formats (ibid., 54). What is needed in the course of this thesis, then, is a clear, consistent, and workable process of drawing the main text’s boundaries.

Genette (1997) conceptualizes paratexts as metaphorical thresholds constituting some ambiguous middle ground between the text and the discourse about it. In her wide-reaching foray into the nature of paratexts in translation, Batchelor (2018, 142) builds on Genette’s framework, broadly defining a paratext as any “consciously crafted threshold” with the potential to influence the reception of texts. Making slight adjustments to Rockenberger’s (2014, 262-263) typology of paratext functions, Batchelor’s (2018, 160-161) framework identifies as paratexts those elements of a text whose functions may be considered referential, self-referential, ornamental, generic, meta-communicative, informative, hermeneutical, ideological, evaluative, commercial, legal, pedagogical, instructive or operational, and personalized or interactive. Each of these types of paratexts contributes in some manner to the formation of a “guiding set of directions” for receiving the main body of the text (Genette 1997, 2). Naturally, this guiding set of directions assumes substantially different forms depending on text type. Therefore, it is the project’s aims and context which dictate the manner in which this broad definition is made workable in order to determine which textual elements are considered paratexts and subsequently removed during pre-processing to create two digital versions of each text: one with all its paratexts included and one without

(Batchelor 2018, 144). (As will be discussed later on in this chapter, paratexts are considered irrelevant to the aims of two of the project's constituent studies yet integral to the third.)

Considering the project's concern with literary prose, paratexts may be generally conceptualized in the course of this project as textual features that are *functionally distinct* from the narrative, much as the text's literal author is able to be distinguished from the narrative persona presenting a fictional account. This criterion applies not only to legal elements such as imprints and copyright information, but also elements whose primary purpose is to organize the text's structure, such as chapter or section titles, considered by Genette (1997, 3-4) to be paratexts. Other extra-narrative elements of texts include quotes introducing chapters, as exemplified in Matilde Serao's *L'anima Semplice* (1901), which introduces its opening chapter by quoting Dante's *Paradiso*.

Although Batchelor (2018, 142) stresses the importance of a function-based rather than location-based approach to classifying paratexts, the scale of this project necessitated a more superficial means of identifying paratexts to be removed. The strategy for identifying paratexts was initiated by focusing on location and typography. Textual elements constituting paratexts were thereafter identified by a quick assessment of their general function. While the distinction between the main text and its paratexts is far from certain, commonly identifiable patterns of features of paratexts appeared over the course of pre-processing.

For instance, legally oriented paratexts such as copyright information consistently appeared before or after the main body of each text, and other paratexts were easily identified by their typographical features, such as uncommon punctuation (e.g., brackets) and indentations. Such distinguishing typographical features were detected by a combination of an extensive manual review of corpus texts and systematic searches for common typographical features of paratexts using the dynamic search queries of text processing software. Using this combined manual and systematic procedure, textual elements were identified as paratexts and removed accordingly, including but not limited to:

- Copyright information
- Author forwards and translator prefaces
- Introductions situated external to the fictional plot
- Quotes introducing chapters or book sections
- Descriptions of illustrations
- Footnotes
- Endnotes

In addition to these identified paratexts, other textual elements were removed in anticipation of their interference in studies' linguistic analyses, such as frequently repeated dates and openings in works structured as a series of letters or journal entries. As described in the following section, code switches constituted a significant conundrum with respect to the project's central research question.

4.4.3. Removing code switches

The task of disentangling the competing influences of specific SLs and TLs in translation is complicated by the fact that original-language literary works frequently contain elements from other languages. Certain multilingual aspects of original texts are, rather confusingly, also referred to as interference in contact linguistics (see Mullen 2012, 19). It must be emphasized that this thesis narrowly focuses on SL influence – interference or foreignization – as a *translation-induced* phenomenon, in which SL features of target texts are directly attributable to the translation process. Therefore, the project disregards original texts' foreign-language elements and linguistic elements of translations that originate from languages other than the SL or TL. Such features are more aptly linked to language contact or multilingualism, referred to collectively as code-switching – "the use of several languages or varieties within the same text" (Gardner-Chloros and Weston 2015, 186). Code-switches may muddle the effects of dichotomous SL/TL power imbalances on the composition of translations, and their removal from texts was therefore warranted for the purposes of this project.

In practice, the identification and subsequent removal of code-switches from texts were contingent upon the ease with which they could be completed. As with paratexts, code-switches were primarily identified via typographical markers, which were most frequently stand-alone paragraphs or indentations. This practical constraint resulted in code-switches that were less obvious – particularly those shorter in length – remaining in the corpus. The study on lexical interference presented in Chapter 4 will further distinguish loanwords from code-switches, also focusing on the length (in tokens) of these related phenomena.

Common types of code-switches removed were foreign-language songs or poems produced in either original or translated texts, typically indented and/or set apart from the surrounding text in separate paragraphs. Paul Ernst's *Der Tod des Cosimo* (1912) included several instances of Spanish verses that were removed in this manner. Foreign-language dialogue was not removed, despite appearing relatively often in the corpus, as there was no means of consistently detecting these occurrences. As an example, the Croatian>Swedish translation *Återkomsten* (1963) contains many instances of foreign dialogue, as its characters frequently speak in European languages other than the novel's native Croatian or the Swedish TL. Overall, texts in the corpus may still exhibit multilingual traits, though certainly not to the extent where it is "impossible to identify the dominant language" (Mullen 2012, 16); this pre-processing step simply mitigates the confounding potential of code-switches on the project's operationalized metrics of SL influence.

4.4.4. Removing typographical markers

As noted by Brooke et al. (2015, 43), the formatting and typographical features of texts in Project Gutenberg are highly inconsistent. Formatting in print texts is represented in the repository's plain texts files using a variety of typographical distinctions, such as underscores or equal signs. Once the common patterns of typographical representations of special formatting were determined, these patterns were removed from texts using the search and replace function. Pre-digitized text files downloaded from Project Gutenberg and other online repositories frequently placed invisible paragraph markers in the middle of sentences so as to create visually-appropriate line breaks in plain texts files; these intrusive paragraph markers were removed by similar means.

4.5. Basic corpus data

One of the foremost criteria in the early stages of corpus construction is the ideal or anticipated size. While there is no agreed-upon standard for the sufficient size of corpora, corpus linguists have traditionally argued that smaller corpora – anywhere from tens of thousands to several million tokens – are adequate when properly tailored to specific purposes (Corpas Pastor and Seghiri Domínguez 2010). In practice, the size of a corpus typically reflects the ease with which the relevant texts are obtained. Over the course of building the corpus, it was determined that an attainable size for nearly all selected languages' comparable subcorpora was roughly two million tokens, as shown in Table 4.

Table 4: Comparable subcorpus sizes (in descending order of language status)

Comparable subcorpus	Texts	Tokens
English	37	3,564,008
French	32	2,899,063
German	38	2,132,158
Italian	45	2,278,424
Swedish	40	1,889,923
Croatian	30	2,041,219
Irish	55	1,188,331
TOTAL	277	15,993,126

The subcorpora for all translations into each TL were more subject to the availability of texts, and therefore much less balanced, as evident in Table 5, which presents the sizes for each subcorpus comprised of all translations into each of the project's selected languages (in descending order of status):

Table 5: Fixed TL subcorpus sizes (in descending order of language status)

Translation subcorpus (fixed TL)	Texts	Tokens
English	38	3,654,300
French	20	1,245,114
German	23	1,771,834
Italian	10	677,664
Swedish	10	1,416,385
Croatian	15	1,248,465
Irish	6	280,442
TOTAL	122	10,294,204

The full details (including metadata) for texts included in the comparable and translation subcorpora are provided in Worksheet 1.1./1.2. on the project's associated Github directory, whose top-level link is provided in Appendix A.

4.5.1. Parallel texts

The comparable corpus design described in the opening to this chapter is depicted in with specific data regarding the number of parallel texts it contains. For each TL, the percentages of translated texts constituting parallel texts – i.e., those whose source text is included in the corresponding comparable SL subcorpus – is provided in Table 6 (see also Worksheet 1.3. for further details).

Table 6: Number of parallel texts in fixed TL subcorpora (in descending order of status)

TL	Parallel texts	Total translations	Percentage
English	2	38	5.26%
French	4	20	20.00%
German	8	23	34.78%
Italian	4	10	40.00%
Swedish	3	10	30.00%
Croatian	8	15	53.33%
Irish	1	6	16.67%

The subcorpus of translations into Croatian contains the highest proportion of parallel texts, with just over half (53.33%) of its translations matching source texts in the comparable subcorpora. Five of these eight parallel texts are translations of French source texts (see Worksheet 1.3.). The English translation subcorpus contains the lowest proportion (5.26%) with just two out of 38 translations constituting parallel texts –*Royal Highness* (1916, from Thomas Mann's 1909 *Königliche Hoheit*) and *A Woman At Bay* (1908, from Sibilla Aleramo's 1906 *Una donna*). In this manner, the Croatian translation subcorpus exhibits a much higher degree of parallelism than the English translation subcorpus.

4.5.2. Language pairs

The numbers of texts and tokens in the subcorpora of translations in specific language pairs are provided in Table 7 below.

Table 7: Language pair subcorpus sizes ranked by size (tokens)

Rank	SL	TL	Tokens	Texts
1	German	English	1,068,859	9
2	French	English	1,020,965	12
3	English	French	820,785	11
4	Swedish	English	764,406	8
5	Swedish	German	680,946	9
6	Italian	English	637,707	7
7	French	Croatian	614,696	7
8	German	Italian	443,299	5
9	French	Swedish	427,539	2
10	German	Swedish	375,134	3
11	English	Swedish	351,759	1
12	English	German	328,784	4
13	French	German	328,726	5
14	Swedish	Croatian	325,891	4
15	English	Irish	258,499	5
16	English	Croatian	217,207	2
17	French	Italian	213,402	4
18	Croatian	Swedish	187,100	3
19	Italian	German	175,872	2
20	Croatian	German	153,361	2
21	Italian	French	143,639	2
22	Swedish	French	137,937	2
23	German	French	123,276	4

Rank	SL	TL	Tokens	Texts
24	Irish	English	108,293	1
25	Irish	German	104,145	1
26	German	Croatian	90,671	2
27	Italian	Swedish	74,853	1
28	Croatian	English	54,070	1
29	German	Irish	21,943	1
30	English	Italian	20,963	1
31	Croatian	French	19,477	1

The corpus does not include translations for the following language pairs:

- French>Irish
- Italian>Croatian
- Italian>Irish
- Swedish>Irish
- Croatian>Italian
- Croatian>Irish
- Irish>French
- Irish>Italian
- Irish>Swedish
- Irish>Croatian

4.5.3. Balance of authors

In the course of determining an appropriate balance of representativeness and practicality, there emerged a general guideline that texts from a single author should not amount to more than 1/5 (20%) of the total tokens in a comparable subcorpus. This guideline holds true for all comparable subcorpora except that of Irish, for which the availability of texts was far more restricted. Tables 8 and 9 show the mean number per author of texts and tokens respectively.

Table 8: Summary statistics for author texts among comparable subcorpora by language

Comparable subcorpus	Total no. texts	Total no. authors	Average texts/author	Median texts/author
English	37	18	2.056	2
French	32	19	1.684	1
German	37	15	2.467	2
Italian	45	21	2.143	2
Swedish	40	24	1.667	1
Croatian	30	14	2.143	2
Irish	58	30	1.933	1

Table 9: Summary statistics for author tokens in comparable subcorpora by language

Comparable subcorpus	Total no. tokens	Total no. authors	Average tokens/author	Median tokens/author
English	3,564,008	18	198,000.444	171,403
French	2,899,063	19	152,582.263	103,878
German	2,098,698	15	139,913.200	103,843
Italian	2,278,424	21	108,496.381	73,465
Swedish	1,889,923	24	78,746.792	59,395
Croatian	2,041,219	14	145,801.357	115,567
Irish	1,241,416	30	41,380.533	23,242

Data regarding the proportion (balance) of authors in the various subcorpora are provided separately (see Worksheet 1.4./1.5.). Aside from the inclusion of parallel texts, the repetition of authors among the comparable and translation subcorpora offers another means of approximating the direct source-target comparability enabled by the parallel corpus design.

In order to avoid the overrepresentation of any single author as mentioned here and in Section 4.2.2., the text selection process attempted to construct the translation subcorpora (by TL) in such a way that no author accounted for more than 1/5 (20%) of the subcorpus, though this objective was heavily constrained by text availability. The distributions of authors and translators in the translation (TL) subcorpora are likewise presented in Worksheet 1.5./1.6., respectively.

4.5.4. Texts' publication dates

During the text selection process, data on texts' publication dates were also collected. This information was often found in paratexts of digital text files corresponding to the book's imprint, or else determined via online encyclopedias or databases. For the sake of consistency, each text's first year of publication was always prioritized in order to account for the possibility of subsequent reprints. Each translation's precise year of (first) publication, as well as its corresponding source text's year of publication, was also recorded, although perfect accuracy is not guaranteed. Tables 10 and 11 show summary statistics for each comparable subcorpus by language.

Table 10: Summary statistics for texts in comparable subcorpora

Comparable subcorpus	Texts	Mean year of publication*	Median year of publication	St. dev.	Range	Min.	Max.
English	37	1903	1910	21.928	79	1847	1926
French	32	1893	1895.5	25.951	79	1844	1923
German	40	1908	1910	15.818	82	1843	1925
Italian	45	1899	1900	16.663	56	1866	1922
Swedish	40	1895	1903.5	24.799	83	1839	1922
Croatian	30	1895	1887	20.798	69	1871	1940
Irish	58	1913	1914	7.540	30	1895	1925

*Rounded to the nearest whole.

Based on the figures in Table 10 above, the comparable subcorpora strongly reflect the project's intention to build a corpus representing literary prose from the mid-19th to early-20th century. Notably, the comparable Irish subcorpus covers a particularly narrow range, reflecting the Gaelic League's programmatic efforts to publish and thus revitalize Irish-language literature starting at the very end of the 19th century (O'Leary 1990, 90).

On the whole, the data indicate a fairly high concentration of texts around both the median and mean year in each comparable subcorpus, with standard deviations all falling below 26 years and all means falling between 1892 and 1914. In contrast, there is much greater variation among publication dates for translations, as evidenced in Table 11 below.

Table 11: Summary statistics for publication dates for texts in fixed TL subcorpora

Translation subcorpus (fixed TL)	Texts	Mean year of publication (tr.)*	Median year of publication	St. dev.	Range	Min.	Max.
English	38	1906	1902	27.755	162	1853	2015
French	20	1909	1905	36.895	139	1854	1993
German	23	1924	1920	33.943	158	1859	2017
Italian	10	1907	1889	58.594	130	1853	1983
Swedish	10	1943	1946.5	29.598	94	1894	1988
Croatian	15	1986	1995	19.242	59	1947	2006
Irish	6	1925	1909.5	36.209	95	1902	1997

*Rounded to the nearest whole.

The wide range of publication dates in the translation subcorpora is perhaps indicative of the historical relationships between each (constellation-bound) language and its associated culture(s). In particular, Croatian and to a lesser extent Swedish translations are significantly more recent than the others, likely due to the inclusion of the more recently published translations between these two languages. Overall, the low numbers of translations for Swedish, Croatian, and Irish fitting squarely within the project's designated historical period may be caused by the combination of a skewed text selection and the simple fact that there were possibly not many translations into these languages at the time. The inclusion of texts from a wide time period gives rise to several methodological and analytical complications, as will be explored in the thesis' conclusion. (For all pertinent metadata regarding texts in the corpus, see Worksheet 1.)

4.6. Statistical approach

The causal relationship posited by Toury's proposed law of interference naturally warrants a deductive or "top-down" research design, in which the "suitability of models or theories to describe specific aspects of translation and interpreting" is hypothesized and tested directly (Mellinger and Hanson 2016, 4). Despite qualifying statements about the inevitability of confounding variables in determining SL influence in translation, the simplicity of Toury's asserted linear correlation implies that language power dynamics are generally expected to override other potential contributing factors. As briefly covered in the introduction to the thesis, literature in related disciplines such as contact linguistics and sociolinguistics strongly points toward the pre-eminence of language status as a predictor of language change. These historical trends among related disciplines may be considered "good reasons to count socio-cultural factors again among the important conditions" of the manifestation of interference in translation (Toury 2012, 311). Evaluating Toury's law of interference as a suitable theory of translation may therefore seem predisposed to a monofactorial research design.

However, cutting-edge research in corpus-based translation studies is increasingly turning toward multifactorial analyses employing advanced statistical methods, with scholars cautioning against the severe limitations of bivariate research designs (see De Sutter and Lefer 2020). In this view, the more comprehensive nature of multivariate research is more conducive to an "emerging, bottom-up translation theory" that gradually illuminates the complex, probabilistic relationships among interrelated variables instead of proposing uncomplicated laws from the outset (ibid., 2). Among the multiplicity of factors worth incorporating into corpus-based translation research, De Sutter and Lefer (ibid., 6) suggest:

- 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

Beyond the factors listed here, editorial policy and intervention are being increasingly recognized for their impacts on translation products (Kruger 2017). The current project's focus on historical literary texts renders impossible the task of confirming the vast majority of these conditions for all 122 translations. In absence of a multifactorial approach, then, the unknown variables – e.g., language attitudes or the translation policy of a given target culture – may complicate any attempt to surmise a straightforward correlation between SL status and SL influence in translated texts. Nevertheless, many of the abovementioned factors are highly interrelated, and may be implicitly – albeit imperfectly – accounted for in the project's methodology and subsequent analysis.

Chapter 2 presented a theoretical distinction between language status and language prestige, depicting the relationship between the two as well as the role of subjective language attitudes in conceptualizing the latter. The translation policy of a given target culture may likewise be categorized as a sociolinguistic condition. Instead of examining cultures' prescriptive translation practices, this project hypothesizes that comparative SL status is the primary determinant of linguistic features of translations, where other potential confounding variables are also subservient to overarching language power dynamics. Still, this project's bivariate research design is ultimately reflective of its inability to account for these potential confounding variables instead of a calculated methodological choice.

Toury (2012, 300-303) acknowledges the probabilistic nature of translation theories ("laws"), where certain variables in certain situations prove more impactful than others, and elsewhere calls for "some combination of 'top-down' and 'bottom-up' movements" (Toury 2004, 23). By testing Toury's hypothesized bivariate correlation directly, this thesis intends to promote a broader research avenue focused on language power dynamics in translation, combining the merits of both top-down and bottom-up approaches. The limitations of the project's bivariate research design will be further explored in the conclusion to the thesis, although the results may indicate which factors – if any – also appear to impact SL influence in translation. The discussion and conclusion sections of the project's constituent monofactorial studies will draw from historical circumstances to speculate on the potential effects of other factors despite the lack of a practical and systematic method of controlling potential confounding variables. Given the current lack of a systematic investigation into the effects of language power on translation, this thesis is intended to lay the groundwork for more data-rich and methodologically complex research on the effects of language status on diverse translation contexts.

It will be crucial to examine the effects of language status in a wide variety of translation contexts in order to scrutinize the validity of Toury's law of interference. As summarized in Chapter 2, research regarding the effects of language power dynamics has been limited to only a few language pairs. In the case of SL influence, there is also the issue of divergent operationalizations of the same concept, which inhibits comparisons and generalizations among different studies (Chesterman 2004, 44). In view of this widespread inconsistency, the operationalizations of the project's variables – language status and the various forms of SL influence in translation – are designed to be language-universal, applicable to any selection of languages in any translation context.

4.6.1. Operationalizing SL influence

The highly context-dependent nature of the project's explanatory and response variables – language status and SL influence in translation, respectively – calls for theoretically rigorous operationalizations of these variables. Where necessary, and as indicated in previous chapters, the thesis borrows concepts and methods from adjacent fields, such as sociolinguistics and contact linguistics.

The diverse forms that Toury's interference may take in translation requires a multifaceted approach designed to account for the possibility that interference manifests differently on different linguistic levels, as anticipated by Toury (2012, 315). To this end, the thesis operationalizes its overarching research question into three response variables, examined in separate studies reflecting distinct dimensions of translations' linguistic features: lexical, syntactic, and paratextual. In devising three typologically distinct measures of SL influence in translation, the project reflects a substantially robust operationalization of the response variable (see Mellinger and Hanson 2016, 6). The three linguistic dimensions examined in this thesis require slightly different theoretical approaches in order to establish evidence of SL influence. On the lexical and syntactic levels, SL influence is framed as interference, as the purely linguistic nature of these dimensions closely aligns with Toury's framing. On the contrary, SL influence on the paratextual level is conceived as foreignization (see Venuti 1995); the theoretical justification for this framing is further discussed in Chapter 7.

Some operationalized metrics may situate interference and normalization as symmetrical polarities on a continuous spectrum, as demonstrated in Chapter 6, where the relative frequency (RF) of translations' features may be compared against those of both comparable SL and TL subcorpora. Other indicators of interference constitute asymmetrical (i.e., positive-only) phenomena, meaning that their presence in target texts has no obvious basis for a correspondent comparison with comparable TL texts, and thus their mere absence in translated texts does not necessarily reflect normalization (see Chapters 5 and 7). What is common among the operationalizations of the response variable, however, is the requirement for measuring and comparing the *degree* of SL influence across samples of different sizes, whether translated texts or the various translation subcorpora. In translation as with other forms of language contact, empirical research most often measures degrees of SL influence in terms of the "relative over- or under-representation of linguistic features in comparison to [unmediated] monolingual language production" (Kotze 2021, 116). Following this precedent, the various forms of SL influence investigated in this thesis are operationalized using relative RFs of observable phenomena.

Given that the corpus' constituent texts and subcorpora vary in size, it is necessary to calculate operationalized metrics of SL influence in translation as RFs for the sake of comparison. However, as will be justified later on, it is necessary to explore multiple ways of grouping texts into subcorpora, each of which requires altering the basis for relativizing absolute frequencies (AFs) of operationalized metrics of SL influence. For instance, the German>Swedish translation *Huset Buddenbrook* (1922) contains 198,365 tokens. This translation may be included in two subcorpora whose necessity will be described in the following section: that of all translations from German into any TL, and that of all translations from any SL into Swedish. In the subcorpus of all translations into Swedish (1,416,385 tokens; n = 10), *Huset Buddenbrook* represents 14.00% of tokens and 10% of texts. On the contrary, in the subcorpus of all translations from German (2,105,763 tokens; n = 24), *Huset Buddenbrook* represents 9.42% of tokens and 4.17% of texts. This imbalance requires readjusting the operationalized metrics of SL influence to the size of each subcorpus in order to compare findings from the data analyses of these distinct subcorpora.

Adjusting RFs according to various subcorpora and thus examining the data from multiple perspectives provides more buffer against the potentially distorting effects of outliers, given the small sizes of many subcorpora. In this manner, the degrees of SL influence may be quantitatively compared in translated texts across the three linguistic dimensions. Chapters 5, 6, and 7 present individual studies on lexical interference, syntactic interference, and paratextual foreignization respectively, with the thesis' conclusion synthesizing the findings of these studies. Each operationalization is briefly foreshadowed:

> *Lexical interference* is conceptualized as an asymmetrical, frequency-based metric, measured here by the RF of translator-attributed loanwords. It is worth emphasizing here that the *absence* of translator-attributed loanwords does not necessarily signal lexical normalization but rather a lack of lexical interference. It is hypothesized that comparatively high-status SLs induce higher RFs of translator-attributed loanwords.
>
> *Syntactic interference* and *syntactic normalization* are conceptualized diametrically, where the inverse of syntactic interference is normalization and vice versa. These concepts are jointly operationalized in a novel metric called the syntactic interference/normalization coefficient (SINC). SINC is calculated using a formula that compares the RF distribution of part-of-speech (POS) n-grams of translated texts with those of their comparable SL and TL subcorpora. In this case, syntactic interference is conceptualized as a positive value (SINC > 0), and syntactic normalization as a negative value (SINC < 0). It is hypothesized that translations with comparatively higher-status SLs produce positive SINC values (syntactic interference), while translations with comparatively lower-status SLs produce negative SINC values (syntactic normalization).
>
> *Paratextual foreignization* is framed as an asymmetrical, frequency-based value, measured by the RF of translator-attributed footnotes. Again, the absence of translator-attributed footnotes merely represents an absence of paratextual foreignization, and not paratextual domestication. It is hypothesized that comparatively high-status SLs induce higher RFs of translator-attributed footnotes.

With SL influence in translation operationalized as continuous variables, a statistical test for rank correlation (i.e., association) is conducted for each study.

4.6.2. Test for rank correlation

Because the nature of the project's explanatory variable is such that the distance between each ranking is unknown, it is not possible to test the data for a linear correlation, i.e., to calculate a Pearson correlation coefficient (Brezina 2018, 142). Instead, the Kendall rank correlation coefficient (τ) suits the nature of the variables. As a non-parametric statistical test, Kendall's rank correlation measures the ordinal association (rank correlation): the tendency for rankings of one variable, such as SL status, to coincide with those of another, such as SL influence in translation. It is thus capable of handling ordinal variables, since it does not purport to calculate the moment-product correlation as Pearson's does (Mellinger and Hanson 2016, 190). The project calculates Kendall's tau using SPSS Statistics. A one-tailed test is conducted, given that the project hypothesizes the direction of the association between variables (ibid., 5). The study draws its scale for determining the strength of association expressed in the Kendall rank correlation coefficient from Dancey and Reidy (2020, 182-183).

The thesis uses Kendall's rank correlation test to conduct its primary data analysis, as described in the following section. Its results are further contextualized and refined by means of secondary analyses, which do not explicitly test for ordinal association.

4.6.2.1. Primary data analysis

The explanatory variable's standing as a ranked variable precludes the possibility of conducting a single test for association in order to determine the relationship between comparative SL status and SL influence simultaneously for all translations. Since comparative SL status necessarily denotes the *difference* between SL status and TL status, such a test would require that the distances between language status rankings could be expressed in absolute terms. That is, the magnitude of the SL/TL status differential characterizing translations from English into German would need to be quantified and compared to that of translations from Italian into Croatian, thus allowing translations between these two language pairs to be included in the same dataset. Instead, the classification of language status as a ranked variable means that the effects of comparative SL status on linguistic features of translations may only be assessed from the vantage point of their common SLs or TLs, whose fixed position allows for the observation of the manner in which SL influence responds to variability in status differences among all language pairs that they are involved in.

For this reason, the primary method of interpreting the data in relation to the thesis' central research hypothesis entails jointly testing two subhypotheses adapted to the operationalizations of each of the project's constituent studies. Subhypothesis I is formulated as follows:

Subhypothesis I:

> As SL status increases relative to the TL status, translations are expected to exhibit an *increasing* degree of the operationalized metric of SL influence in translation. Therefore, as the TL remains constant and the SL status increases, it is expected that there is a *positive association* between SL status and the operationalized metric.

Subhypothesis I is tested by means of a *fixed TL analysis*. In the fixed TL analysis, all translations into a given TL are grouped into a single subcorpus (fixed TL subcorpus). Levels of SL influence for individual texts in this subcorpus are visualized in a scatter plot, where the x-axis displays SLs in order of increasing status, and the y-axis displays the study's operationalized metric of SL influence expressed as a RF of the subcorpus in question. The data are subsequently tested for a hypothesized positive association using Kendall's tau. This process is repeated while holding each of the TLs in the corpus constant. Note, however, that no fixed TL analysis is performed for the fixed Irish TL subcorpus, as it does not contain a wide enough variety of translations to produce meaningful results, with five English>Irish translations and one German>Irish translation.

The positive association anticipated by subhypothesis I is the most straightforward expression of the project's hypothesized correlation between comparative SL status and the degree of SL influence in translation. However, these cross-sections of the data do not reflect the full extent of the relationship between SL status and SL influence in translation; the data must be tested for their ordinal association with all translations from each SL grouped together. Sub-hypothesis II is a formally inverse yet logically consistent expression of Toury's law of interference: as TL status increases in relation to a fixed SL, it is expected that translations exhibit comparatively lower degrees of influence from the SL, given its diminished power.

Subhypothesis II:

> As TL status increases relative to the SL status, translations are expected to exhibit a *decreasing* degree of the operationalized metric of SL influence in translation. Therefore, as the SL remains constant and the TL status increases, it is expected that there is a *negative association* between TL status and the operationalized metric.

Subhypothesis II is tested by means of a *fixed SL analysis*. In the fixed SL analysis, all translations from a given SL are grouped into a single subcorpus (fixed SL subcorpus). Levels of SL influence for individual texts in this subcorpus are visualized in a scatter plot, where the x-axis displays TLs in order of increasing status, and the y-axis displays the study's operationalized metric of SL influence expressed as a RF of the subcorpus in question – either the fixed TL or fixed SL subcorpus. The data are subsequently tested for a hypothesized negative association using Kendall's tau. This process is repeated while holding each of the SLs in the corpus constant. Note once again that no fixed SL analysis is performed for the fixed Irish SL subcorpus, as it does not contain a wide enough variety of translations to produce meaningful results, with merely one Irish>English translation and one Irish>German translation.

In order to generalize an overall tendency for translations from comparatively higher-status SLs to exhibit more lexical interference, it is necessary to confirm Sub-hypotheses I and II in tandem – i.e., to synthesize the results of the fixed TL analysis with those of the fixed SL analysis. In this manner, the thesis accounts for the language status differences of all translations in all language pairs without making any undue assumptions. Together, the fixed TL and fixed SL analyses form the primary means of testing the hypothesized correlation between SL status and SL influence in translation in the context of the lexical, syntactic, and paratextual dimensions.

4.6.2.2. Secondary data analyses

Ranking the data points according to various status-related groupings may also contextualize any potentially observed trends. These alternative perspectives further account for the potentially distorting effects of differing sizes of subcorpora, as operationalized metrics of SL influence are readjusted (re-relativized) to each relevant subcorpus:

Status pair analysis:

> As put forth in the subsequent chapter, the seven selected languages may be sorted into general status groups (SGs) – namely, high-status, medium-status, and low-status languages[11]. This manner of grouping the languages has the benefit of accounting for an operationalization of language status perceived to be overly precise, where the validity of the exact rankings of languages status may be called into question. Subcorpora formed around status pairs (SPs) contain all translations whose language pairs reflect the same combination of SGs – e.g., all translations from high-status languages into low-status languages. In this analysis, the operationalized metrics of SL influence are re-relativized according to the size (in tokens) of each SP subcorpus. SPs are subsequently ranked according to their level of SL influence.

[11] No status group is formed for the lone very low-status language (Irish), given the lack of sufficient variety of translations into and from Irish described earlier.

Language pair analysis:

> In this analysis, subcorpora are formed around each specific language pair, and operationalized metrics of SL influence are relativized according to the size of each language pair subcorpus. Language pairs are subsequently ranked according to their level of SL influence.

Ranked-text analysis:

> In this analysis, operationalized metrics of SL influence are relativized according to the size of each individual text. Texts are subsequently ranked according to their level of SL influence.

Because it is not possible to compare language status differentials between entirely discrete language pairs (i.e., those with no overlapping SL or TL, such as French>Croatian and Swedish>English), the expected results for these alternative vantage points are not formally hypothesized. Still, trends observed in these secondary analyses may contextualize or provide nuance to evidence either for or against the two subhypotheses. For example, when texts are ranked according to the operationalized SL influence metric, those with comparatively higher-status SLs are generally expected to be ranked near the top and those with comparatively lower-status SLs near the bottom, as predicted by subhypotheses I and II. Furthermore, alternative perspectives of the data may reveal unexpected trends that point toward specific confounding variables or a more complex relationship between SL status and SL influence in translation.

5. Lexical interference

5.1. Chapter introduction

In order to determine a potential positive association between SL status and SL influence in translation, it is fitting to begin on the lexical level. Lexical interference is often considered the most prominent or obvious form of interference in translation (Gómez Capuz 1997, 83; Görlach 2003, 1). The study presented in this chapter hypothesizes that translations from comparatively high-status languages into low-status languages exhibit more lexical interference than translations with an inverse power imbalance between SLs and TLs. In determining how to operationalize lexical interference in this corpus-based study, it is necessary to observe that comparisons between different TLs' tolerances for loanwords and other forms of lexical borrowings in translation have been conducted since translation studies' earliest corpus-based investigations (see Baker 2018, 36).

5.2. Lexical borrowing as translation strategy

Linguistic approaches to translation have long recognized lexical borrowings as a common and source-oriented translation strategy. Vinay and Darbelnet (1958/2000, 85) characterize a "borrowing" as an operation that entails source-language lexical items being imported into target texts in their *unaltered* forms, constituting neologisms. They claim that translation is the source of many borrowings that eventually integrate into the lexicons of recipient languages, thus translators (and, implicitly, translation scholars) are "particularly interested in the newer borrowings, even personal ones" (ibid., 85). They also identify "calques" as a "special type of borrowing" which "translates literally each of [the source word's] elements" (ibid., 85-86).

Taking a different approach, Catford (1965, 43) presents "transference" as a translation strategy in which "the TL text, or, rather, parts of the TL text, have values set up in the SL: in other words, have SL meanings." He does mention that "an SL lexical item embedded in a TL text [may seem to be] pure transference," but reasons that such items may not "fully retain [their] SL meaning" (Catford 1965, 46). For Catford, transference may take place on different levels (e.g. phonological, orthographical, or grammatical), but it primarily denotes the similarity of a lexical item's TL meaning with its corresponding SL meaning. He therefore posits dichotomous translation strategies; transference is the "implantation of SL meanings into the TL text", whereas translation is the "substitution of TL meanings for SL meanings" (ibid., 48). Here, the phenomenon of translators importing unaltered SL word forms into target texts is presented with a further dimension in which the underlying meaning of these imported terms may change. Catford refrains from labeling the translation strategy of transference – or any of its subtypes – as borrowings or loanwords. He seems to reserve the term "loanword" to refer to the phenomenon in which prior language contact has resulted in a word form from a donating language being firmly established in the lexicon of a receiving language, such as the Japanese *kimono* in English (ibid., 100).

Newmark (1988, 81) asserts that loanwords belong to Catford's idea of transference in translation. He also defines the term "naturalisation", which "succeeds transference" and constitutes an adaptation of the SL word form's pronunciation and morphology (ibid., 82). He further explains that a "through-translation" (synonymous with "calque" or a "loan translation") constitutes a literal translation of multi-word units and phrases such as compound nouns or frequently observed collocations (ibid., 85).

Baker (2018, 23) notes that the use of loanwords – by which she means foreign-language word forms preserved and unaltered in the target or receiving language – also occur in source (or original) texts, wherein the source and target languages conceptualized by translation scholars correspond to the donating and recipient languages conceptualized by sociolinguists. The appearance of loanwords in original texts poses a challenge for translators, who must consider how to render already borrowed items into yet another language. There is therefore a crucial distinction between loanwords appearing in original texts as a result of prior language contact versus those produced in translated texts and attributable to the translation process. This distinction will be revisited later on.

It is clear that the terminology on loanwords and related phenomena in translation studies literature is rather unsettled. Sociolinguistic approaches to the study of loanwords may then provide a clearer picture. Before proceeding to the project's operationalization of lexical interference in translated texts, it is worth turning to literature in sociolinguistics and contact linguistics to establish specific definitions of lexical borrowing phenomena and disentangle these closely related concepts.

5.3. Typology of lexical borrowings

While lexical borrowings may be examined as a category of translation strategies in specific contexts, these phenomena are also typified in language contact scenarios more broadly. Lexical borrowings may be examined as both catalysts and markers of language change brought about by not only translation but also any form of language contact between language communities. This broader and more historically oriented form of research on lexical borrowings, most frequently associated with contact linguistics and sociolinguistics, enjoys a much deeper history of thought and debate.

Einar Haugen's inaugural research on lexical borrowing phenomena in the mid-twentieth century has served as the core framework for subsequent developments (Hoffer 2002, 5; Serigos 2017, 7). Haugen (1950, 212-213) defines a borrowing as "the attempted reproduction in one language of patterns previously found in another." Although the terms "borrowing" and "loanword" are sometimes difficult to interpret in the literature (Van Poucke 2011, 103), the "borrowings" category tends to serve as the overarching term for more specific phenomena of cross-lingual influence such as "loanwords" (Haugen 1950, 212; Haspelmath 2009, 38). This section primarily adopts Gómez Capuz's (1997, 87-89) typology of lexical borrowings, highlighting corresponding alternative terminology and definitions for the described phenomena provided by Haugen (1950). These phenomena – loan translations (or calques), loanblends (or hybrids), and importations (or loanwords) – are presented in ascending order of the resultant lexical item's similarity to the word form in the donating (source) language. Here, the similarity of each type of lexical borrowing to the corresponding lexical item in the donating language is categorized according to its level of "morphemic substitution" –

whether the donating language’s morphemes have been fully, partially, or not at all replaced by morphemes from the receiving language (Haugen 1950, 230). The inverse of morphemic substitution is “morphemic importation” (ibid., 212), where the donating language’s morphemes are left unchanged in the receiving language.

5.3.1. Loan translations or calques

“Loan translations” or “calques” – the latter taken from French – constitute a borrowing phenomenon in which all of the individual morphemes of the (polymorphemic) lexical item from the donating language are replaced with morphemes in the receiving (target) language (Gómez Capuz 1997, 88). One example of a loan translation is the Spanish term *auto-defensa*, derived from “self-defense” in English (ibid., 89), in which “self” is replaced by *auto*, and “defense” is replaced by *defensa*. Haugen (1950, 230) refers to this piece-by-piece replacement as a “complete morphemic substitution.” This phenomenon occurs not only as a possible symptom of sustained language contact over time but also as a conscious translation strategy. Vinay and Darbelnet (1958/2000, 85) describe this translation technique as a “special kind of borrowing whereby a language borrows an expression form of another, but then translates literally each of its elements.” Of the three borrowing types examined in this section, loan translations have the weakest source-language influence.

5.3.2. Loanblends or hybrids

“Loanblends” or “hybrids” blend both morphemic substitution and morphemic importation (Haugen 1950, 215). The Spanish word *boxeo* is a loanblend adapted from “boxing” in English (Pratt 1980, 157-158 cited in Gómez Capuz 1997, 88); here, the root “box” is imported directly, while the suffix “-ing” is replaced by the Spanish morpheme -*eo*. Loanblends are not as common as other lexical borrowing phenomena stemming from language contact (Haspelmath 2009, 39), and may occur even less frequently in translation. Given their mixture of receiving-language morphemes with donating-language morphemes, loanblends have a slightly stronger influence from the source language than loan translations.

5.3.3. Importations or loanwords

Gómez Capuz (1997, 87) defines an importation as a “direct transference of a lexeme, that is, both meaning and form.” However, it is anticipated that this term may not be as widely recognizable as an alternative to “loanword”. Haugen (1950, 213-214) uses the label “loanword” to denote a lexical borrowing that employs full morphemic substitution, yet concedes that the term is sometimes applied to similar phenomena. Conversely, for Haspelmath (2009, 36), loanwords cover a broader range of related phenomena which “at some point in the history of a language entered [a recipient language’s] lexicon as a result of borrowing.” His definition includes phonological, orthographic, morphological, and/or syntactic adaptations of donor-language words (ibid., 42). Further back in history, yet another distinction emerges. German linguists in the 19th century made a “practical (but theoretically fuzzy) distinction” between *Lehnwörter* (‘loanwords’) and *Fremdwörter* (‘foreign words’); the former refers to recipient-language adaptations (such as the Spanish word *cóctel* adapted from the English word “cocktail”) while the latter constitutes exact reproductions (or full morphemic importation) (Gómez Capuz 1997, 87).

5.3.4. Code switches

Code switches are excluded from Haugen’s initial 1950 taxonomy of lexical borrowings as well as that of Gómez Capuz. Still, it is necessary to discuss this phenomenon, as it has been a central focus in sociolinguistics and contact linguistics, and will need to be dealt with in the present study’s methodology. The distinction between loanwords and code switches is a well-trodden yet ongoing debate among scholars (see Backus 2015, 25-26), with research since Haugen moving beyond the narrowly formal or structural categorizations of lexical borrowing phenomena by emphasizing more diachronic, usage-based approaches.

As Poplack and Dion (2012, 279) note, much scholarship on language contact conceptualizes code switches as lexical items first incorporated from donor languages in their unaltered form, which typically then "gradually assume more and more characteristics of the recipient language" until their full incorporation into the recipient language's lexicon, at which point they are considered loanwords. Contrary to Haugen, in this framing loanwords are those borrowed forms that adopt characteristics (whether phonological or morphological, etc.) of the receiving language. Still, other scholars contend that single-word units are exceedingly difficult to distinguish as code switches or loanwords without the proper resources for a diachronic comparison of this gradual process of adaptation (ibid., 279)

According to Backus (2015, 23-24), sociolinguistic research on code-switches tends to focus on "alternational code-switching, in which a speaker alternates larger chunks, usually full clauses or at least stretches of language that represent some degree of grammatical construction." While hardly reflective of the depth or variety of perspectives of code switches in contact linguistics and related disciplines, this tendency for code switches to be viewed as "full clauses or at least stretches of language" may benefit the present study's conceptualization of code switches, particularly when devising the methodology. In fact, much of the terminological and conceptual instability surrounding code switching in the subdisciplines of linguistics may be simply attributed to differences in methodology (Gardner-Chloros 2009, 7-8). From a technical standpoint, contact linguists seeking to identify these phenomena employ different automated techniques for (non-translated) texts containing isolated loanwords ("sporadic foreign word insertions") and extended, multi-word code switches (Serigos 2017, 52). In consideration of the present study's aims and methodological approach, the category of code switches is reserved for extended passages ("larger chunks") of foreign-language usage rather than single- or few-word lexical items.

5.3.5. Operationalizing lexical interference in translation

As with translation studies, it is evident that the term “loanword” has been used at various times in contact linguistics literature to refer to related but distinct phenomena. Still, the concept underlying Gómez Capuz’s importation and Haugen’s loanword is clearly distinguishable from the other three borrowing phenomena detailed in this section, and constitutes the most striking marker of language contact. The full and unaltered morphemic importation of a donor-language term into a receiving language’s lexicon is paralleled in translated texts by the reproduction of an SL term in its unaltered form. Because this type of borrowing constitutes the strongest possible SL interference in translation, and because these unaltered word forms may be readily cross-checked across corpora, the present study takes loanwords as its metric of lexical interference in translation. Previous translation research has operationalized lexical interference in the same manner, and has demonstrated the necessity of methodologies that identify loanwords attributable to translator decisions in target texts’ production processes and exclude those originating from prior language contact.

5.4. Related work

The current project follows numerous other corpus-based studies in taking loanwords as clear evidence of interference at the lexical level. These previous works primarily base their methods of loanword identification on the corpora available and the researchers’ linguistic expertise. Methods of identifying loanwords among the previous literature are therefore varied and highly context-dependent.

Winters’ (2004) study examines two German translations of the same English source text. She adopts Görlach’s (2003) typology of loanwords in German, which includes subcategories such as gallicisms (words loaned from French), and anglicisms (words loaned from English). According to Görlach (ibid., 1), gallicisms and anglicisms are those words that retain some element(s) – whether spelling, pronunciation, morphology, or some combination thereof – of the donating language even as they are established in the German lexicon. Winters (2004, 251) uses WordSmith Tools’ KeyWord function to identify loanwords. In corpus linguistics, (positive) keywords are those words which appear more frequently than expected in a target text or corpus relative to their frequency in a selected reference text or corpus. Winters' (2004, 253) study determines which translation contains more interference by generating a list of keywords for one by using the other as a reference text, then determining how many loanwords appear as keywords (thus highlighting them as more frequent relative to the other translation). The study determines that one contains fewer loanwords and code switches – both indicators of source-language interference – than the other.

Delaere and De Sutter (2017) present a corpus-based study involving English loanwords in translated and non-translated texts in Belgian Dutch. They adopt the approach of Zenner (2013), referring to a Dutch dictionary to determine the status of English loanwords based on words' pronunciations and etymological roots. In this paradigm, words borrowed from English into Dutch are not necessarily recognized as loanwords. Loanwords are identified based on a profile-based correspondence analysis that notes when SL lexical items are used in place of possible TL alternatives. Pronunciation is also taken into account (Delaere and De Sutter 2017, 85). When the Dutch pronunciation of a loanword candidate exactly matches the word's pronunciation in English, it is considered an English loanword; when the word's pronunciation in Dutch is slightly different from its pronunciation in English, it is considered an endogenous word (ibid., 86).

Van Poucke (2011) identifies loanwords in a corpus of 20 Dutch translations of Russian literature, published at various points over the course of the late 20th and early 21st centuries. His study takes loanwords to mean SL words whose unaltered forms are reproduced exactly in the target text, constituting evidence of source-language influence in translation (ibid., 3). Van Poucke (2011, 108-109) explicitly excludes related lexical borrowings such as calques and loanblends, toponyms, proper nouns, as well as loanwords that "have been entirely assimilated into Dutch and are no longer recognised as loanwords by the audience" (i.e. those stemming from prior language contact), and loanwords from languages other than Russian. It is unclear how the study identifies these various lexical borrowings. It may be the case that the researcher applied his personal expertise in the language pair to manually search the Dutch translations for recognizably Russian loanwords, since loanword totals are presented per 100 pages of translated text (Van Poucke 2011, 111). Like the current study, Van Poucke (ibid., 106) hypothesizes that translations from SLs with a comparatively higher status or prestige tend to exhibit a higher frequency of loanwords. He finds a general decrease in the number of loanwords used over time, attributing this trend to a decline in perceived relative status of Russian (ibid., 118).

Another study by Van Poucke (2012) examines three Dostoyevsky novels in translation – two translated into Dutch and one into English – as a proof of concept for a novel method of operationalizing foreignization and domestication (Venuti 1995). This examination includes the manual identification of loanwords and related forms of borrowing (Van Poucke 2012, 8). The study conducts a "detailed microstructural comparison" of limited samples taken from the source text and the target texts, generalizing translators' overall tendencies toward foreignization or domestication (i.e. interference or normalization) based on these limited excerpts (ibid., 12-13). This limited study is simply intended to exemplify a quantitative method for gauging domestication and foreignization in small corpora.

Frankenberg-Garcia (2005, 2) compares loanwords across original and translated fiction in English and Portuguese, where English is taken as the higher-status language. In her study, the loanwords identified in both original and translated texts include those borrowed from *any* language, not just the translated texts' source languages (ibid., 7). Frankenberg-Garcia (ibid., 2) similarly takes lexical borrowings as textual evidence of Toury's interference. She uses a corpus whose foreign words have been previously tagged as such by the texts' author or translator, making loanwords straightforward to identify for other researchers (ibid., 6). Her study concludes that translated texts tend to contain more loanwords than source texts, and that the "relative status of the source-text language and culture" plays a determining role, regardless of loanwords' linguistic origins (ibid., 19).

Bernardini and Ferraresi (2011) identify and compare anglicisms in original and translated Italian computing texts, referring to the translations' English source texts. The authors adopt Gottlieb's (2004, 44-47) typology of anglicisms, including those both unaltered and adapted to the conventions of Italian morphology. In order to identify unaltered anglicisms, they use Log-Likelihood to identify word forms that are significantly more frequent in either Italian subcorpus (translated or original texts) when compared to the other, scanning the resultant lists for “English-looking” word forms (Bernardini and Ferraresi 2011, 233-234). They find that Italian translations of English source texts tend to use fewer unaltered anglicisms than comparable original Italian texts (ibid., 241).

Previous works such as that by Frankenberg-Garcia (2005) capture loanwords that are not only borrowed from a target text’s source text in the course of translation, but also those which result from prior language contact and are thus independent of the local translation process. Researchers using this broader definition therefore include loanwords donated from *any* language, also identifying these occurrences in source or original texts. The overall aim of the current project is to quantify *translation-induced* interference as a variable that responds to differentials in the statuses of translations’ source and target languages. The characterization of loanwords put forth in this chapter is therefore more intentional and targeted with respect to translation theory. Following Van Poucke (2011, 109), the present methodology necessarily excludes loanwords attributable to prior language contact, as this phenomenon is external to the local translation process. The current study aims to capture a narrower form of translation-induced interference.

Some studies also consider the manner in which assumed imbalances in language status or prestige might influence the frequency of loanwords in translated texts. Frankenberg-Garcia (2005, 2) and Van Poucke (2011, 106) assume language power relations to explain observed differences between language pairs in the frequency of loanwords in translation; however, both studies involve only one language pair and apply an intuitive and thus non-transferrable categorization of the status differential between the source and target languages.

The current project aims to put forth a method of loanword identification that both 1) includes only those loanwords that are most likely attributable to translation-induced interference, and 2) is replicable in future studies using comparable corpus methodology, regardless of the language pair. Preliminary tests conducted for the present study applied a similar methodology to Frankenberg-Garcia (2005) and Bernardini and Ferraresi (2011), but keyword lists appeared to be unreliable in capturing loanwords. The failure of this approach is perhaps due to the stipulation that keywords are those words that most dramatically overshoot their expected frequencies, discounting loanwords that occur only once or very infrequently. Moreover, the "very labor-intensive and time-consuming" method used by Van Poucke (2012, 13) is not feasible for the current project's aim of quantifying lexical interference across the full multilingual corpus. Given the inadequacy of these previous methods in the present context, a novel method of identifying loanwords using monolingual comparable corpus methodology is designed.

5.5. Methodology

In translated literature, lexical borrowings are sometimes distinguished typographically, such as with italics or quotation marks. However, the variety of texts included in the current study's multilingual corpus prevents the typological consistency that may otherwise make loanwords uniformly identifiable. Thus, in order to identify suitable loanword candidates, the present study makes use of frequency statistics, which is perhaps the most elemental application of corpus methodology.

The methodology section is structured as follows. First, it defines a specific type of loanword constituting translation-induced lexical interference, then outlines the process of identifying these lexical items in the corpus. It subsequently reiterates the project's explanatory variable and describes the method used to draw comparisons between the frequencies of loanwords across the various texts and subcorpora. Finally, the section justifies the use of its selected statistical analysis.

5.5.1. Defining translational loanwords (TLWs)

This project measures instances of (full morphemic) importation of source terms into target texts as an indicator of lexical interference in translation, as this phenomenon constitutes the strongest possible source-language influence out of the three lexical borrowing phenomena examined previously.

As already indicated, there is an important distinction to be made between loanwords firmly implanted in the receiving language's lexicon and those retaining their perceived “foreignness”. The former may be the result of historical language contact, whereas the latter constitutes a more recent interlinguistic influence that may be more attributable to recent translation activity (Vinay and Darbelnet 1958/2000, 85). Although contact linguistics literature likewise refers to donating-language influence as interference (see Haugen 1950, 223; Gómez Capuz 1997, 81; Haspelmath 2009, 36), Toury’s (2012, 314) law of interference posits interference specifically as a *translation-induced* phenomenon attributable to SL influence exerted in the local translation process. Therefore, this project seeks to identify only those loanwords that are suspected to reflect interference from target texts’ respective source texts instead of from (distantly) prior language contact.

Adopting the work of Humbley (1974), Gómez Capuz (1997, 83) asserts that “‘lexical borrowing’ (meaning and form) makes up the ‘core’ of interlinguistic phenomena, the other categories being peripheral by nature.” A lexical borrowing is “the transference of a whole lexical unit, meaning and form” (ibid., 83-84).

Therefore, this study introduces the term *translational loanwords* (TLWs) to specify those loanwords that are apparently borrowed over the course of the target text’s local translation process instead of via prior language contact. Unlike TLWs, loanwords attributable to prior language contact become well-established in the target lexicon. Due to methodological and practical constraints, the determination of which process a loanword candidate is more likely attributable to relies on assumptions rather than certainties. Because the project’s use of comparable corpus methodology means that the respective source texts of texts comprising the translation corpora are not necessarily all available, all markers of source-language interference are necessarily inferential. That is, these markers are determined primarily by comparing translated texts to the comparable source-language and target-language corpora, and secondarily by referring to other language resources.

SL word forms other than TLWs in the translation subcorpora may not result from prior language contact, but still constitute a type of translation interference, albeit one that is categorically different from TLWs. The previous discussion of the lexical borrowing taxonomy included a brief foray into the theoretical difficulties of distinguishing loanwords from code switches in contact linguistics. These complications have little bearing on translation studies research. Winters (2004, 249) distinguishes code switches from loanwords in translated texts, defining the former as “a *superordinate* [emphasis added] category comprising words, proper names, phrases and quotations” in the source language that are assumed to be somewhat understandable to the target audience. Here, Winters introduces an overtly psycholinguistic presupposition, stipulating that code switches are generally understandable to a target audience. In order to distinguish TLWs from code switches in the context of the present study, it is more straightforward to resort to a conceptual framework that readily identifies code switches by their length in words.

For instance, a translator may choose to leave a 20-line poem appearing in the source text in its original language in their translated text. Rather than counting each word in the poem as an individual TLW and thereby skewing results (as the translator did not choose to borrow each individual word but rather the poem as an entire unit), these instances are excluded from the analysis of TLWs. The loanword typologies described earlier generally conceptualize loanwords as individual words, although short multi-word units such as *déjà vu* may be considered loanwords as well (see, for example, Vinay and Darbelnet 1958/2000, 85; Newmark 1988, 81; Gómez Capuz 1997, 87; Chesterman 2016, 92). The literature is unclear on the acceptable length (in number of words) of a loanword as a single unit; multi-word phrases exceeding this theoretical word limit may be distinguished as code-switches, constituting quotes instead of singular lexical units. For the purposes of this study, it must therefore be determined what the cut-off point is for a multi-word TLW as a single unit; units comprised of source-language words that exceed this delineated boundary may be referred to as “code switches”.

What then is the maximum length of a TLW? TLW candidates that are deemed “too long” may be more akin to a quotation, and therefore a code switch. In Winters’ (2004, 255-256) study, code switches are categorically different from loanwords, and are thus calculated and analyzed separately. Following this precedent, the present study views code switches as a related yet distinct phenomenon from loanwords in translation. The methodology for identifying TLWs in the translation corpora thus deliberately excludes all untranslated source-language words comprising e.g. songs, poems, or character-specific honorifics. The current study therefore makes the judgment that a TLW is comprised of five words (5-grams) or fewer; the review of TLW candidates revealed a relatively naturally-occurring boundary between contiguous multiword TLWs up to this range and much lengthier SL passages (i.e., code switches).

As in Winters' (2004) study, proper nouns are considered to be code switches, or merely categorically different from TLWs. The present study also includes honorifics and currencies among code switches. Honorifics used for specific people may be viewed as extensions of proper nouns. However, honorifics that are not linked to proper names or specific characters and that may otherwise be replaced by a target-language lexical item constitute markers of interference, and are therefore counted as TLWs. Additionally, foreign currencies are assumed to be generally established in the lexicons of other languages due to international trade. Character-linked honorifics and currencies are judged to be code switches, and are therefore not counted as TLWs.

Unifying the criteria laid out earlier, the present study defines a TLW as the following:

> A **translational loanword** (TLW) is an unaltered word form or collocation (up to five grams) in a translated text that is suspected to be borrowed from the text's respective source language as a direct result of the relevant text's translation process.

The following lexical items are discounted as TLWs:

- Loanwords already established in the target language's lexicon
- Orthographically adapted lexical borrowings
- Loanwords originating from a language other than the text's source language
- Proper nouns (e.g. places, characters)
- Honorifics, titles
- Currencies

Having operationalized and clearly delineated what constitutes textual evidence of lexical interference, it is now necessary to define the study's process for identifying TLWs among the subcorpora.

5.5.2. Identifying TLWs

The process of identifying TLWs among the word frequency lists of the various language pairs' subcorpora involved three relevant subcorpora: the translation corpus for the language pair in question ($C^{SS>TT}$, where SS and TT represent the two-letter language codes of the source and target languages), the source reference corpus (C^{SS}), and the target reference corpus (C^{TT}). The following process was repeated for each language pair appearing in the corpus. First, a word frequency list for the translation corpus was generated using WordSmith 8.0. This word frequency list served as the preliminary pool of translational loanword candidates; before any other information was determined, the project needed to assume that any word in the word frequency list could be a TLW. Next, word frequency lists were generated for the corresponding source and target reference corpora. Then, the three word frequency lists were incorporated into a single spreadsheet, which included each word's absolute frequency (AF) and relative frequency (RF) in its respective corpus.

Given the unmanageable size of the word frequency list for each translation corpus, there needed to be a method of bringing into focus word forms that were most likely to be TLWs. The project thus devised two types of TLWs with unique identification procedures, ordered according to which procedures are judged most likely to capture strong TLW candidates.

5.5.2.1. Type I TLWs

Using these data, the project makes two key assumptions about the AFs of (Type I) TLWs across the subcorpora:

1. TLWs do not appear in the target reference corpus ($AF^{TT} = 0$).

2. The AF of TLWs in the source reference corpus is greater than their AF in the SL>TL translation corpus ($AF^{SS} > AF^{SS>TT}$).

The first assumption reflects the fact that, as previous linguists have pointed out, loanwords are not established in the target language. If a word is borrowed in the process of translation, it would therefore not be expected to be found in texts originally written in the target language. Conversely, if a source-language word happened to be established in the target language due to prior language contact, it may be expected to appear in texts originally written in the target language.

The second assumption reflects the fact that a source-language word is expected to occur more frequently as a word in the SL lexicon than as a "newer borrowing" in translated texts (Vinay and Darbelnet 1958/2000, 85). This expectation is formed because a word occurring in the course of a language's regular usage is available to any author, while translators may only sometimes choose to borrow a word if triggered by its appearance in their particular source text – an event they have no control over. A word borrowed from a source language may thus be expected to occur more frequently in texts originally written in the source language than in texts translated from the source language.

The experiment used spreadsheet formulae to generate a list of all word forms satisfying these two criteria, which constituted a list of Type I TLW candidates. Concordances were then reviewed for each candidate in the list, in order to determine whether evidence in the surrounding text could assist in determining the candidate's status as a TLW. It bears repeating that the current study identifies loanwords that are likely attributable to the process of translation. This means that loanwords that have already been established in the receiving (target) language's lexicon are not relevant. In order to determine whether loanwords had been already firmly established in the target language's lexicon, TLW candidates were searched in online dictionaries for each target language[12]. If a word form was found in a target-language dictionary, and already had a dictionary entry corresponding to its use in the concordance, the TLW candidate was determined *not* to be a TLW. That is, the appearance of a (non-translational) loanword in the comparable monolingual corpus or dictionary of the target/receiving language is

[12] English: Cambridge Dictionary (https://dictionary.cambridge.org/); Le Dictionnaire (https://www.le-dictionnaire.com/), The Free Dictionary (https://fr.thefreedictionary.com/); German: The Free Dictionary (https://de.thefreedictionary.com/); Italian The Free Dictionary (https://it.thefreedictionary.com/); Swedish: Ordlista.se (https://www.ordlista.se/ordbok/); Croatian: Hrvatski jezični portal (https://hjp.znanje.hr/); Irish: Foclóir Mháirtín Uí Chadhain (https://focloiruichadhain.ria.ie/), An Foclóir Beag (https://www.teanglann.ie/en/fb/)

taken as evidence of its assimilation into that language's regular lexicon, and therefore rules it out as a TLW. If a word form was not found in a target-language dictionary, it was somewhat likely to be a TLW. In some cases, the word form was searched in a SL dictionary, in order to confirm that it likely originated from the source text and/or not from e.g., a language other than the target text's SL. Van Poucke (2011, 107) raises this method as a possibility for translators to "check the level of assimilation of particular words in the target culture." Sometimes whether or not the given word form would appear in a given language's dictionary seemed obvious enough to bypass this process, based on personal knowledge of the language(s) in question. Of course, this method involved a significant amount of subjectivity on the part of the researcher. (Additional methodological shortcomings regarding the use of these online dictionaries and the reliance on subjective judgments will be emphasized later on in the chapter.) For each candidate judged to be a Type I TLW, its AF in the translation corpus was noted.

5.5.2.2. Type II TLWs

The method for determining Type I TLWs may unintentionally exclude other strong loanword candidates, particularly given its first assumption; a SL reference corpus and a TL reference corpus may contain identical word forms as a result of pure coincidence rather than lexical borrowing stemming from prior language contact or the text's translation process. Word forms that are orthographically identical yet semantically distinct across languages are referred to as interlingual homonyms. For instance, an interlingual homonym between English and Spanish is the word form "sin". In English, the term refers to a transgression; in Spanish, the term simply means "without". The possibility of interlingual homonyms occurring between SLs and TLs – especially those that are linguistically related such as French and Italian – interferes with the assumptions laid out previously, since TLWs may also happen to be interlingual homonyms.

In order to correct this methodological deficiency, a separate process for identifying TLWs among interlingual homonyms (Type II TLWs) was devised and conducted. The process for determining Type II TLWs was identical to that of the Type I TLWs, except that it generated a list of TLW candidates by identifying word forms appearing in all three of the relevant corpora (AF^{SS}, AF^{TT}, $AF^{SS>TT} > 0$). The Type II TLW candidate list thus covered word forms in the source reference corpus ($AF^{SS} > 0$) that were borrowed in translation ($AF^{SS>TT} > 0$) and happened to have an interlingual homonym in the target reference corpus ($AF^{TT} > 0$). False positives were consequently filtered out over the course of the manual concordance review. (See Worksheet 2. for an example of the TLW identification process put to practice for a given language pair.)

5.5.3. Calculating TLW RF for various subcorpora

Lexical interference is operationalized as the combined RF of all tokens identified as TLWs in the relevant subcorpus. In order to analyze the data from different vantage points, TLW RFs are calculated relative to subcorpora formed around the translated texts' various possible metadata (pertaining to language status): fixed SL, fixed TL, and status pair (SP).

For each subcorpus, the TLW RF is calculated as the AF (tokens) of TLWs divided by the total number of tokens in the subcorpus. Finally, TLW RFs are calculated relative to each individual text's total number of tokens.

5.5.4. Hypothesis testing

The two complementary subhypotheses constituting the project's primary data analysis (see Section 4.6.2.) are adapted to this study as follows:

Subhypothesis I:

> As SL status increases relative to the TL status, translations are expected to exhibit an *increasing* degree of lexical interference. Therefore, as the TL remains constant and the SL status increases, it is expected that there is a *positive association* between SL status and TLW RF.

Subhypothesis II:

> As TL status increases relative to the SL status, translations are expected to exhibit a *decreasing* degree of lexical interference. Therefore, as the SL remains constant and the TL status increases, it is expected that there is a *negative association* between TL status and TLW RF.

The results of this study are presented in the following section.

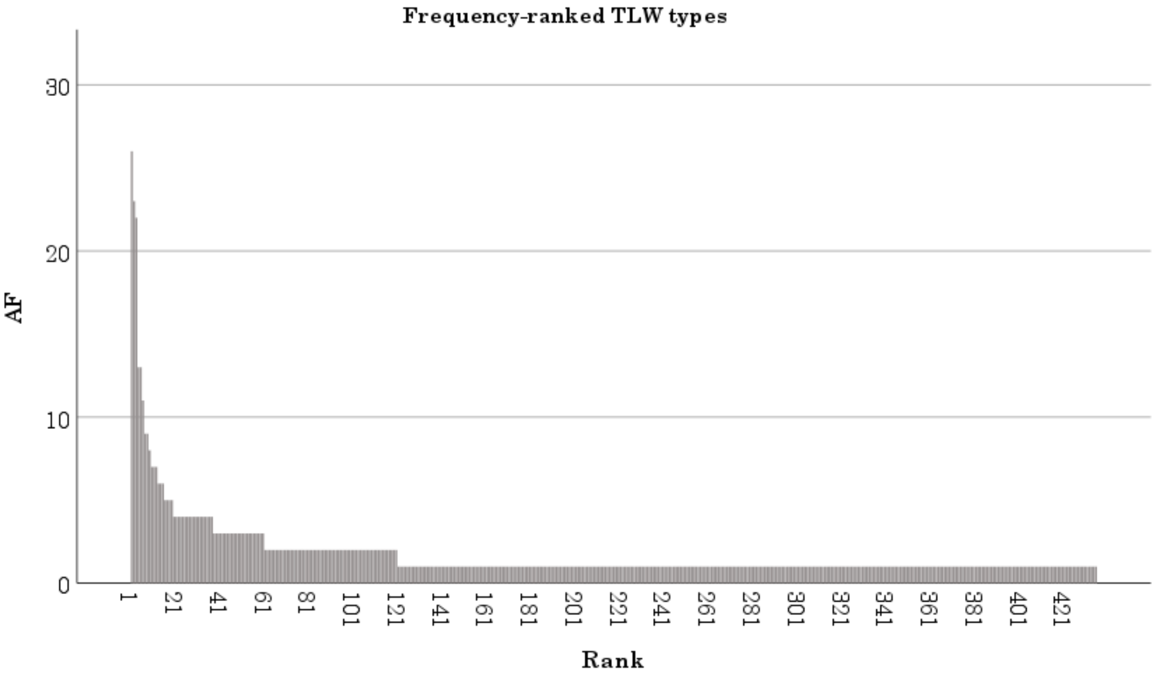
5.6. Results

Firstly, a general overview of the data is provided by way of basic summary statistics. The study then conducts the fixed TL and fixed SL analyses in order to test the two subhypotheses. Lastly, TLW RFs are presented as they correspond to each status pair (SP), language pair, and individual text, relativized to each subcorpus in question. All RFs presented are normalized per 100,000 tokens.

5.6.1. Summary statistics

In total, 769 TLWs (TLW tokens) and 440 unique TLWs (TLW types) were identified in the multilingual corpus. The frequency list of all TLWs – along with their corresponding texts, SLs, and TLs – is provided separately (see Worksheet 3.1.). Figure 1 below presents a ranked frequency histogram of all TLWs identified across the entire multilingual corpus:

Figure 1: TLWs ranked by frequency



Ranking the frequencies of the identified TLWs in this manner, it is apparent that the shape of the TLW frequency distribution somewhat resembles that of Zipf's Law. Of the 440 unique TLWs, 315 (71.59%) occur exactly once (TLW AF = 1), and 425 (96.59%) occur five or fewer times (AF ≤ 5). Only six unique TLWs (1.36%) – dispersed across six different texts and five language pairs – occur ten or more times (AF ≥ 10), meaning that 434 unique TLWs (98.64%) occur fewer than 10 times in the corpus. The AFs of the top six unique TLWs are 26, 23, 22, 13, 13, and 11. While there is clearly a heavy skew to the data, to the point where the most fitting measures of central tendency are equal to the minimum value (median = 1; mode = 1), there are no glaring outliers among the AFs of unique TLWs. Measurements of TLW RFs will, however, be relativized to each subcorpus, allowing for more variation according to the subcorpus sizes and text lengths.

5.6.2. Fixed TL and fixed SL analyses

Tables 12 and 13 present the results of the fixed TL and fixed SL analyses, displaying Kendall's tau value, p-value, and population size for each fixed TL or fixed SL analysis as the variable language increases in status. If detected, statistically significant findings (p ≤ .05) confirming the expected association between the study's two variables are highlighted in gray, and statistically significant findings contradicting the expected association are italicized.

Table 12: Fixed TL analysis for lexical interference

TL (fixed)	n (texts)	Kendall's tau* (increasing SL status)	p value
English	38	.298	.010
French	20	.550	.001
German	23	.375	.014
Italian	10	.220	.231
Swedish	10	.713	.005
Croatian	15	.469	.019
Irish	6	–	–

*Hypothesized positive association.

Table 13: Fixed SL analysis for lexical interference

SL (fixed)	n (texts)	Kendall’s tau** (increasing TL status)	p value
English	24	.161	.163
French	30	*.542*	< *.001*
German	24	*.368*	*.016*
Italian	12	*.530*	*.015*
Swedish	23	.204	.141
Croatian	7	.146	.345
Irish	2	–	–

**Hypothesized negative association.

5.6.3. Scatter plots for fixed target languages

In the scatter plots in Figures 2 to 8, each point represents a text in a subcorpus composed of all translations into the specified TL. For each text, TLW RF is calculated as the frequency of TLWs relative to the total number of tokens in the subcorpus consisting of all translations into the fixed TL. SLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish TL subcorpus, its scatter plot is included here.

Figure 2: TLW RFs for all translations into English

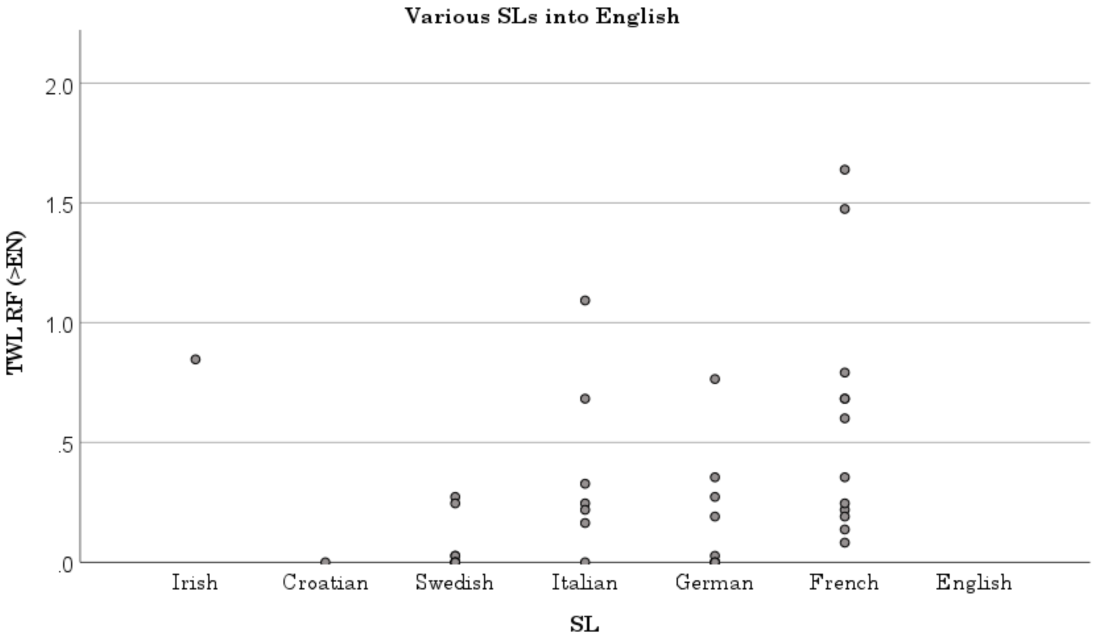


Figure 3: TLW RFs for all translations into French

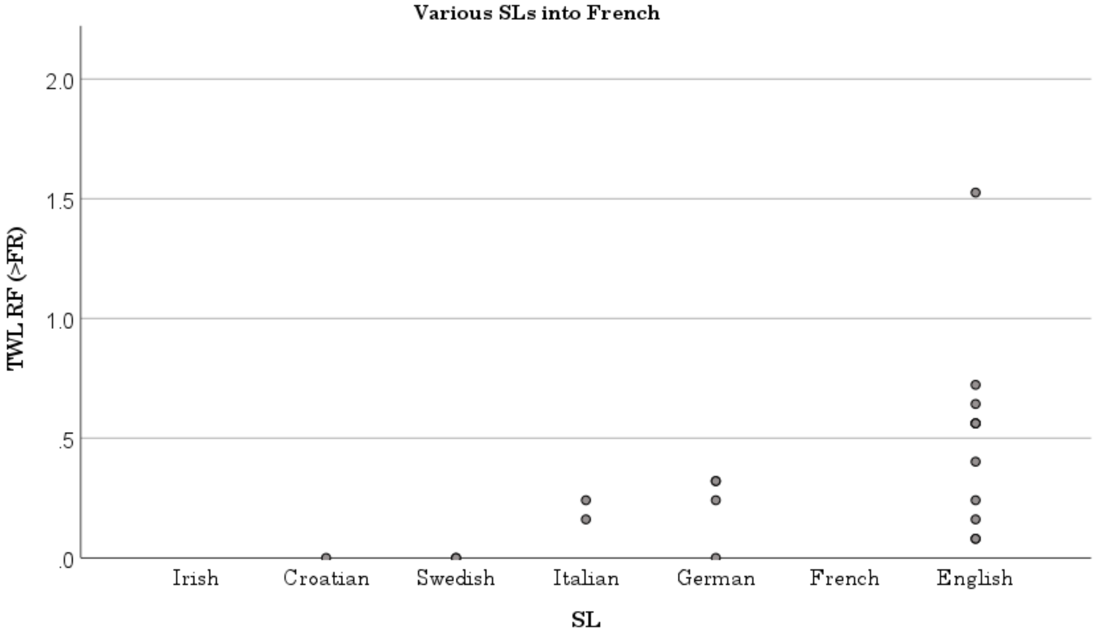


Figure 4: TLW RFs for all translations into German

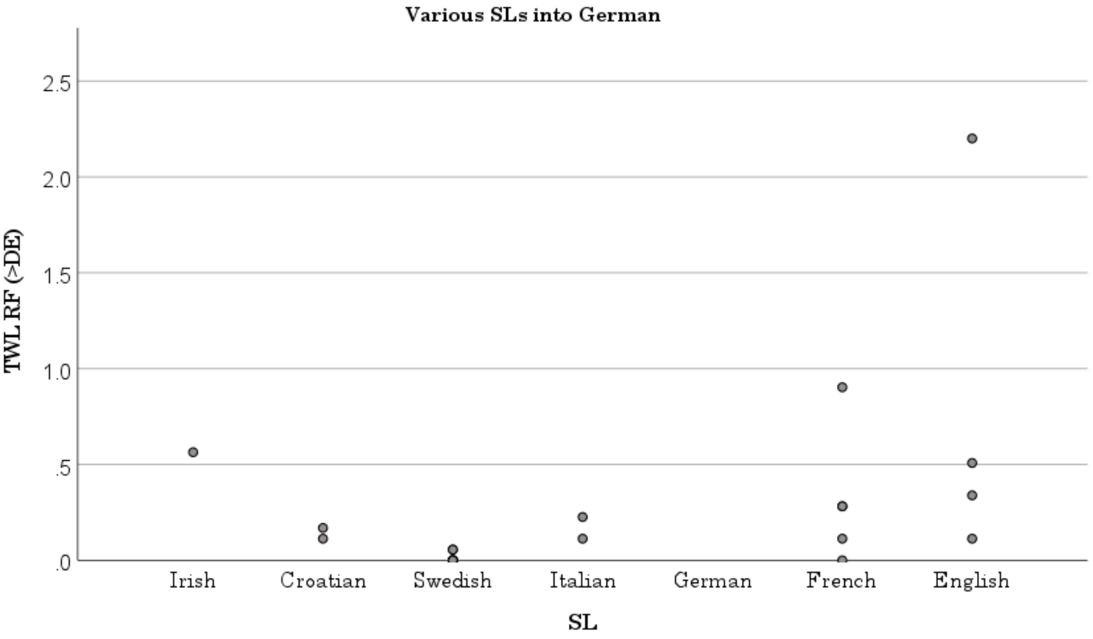


Figure 5: TLW RFs for all translations into Italian

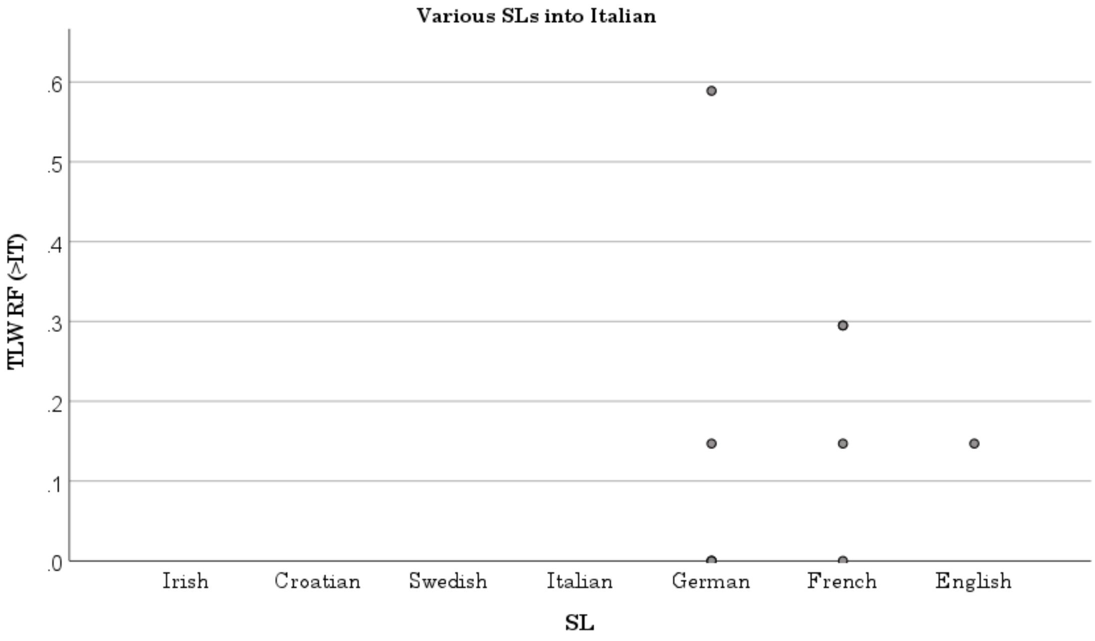


Figure 6: TLW RFs for all translations into Swedish

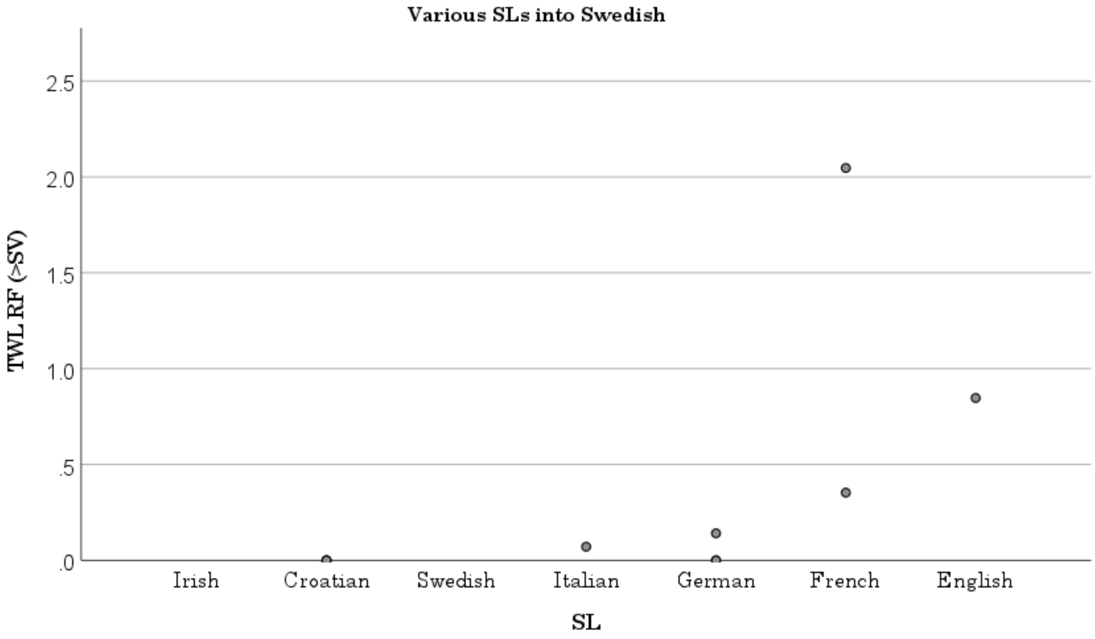


Figure 7: TLW RFs for all translations into Croatian

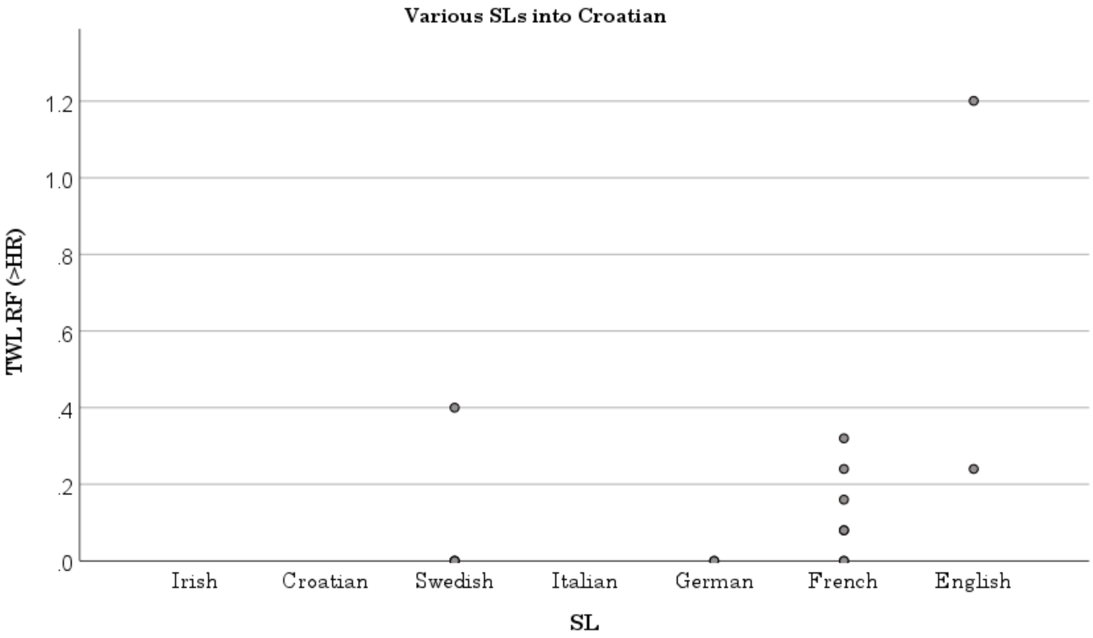
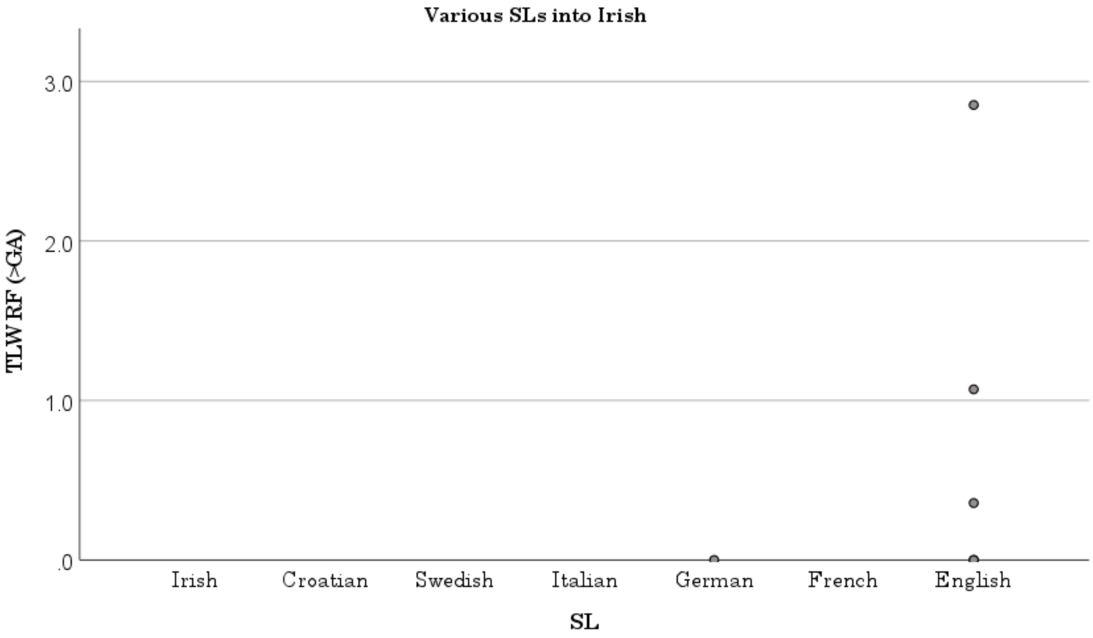


Figure 8: TLW RFs for all translations from Irish



5.6.4. Scatter plots for fixed source languages

In the scatter plots shown in Figures 9 to 15, each point represents a text in a subcorpus composed of all translations from the given SL. For each text, TLW RF is calculated as the frequency of TLWs relative to the total number of tokens in the subcorpus consisting of all translations from the fixed SL. TLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish SL subcorpus, its scatter plot is included here.

Figure 9: TLW RFs for all translations from English

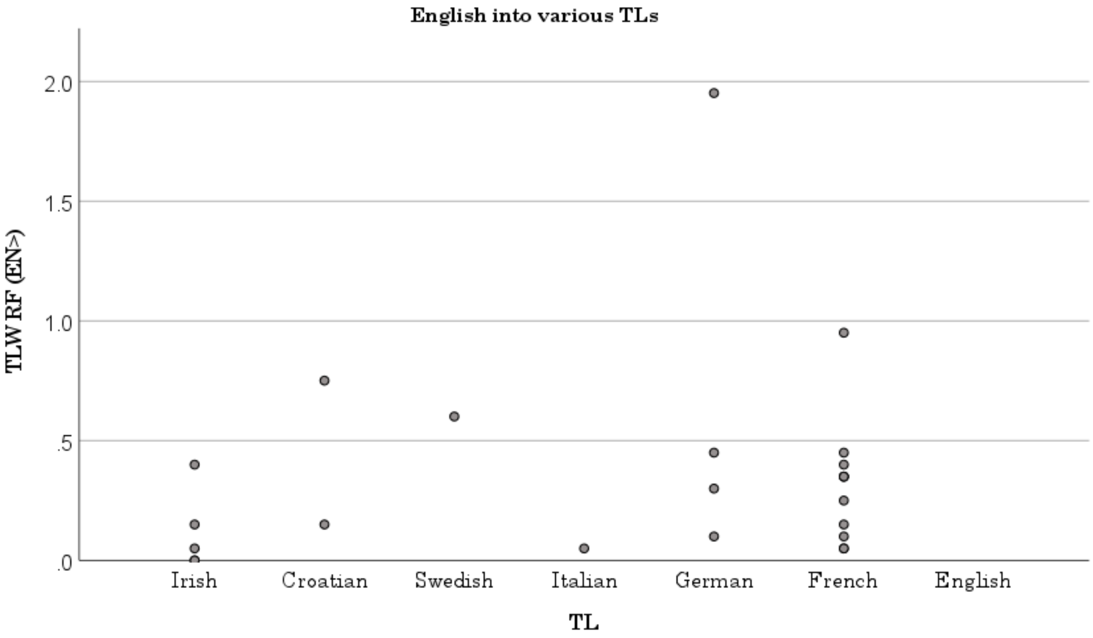


Figure 10: TLW RFs for all translations from French

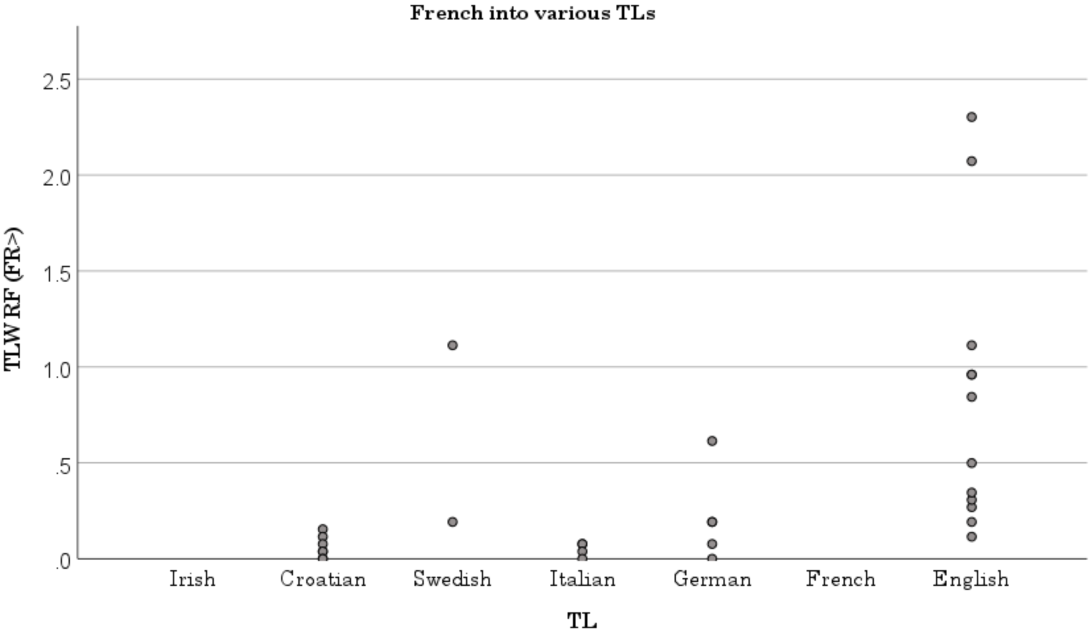


Figure 11: TLW RFs for all translations from German

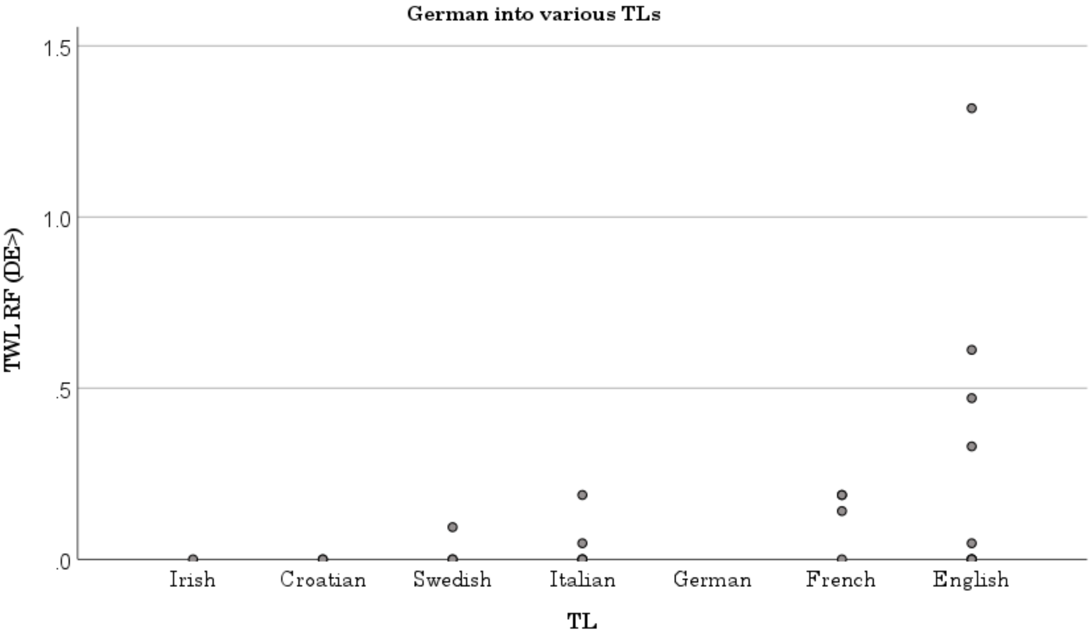


Figure 12: TLW RFs for all translations from Italian

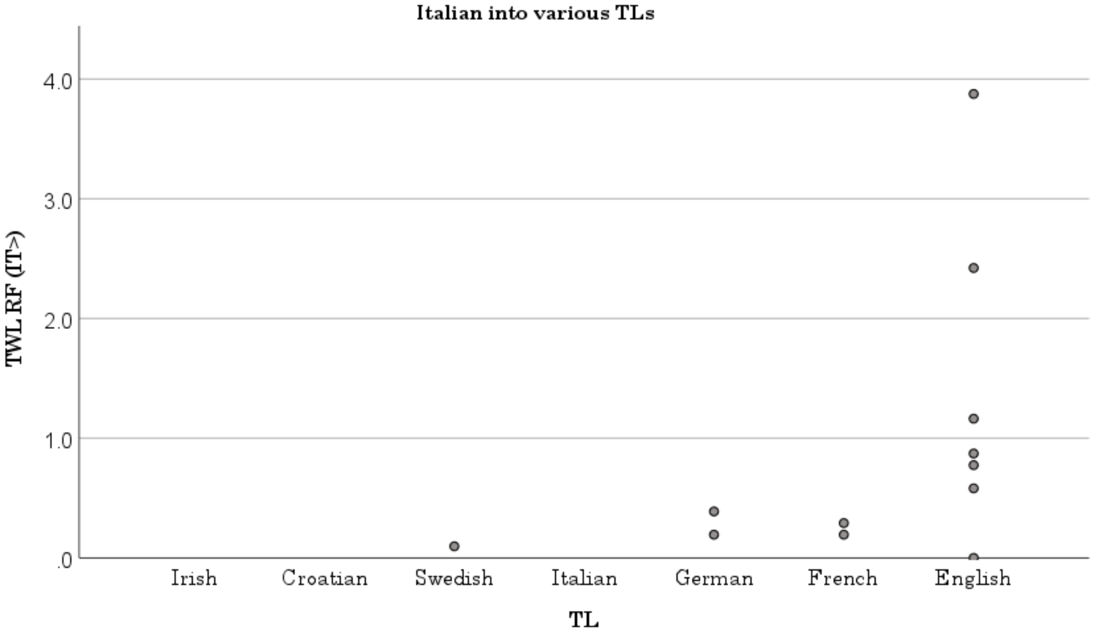


Figure 13: TLW RFs for all translations from Swedish

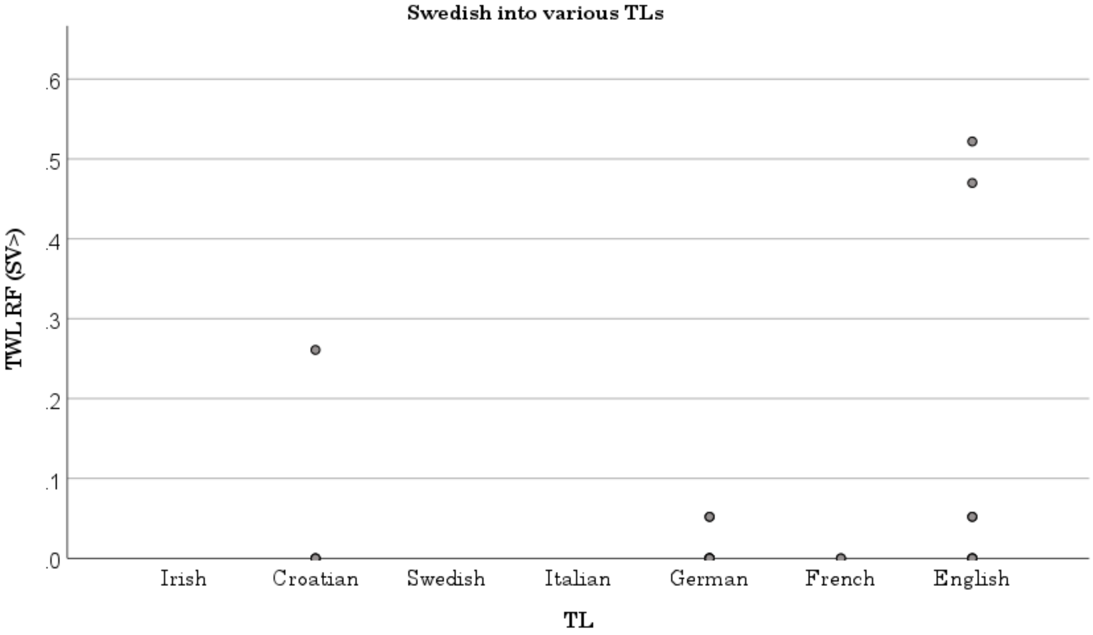


Figure 14: TLW RFs for all translations from Croatian

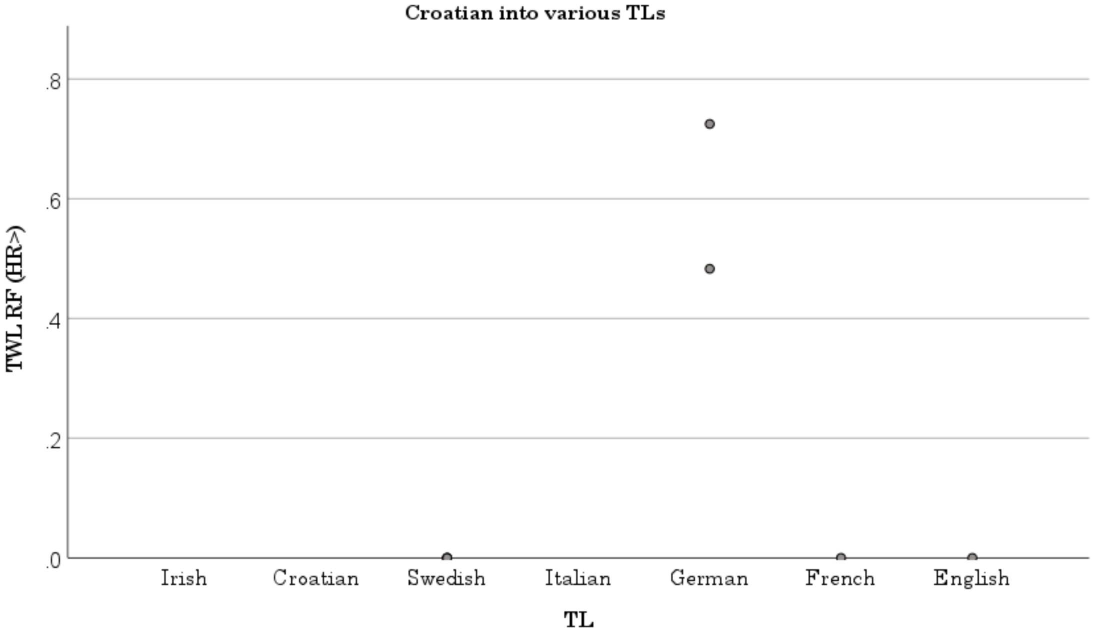
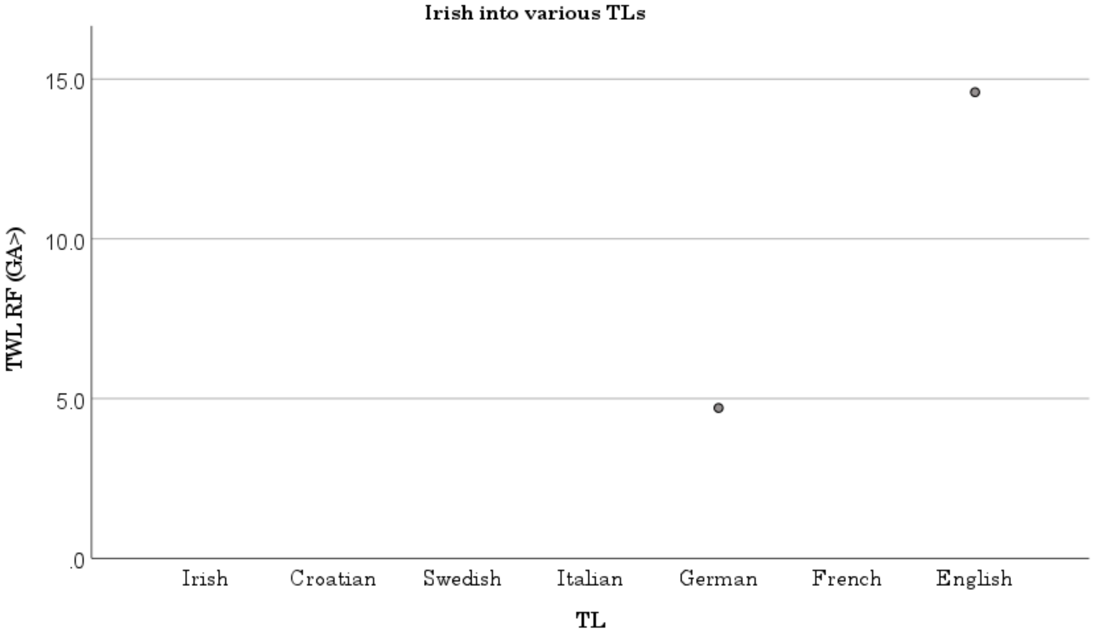


Figure 15: TLW RFs for all translations from Irish



5.6.5. Status pairs (SPs) ranked by TLW RF

In Table 14, English and French are categorized as high-status languages, German and Italian are categorized as medium-status languages, and Swedish and Croatian are categorized as low-status languages. Irish is categorized as an outlying very low-status language; given the small population of texts translated into and from Irish, it is excluded from the SPs. TLW RF is calculated relative to the subcorpus formed around each respective SP.

Table 14: Status pairs (SPs) ranked by TLW RF

Rank	Status Pair (SP)	TLW RF (/SP subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
1	high>high	17.863	1,841,750	23
2	high>medium	10.091	891,875	14
3	medium>high	8.868	1,973,481	22
4	high>low	4.655	1,611,201	12
5	low>high	2.152	975,890	12
6	medium>medium	1.777	619,171	7
7	low>low	0.975	512,991	7
8	low>medium	0.839	834,307	11
9	medium>low	0.555	540,658	6

5.6.6. Language pairs ranked by TLW RF

Table 15: Language pairs ranked by TLW RF

Rank	SL	TL	TLW AF	TLW RF (/lang pair subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
1	Irish	English	31	28.626	108,293	1
2	French	English	260	25.466	1,020,965	12
3	English	German	56	17.032	328,784	4
4	Italian	English	100	15.681	637,707	7
5	Irish	German	10	9.602	104,145	1
6	German	French	11	8.923	123,276	4
7	French	German	28	8.518	328,726	5
8	English	French	69	8.407	820,785	11
9	English	Croatian	18	8.287	217,207	2
10	French	Swedish	34	7.952	427,539	2
11	German	English	59	5.520	1,068,859	9
12	English	Italian	1	4.770	20,963	1
13	English	Irish	12	4.642	258,499	5
14	Italian	French	5	3.481	143,639	2
15	Italian	German	6	3.412	175,872	2
16	English	Swedish	12	3.411	351,759	1
17	Croatian	German	5	3.260	153,361	2
18	Swedish	English	21	2.747	764,406	8
19	French	Italian	5	2.343	213,402	4
20	French	Croatian	11	1.790	614,696	7
21	Swedish	Croatian	5	1.534	325,891	4
22	Italian	Swedish	1	1.336	74,853	1
23	German	Italian	5	1.128	443,299	5
24	German	Swedish	2	0.533	375,134	3
25	Swedish	German	2	0.294	680,946	9
26	German	Croatian	0	0.000	90,671	2

Rank	SL	TL	TLW AF	TLW RF (/lang pair subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
27	Croatian	English	0	0.000	54,070	1
28	Croatian	French	0	0.000	19,477	1
29	Swedish	French	0	0.000	137,937	2
30	German	Irish	0	0.000	21,943	1
31	Croatian	Swedish	0	0.000	187,100	3

5.6.7. Translated texts ranked by TLW RF

Table 16: All translated texts ranked by TLW RF

Rank	Translation	SL	TL	TLW AF	TLW RF (/text)
1	Hermann	German	English	28	88.608
2	The Devil's Pool	French	English	25	65.551
3	Mother of Pearl	French	English	29	56.331
4	The Romance of a Poor Young Man	French	English	22	41.598
5	Very Woman	French	English	25	37.397
6	The Triumph of Death	Italian	English	40	35.455
7	A Cardinal Sin	French	English	13	29.873
8	Le grillon du foyer	English	French	9	29.120
9	Zwei Städte	English	German	39	29.095
10	The Countess of Rudolstadt	French	English	54	29.067
11	The Dirty Dust	Irish	English	31	28.626
12	Twenty Years After	French	English	60	24.783
13	Tristan	German	French	3	22.080
14	The Patriot	Italian	English	25	20.426
15	Sretni vladar - Slika Doriana G	English	Croatian	15	20.273
16	Strife and Peace	Swedish	English	10	20.043
17	A Virgin Heart	French	English	7	18.141

Rank	Translation	SL	TL	TLW AF	TLW RF (/text)
18	Bouvard und Pécuchet	French	German	16	17.467
19	Les chasseurs de chevelures	English	French	19	15.163
20	The Desire of Life	Italian	English	12	12.599
21	Vicomte de Bragelonne	French	Swedish	29	12.020
22	Gegen den Strich	French	German	5	11.770
23	The House by the Medlar-Tree	Italian	English	9	11.563
24	Das Bildnis des Dorian Gray	English	German	9	11.517
25	Royal Highness	German	English	13	11.054
26	La guerre des mondes	English	French	7	10.956
27	A Tale of Brittany	French	English	8	10.760
28	Le portrait de Dorian Gray	English	French	8	10.401
29	Les trois hommes en Allemagne	English	French	7	10.398
30	The Dream	French	English	9	10.130
31	Cré na Cille	Irish	German	10	9.602
32	The Intruder	Italian	English	8	9.147
33	La Mère de Dieu	German	French	4	8.919
34	Cnoc na nGabha III	English	Irish	3	8.674
35	A Woman At Bay	Italian	English	6	7.742
36	Ja i moj sin	Swedish	Croatian	5	7.618
37	Der Weihnachtsabend	English	German	2	7.421
38	Die Frau von dreißig Jahren	French	German	5	7.307
39	Dans l'abîme	English	French	1	7.266
40	Dracula	English	Irish	8	7.140
41	Bübü vom Montparnasse	French	German	2	7.092
42	La débâcle impériale - Juan Fernandez	German	French	4	6.948
43	Der Amateursozialist	English	German	6	6.693
44	Un amant	English	French	7	6.587
45	Put u srediste zemlje	French	Croatian	4	6.575

Rank	Translation	SL	TL	TLW AF	TLW RF (/text)
46	La Morte a Venezia - Tristano - Tonio Kroger	German	Italian	4	6.318
47	Under Sentence of Death; Or, a Criminal's Last Hours	French	English	5	6.154
48	The Home; Or, Life in Sweden	Swedish	English	9	5.939
49	Le crime de Lord Arthur Savile	English	French	2	5.766
50	Round the World in Eighty Days	French	English	3	5.246
51	Il fantasma di Canterville e il delitto di Lord Savile	English	Italian	1	4.770
52	Le magasin d'antiquités, Tome II	English	French	5	4.219
53	Die Rückkehr des Filip Latinovicz	Croatian	German	3	4.151
54	Frederick the Great and His Family	German	English	10	3.926
55	Feu Mathias Pascal	Italian	French	3	3.879
56	Ich und Er	Italian	German	4	3.674
57	I tre moschettieri, vol. II	French	Italian	2	3.540
58	I tre moschettieri, vol. I	French	Italian	2	3.446
59	David Copperfield	English	Swedish	12	3.411
60	Une femme	Italian	French	2	3.017
61	Una donna - Geschichte einer Frau	Italian	German	2	2.985
62	Den Hemlighetsfulla ön	French	Swedish	5	2.684
63	20.000 milja pod morem	French	Croatian	3	2.539
64	Le magasin d'antiquités, Tome I	English	French	3	2.520
65	Hände	Croatian	German	2	2.466
66	Die Inselbauern; oder, Die Leute auf Hemsö	Swedish	German	1	2.164
67	Oliver Twist	English	Croatian	3	2.095
68	Joseph II. and His Court	German	English	7	1.935
69	I tre moschettieri, vol. IV	French	Italian	1	1.881
70	Cnoc na nGabha II	English	Irish	1	1.629
71	Le mort vivant	English	French	1	1.558
72	The Chief Justice	German	English	1	1.504

Rank	Translation	SL	TL	TLW AF	TLW RF (/text)
73	Invisible Links	Swedish	English	1	1.437
74	Izabrane novele - Guy de Maupassant	French	Croatian	1	1.374
75	Germinal	French	Croatian	2	1.368
76	En kvinnas liv	Italian	Swedish	1	1.336
77	Huset Buddenbrook	German	Swedish	2	1.008
78	Gospođa Bovary	French	Croatian	1	0.986
79	Il Messaggio dell'Imperatore	German	Italian	1	0.975
80	Der Sohn einer Magd	Swedish	German	1	0.952
81	Downstream	Swedish	English	1	0.790
82	Put oko svijeta u 80 dana	French	Croatian	0	0.000
83	Thérèse Raquin	French	Croatian	0	0.000
84	Preobrazaj	German	Croatian	0	0.000
85	Proces	German	Croatian	0	0.000
86	Pakleni Stroj	Swedish	Croatian	0	0.000
87	Gösta Berling (HR)	Swedish	Croatian	0	0.000
88	Legende o Kristu	Swedish	Croatian	0	0.000
89	On the Edge of Reason	Croatian	English	0	0.000
90	Blanche - The Maid of Lille	German	English	0	0.000
91	Gertrude's Marriage	German	English	0	0.000
92	The Merchant of Berlin	German	English	0	0.000
93	The Wish	German	English	0	0.000
94	Farewell Love	Italian	English	0	0.000
95	Christ Legends	Swedish	English	0	0.000
96	Married	Swedish	English	0	0.000
97	The Miracles of Antichrist	Swedish	English	0	0.000
98	The Story of Gösta Berling	Swedish	English	0	0.000
99	Enterrement à Thérésienbourg	Croatian	French	0	0.000
100	La Pantoufle de Sapho	German	French	0	0.000
101	La légende de Gösta Berling	Swedish	French	0	0.000

Rank	Translation	SL	TL	TLW AF	TLW RF (/text)
102	Au bord de la vaste mer	Swedish	French	0	0.000
103	Salambo	French	German	0	0.000
104	Christuslegenden	Swedish	German	0	0.000
105	Das Buch vom Brüderchen	Swedish	German	0	0.000
106	Die Gotischen Zimmer	Swedish	German	0	0.000
107	Ein Stück Lebensgeschichte und andere Erzählungen	Swedish	German	0	0.000
108	Gösta Berling: Erzählungen aus dem alten Wermland	Swedish	German	0	0.000
109	Pastor Hallin	Swedish	German	0	0.000
110	Unsichtbare Bande	Swedish	German	0	0.000
111	Blátha Bealtaine	English	Irish	0	0.000
112	Cnoc na nGabha I	English	Irish	0	0.000
113	Eachtra Pheadair Schlemihl	German	Irish	0	0.000
114	I tre moschettieri, vol. III	French	Italian	0	0.000
115	I Buddenbrook	German	Italian	0	0.000
116	Siddharta	German	Italian	0	0.000
117	Silvia ossia - La povera signorina - La gioventù provetta	German	Italian	0	0.000
118	Återkomsten	Croatian	Swedish	0	0.000
119	Händer	Croatian	Swedish	0	0.000
120	Begravning i Teresienburg och andra noveller	Croatian	Swedish	0	0.000
121	Amerika	German	Swedish	0	0.000
122	Slottet	German	Swedish	0	0.000

5.7. Discussion

5.7.1. Fixed TL analysis

Among the results for the fixed TL analysis, positive associations between SL status and TLW RF$_{>\mathrm{TL}}$ are detected in five of the six fixed TL subcorpora examined, lending a substantial level of support to the trend predicted in subhypothesis I:

> <u>Fixed TL subcorpora</u>:
> English ($\tau = 0.298$; weak; $p = .010$)
> French ($\tau = 0.550$; moderate to strong; $p = .001$)
> German ($\tau = 0.375$; moderate; $p = .014$)
> Swedish ($\tau = 0.713$; strong; $p = .005$)
> Croatian ($\tau = 0.469$; moderate; $p = .019$)

While a statistically significant, positive association is determined in the fixed English TL subcorpus, Figure 2 makes clear that there is not a uniformly increasing trend between SL status and TLW RF$_{>\mathrm{EN}}$, as an Italian>English translation and an Irish>English translation contain the third- and fourth-highest TLW RFs$_{>\mathrm{EN}}$, respectively. Conversely, the scatter plot for the fixed French TL subcorpus (Figure 3) reveals that the texts with the five highest TLW RFs$_{>\mathrm{FR}}$ are English>French translations – the only language pair in which French is outranked in terms of language status. The positive association confirmed for French as a fixed TL contextualizes the frequent claims of the Francophone community's linguistically conservative or protective nature (see Casanova 2002, 11), indicating that the supposed conservatism of French translations capitulates as SL status increases.

As seen in Figure 4, the texts with the two highest TLW RFs$_{>\mathrm{DE}}$ are English>German and French>German translations, yet unexpectedly, the third-ranked text is the lone Irish>German translation – the text with the lowest-status SL. Figure 5 displays the scatter plot of the one fixed TL subcorpus for which the hypothesized positive association between the variables is not found – Italian. In the fixed Italian subcorpus, a text with the lowest-status SL in the subcorpus – a German>Italian translation – surprisingly contains the highest TLW RF$_{>\mathrm{IT}}$, nearly twice that of the next highest.

In the fixed Swedish TL subcorpus, translations from French and English contain the three highest TLW RFs$_{>\mathrm{SV}}$, with texts from the rest of the SLs containing comparatively negligible levels of lexical interference (see Figure 6). Figure 7 portrays results for the fixed Croatian TL subcorpus, where only English>Croatian and French>Croatian translations register levels of lexical interference, with the exception of one Swedish>Croatian translation (*Ja i moj sin*) achieving the second-highest TLW RF$_{>\mathrm{HR}}$.

Taking the fixed TL analysis as a whole, the finding that positive associations are detected in five of the six fixed TL subcorpora examined provides considerable evidence for subhypothesis I. Still, the data point toward the necessity of the more granular analyses that will be presented later on, including the ranking of specific texts.

5.7.2. Fixed SL analysis

Among the results for the fixed SL analysis, the hypothesized trend of a negative association between TL status and lexical interference is not found in any of the fixed SL subcorpora. These findings indicate that translators working in various TLs do not uniformly respond to the status of a given SL. In fact, translations from three SLs actively contradict the negative association predicted in subhypothesis II, exhibiting a positive association:

Fixed SL subcorpora:
French ($\tau = 0.542$; moderate to strong; $p < .001$)
German ($\tau = 0.368$; moderate; $p = .016$)
Italian ($\tau = 0.530$; moderate to strong; $p = .015$)

For translations from French source texts ($n = 30$), it is perhaps not surprising to find consistently high levels of lexical interference, given the SL's high status. However, it is unexpected and noteworthy that lexical interference in translations from French increases as the TL status increases ($\tau = 0.542$, $p < .001$), since translators working in lower-status TLs would be expected to borrow TLWs from French more often than translators in comparatively higher-status TLs.

As seen in Figure 10, the texts with the five highest TLW $\text{RFs}_{\text{FR>}}$ in the fixed French TL subcorpus are English translations, which are expected to contain the lowest level of lexical interference, given that the language pair is the only one in which French is lower-status than the TL. On the opposite end of the TL status spectrum, French>Croatian translations exhibit comparatively low levels of lexical interference. This result is highly unexpected, given that translations into the lowest-status TL in the subcorpus would be predicted to exhibit the highest degrees of lexical interference from the high-status SL.

For translations in the fixed German SL subcorpus (n = 24), there is a moderate positive association between TL status and TLW $\text{RF}_{\text{DE>}}$ ($\tau = 0.368$, $p = 0.016$). Likewise, for translations from Italian (n = 12) into the various TLs, there is unexpectedly a moderate to strong positive association between TL status and TLW $\text{RF}_{\text{IT>}}$ ($\tau = 0.530$, $p = 0.015$). As Figure 12 illustrates, translations into English clearly exhibit the highest levels of lexical interference in the fixed Italian SL subcorpus, which is surprising, these translations have the highest-status TL. The levels of lexical interference in the fixed Italian SL subcorpus' other texts, however, are comparatively negligible. It is unclear why a positive association between the variables is detected for Italian and German as fixed SLs.

It is particularly noteworthy that the fixed SL analysis finds no statistically significant negative association between TL status and TLW $\text{RF}_{\text{EN>}}$ among texts in the fixed English SL subcorpus, as would be particularly expected given translation studies' persistent assertions about the universally distorting influence of English as an SL (see Venuti 1995). Translators working with English source texts therefore do not respond uniformly to power disparities between English and their respective TLs on the lexical level.

5.7.3. Status pair (SP) analysis

Ranking the various status pairs (SPs) according to TLW RF reveals a striking pattern. The top three SPs are the three possible combinations of high- and medium-status languages, with the TLW RF of the high>high SP being nearly twice that of the next highest SP (high>medium). This result indicates that translations between the high-status languages (i.e., English and French) are much more likely to contain higher levels of lexical interference than translations in other SPs. The translation direction does not seem to influence lexical interference as much as the mere presence of a high-status language in the language pair. Further supporting this observation is the fact that the three possible combinations of low- and medium-status languages constitute the three lowest-ranking SPs in terms of TLW RF. It would thus seem that high-status languages both induce (as SLs) and accommodate (as TLs) lexical interference in translation more so than low-status languages.

5.7.4. Language pair analysis

The results for the ranked language pairs somewhat complicate the pattern described previously, primarily due to one particular source text. The Irish>English subcorpus has the highest TLW RF, while the Irish>German subcorpus has the fifth-highest TLW RF. Both of these subcorpora consist of just one translation, both from the same source text – an outlier that will be further explored later on. Removing these two language pairs from the top five reestablishes the trend found among the ranked SPs, where the inclusion of a high-status SL or TL drastically increases the level of lexical interference. Translations of Swedish source texts into Croatian, German, French, and English all fail to register a single TLW. The findings in this narrow context fit the expected trend, given the SL's low status and the hypothesized reluctance for translators of Swedish source texts to render TLWs into their primarily higher-status TLs. In total, six of the multilingual corpus' language pairs are not found to contain any TLWs, and four of these language pairs have a low-status SL (Swedish or Croatian), while the other has a very low-status TL (Irish).

The only text in the Irish>English subcorpus is Alan Titley's translation of Máirtín Ó Cadhain's 1949 Irish novel *Cré na Cille* (trans. *The Dirty Dust*, 2015). Gabriele Haef's German translation of the same novel (trans. *Grabgeflüster*, 2017) also makes up the sole text in the Irish>German subcorpus. Given the source text's monumental importance among Irish-language literature, it may be that translators approach the novel with greater reverence for the lexical nuances of Irish, particularly given the language's critical role in Ireland's history and national identity. In this particular scenario, the source text's specific political or historical conditions may override the low status of Irish in terms of their influence on the translators' lexical choices.

5.7.5. Ranked-text analysis

An analysis of the data at the text level reinforces the emerging tendency for translations involving high-status SLs or TLs to exhibit higher levels of lexical interference. The text with the highest level of lexical interference relative to its total tokens is the German>English translation *Hermann* (TLW RF$_{\text{text}}$ = 88.608). These TLWs mostly pertain to the novel's preoccupation with the social identifiers of German nobility. For instance, eight of the TLWs found in *Hermann* are inflectional variants of the adjective *bürgerlich* ("civil"), used in reference to commoners outside the noble class. The next four texts with the highest TLW RFs$_{\text{text}}$ are French>English translations: George Sand's *The Devil's Pool* (trans. 1851 by George B. Ives), Anatole France's *Mother of Pearl* (trans. 1922 Frederic Chapman), Octave Feuillet's *The Romance of a Poor Young Man* (trans. 1907 Henry Harland), and Remy de Gourmont's *Very Woman* (trans. 1922 J. L. Barrets). Common TLWs among these French>English translations are salutations (e.g., *bonjour*, *au revoir*, and *a bientôt*) and interjections (e.g., *mon Dieu* and *peste*). Of the top 20 texts with the highest TLW RFs$_{\text{text}}$, exactly half (10/20, 50%) are translations between the project's two high-status languages, while *all* of them (20/20, 100%) include at least one high-status language in their language pair.

English and French translators appear far less likely to adopt TLWs from Swedish and Croatian, and these low-status languages generally seem to induce very little lexical interference. Out of the 40 translations in which no TLWs were identified, 72.5% (29/40) involve a low-status SL or TL, and 15% (6/40) are translations between Swedish and Croatian – the two low-status languages. These granular findings combine with the higher-level analyses described previously to confirm an emergent pattern. Although the study's results do not support its hypothesis, the data reveal a strong tendency for high-status languages to both induce and accommodate lexical interference in translation, particularly when paired with their high-status counterpart.

5.8. Limitations

One potential confounding variable that the present study does not account for is what Haugen (1950, 223) terms a "structural resistance to borrowing". Orthographic or phonetic similarities between SLs and TLs may very well play an important role in translators' decisions to produce TLWs in the target text. Relatedly, the study's rather narrow operationalization of lexical interference – defining TLWs as word forms that are *exactly* the same in the SL as in the TL – likely excludes a number of lexical borrowings. Calques or loanblends, for example, may be slightly adapted toward the structural conventions of the TL. Although the present study does not identify these types of borrowings, they nonetheless constitute evidence of lexical interference, albeit to a lesser degree than lexical borrowings that reproduce the exact SL word form.

The study's use of online dictionaries to determine the established presence of SL word forms in TL lexicons presents another set of complications. Like language status, language contact is necessarily historically and thus temporally situated, meaning that its effects on receiving languages' lexicons may be observed in one year (or period) but not the other. The dictionaries used as references for established TL lexicons reflect modern lexicons, while many of the corpus texts reflect language from an earlier period. In theory, the comparison of TLW candidates to their presence in both the comparable SL and TL corpora may indicate an established presence of an SL word form in the TL lexicon, though it is worth emphasizing that the relatively small size of each comparable corpus reduces the reliability of this method. Moreover, dictionaries for the same language may have different lexical coverages; the *Merriam-Webster* English dictionary, for example, includes a number of French words and phrases that the *Cambridge Dictionary* does not. These cases were resolved via additional research and subjective judgment. Such discrepancies and subjective assessments thus played a role in determining which TLW candidates had already been established in the TL lexicon, and consequently may have skewed the data. In light of this particular limitation, it is important to emphasize once again that the definition of TLWs presented here is tailored toward studies involving monolingual comparable corpora. The availability of parallel corpora – i.e. translations aligned with their source texts – would remove the current method's reliance on the subjective determination of which TLWs are "reasonably suspected" to be borrowed over the course of translation instead of prior language contact.

5.9. Chapter conclusion

The study presented in this chapter did not produce evidence in support of the hypothesized positive association between comparative SL status and lexical interference, operationalized as the RF of translator-attributed loanwords – translational loanwords (TLWs) – in a given subcorpus or text. The results of the fixed TL analysis provided considerable evidence for subhypothesis I, as positive associations are determined for five of the six TLs. However, the results of the fixed SL analysis contradicted subhypothesis II: no negative associations were detected, yet positive associations were determined for three of the six SLs. Therefore, the data did not constitute evidence of a consistently positive association between SL status and lexical interference in translation.

However, examining the data from alternative perspectives revealed a consistent pattern: language pairs involving high-status languages (i.e. English or French) tended to exhibit higher levels of lexical interference. This pattern seemed to hold true regardless of whether the high-status language served as the SL or TL. In fact, translations between the two high-status languages exhibited far higher levels of lexical interference than translations between any other status groups. The texts with the three highest TLW RFs – and four of the top five texts – were French>English translations. Language pairs involving low-status languages also appeared to be far less likely to contain TLWs. The three possible combinations of low- and medium-status languages contained the lowest levels of lexical interferences among status pairs, and nearly three-fourths of the 40 translations for which no TLWs were identified involved a low-status SL or TL.

The high levels of lexical interference found in the English and German translations of *Cré na Cille* indicated that certain political or historical considerations may take precedence over the influencing factor of language status in translation. This particular case may enhance the discussion of the distinction between language status and language prestige, which will be revisited in the conclusion to the thesis (Chapter 8). Similar studies may develop methods to distinguish these closely related concepts as separate variables and subsequently gauge their individual effects on the linguistic features of translated texts.

The results open many other avenues for further research. Future studies may operationalize and calculate frequencies for additional lexical borrowing phenomena such as calques and loanblends, stratifying these borrowings by their strength of interference and calculating composite levels of lexical interference accordingly. Different types of lexical borrowings may be said to represent lexical interference at varying levels of strength based on whether they constitute full, partial, or zero morphemic importation (Haugen 1950, 230). This expanded operationalization of lexical interference would entail the challenge of systematically identifying non-exact matches of word forms across various language pairs.

Other studies may quantify levels of interference using parallel corpus methodology, as exemplified in Hansen-Schirra (2011). Such research would move beyond the present study's inferential means of measuring source-language interference by explicitly identifying and aligning word forms that occur in both source and target texts. The data presented here further suggest that other textual metadata – such as author or translator – could influence lexical interference in translation. For instance, Zlatko Goran translated both English and German into Croatian. His translation of Oscar Wilde's *The Portrait of Dorian Gray* (trans. *Slika Doriana Graya*) contained numerous TLWs (AF = 15), while his translation of Franz Kafka's *Die Verwandlung* (*Preobražaj*) contained none. In order to make such determinations, it would be necessary to design a corpus that repeats certain authors and/or translators in a systematic fashion, as the current corpus is not suitable for this aim.

Further research may also account for the role of linguistic typology in facilitating or discouraging lexical interference in translation. The tendency for lexical interference to be more pronounced in translations between the two high-status languages could be partially attributable to the etymological ties between English and French. While this avenue would necessitate an advanced approach toward codifying linguistic similarity as an additional explanatory variable, the overlapping lexicons of the study's selected languages may be characterized in part by the number of Type II TLW candidates identified for each language pair. That is, the number of word forms found in all three relevant corpora – the SL and TL comparable corpora as well as the SL>TL translation corpus – may indicate the extent of languages' overlapping lexicons and hence linguistic similarity. This proposed research avenue may also consider a TL's "structural resistance to borrowing" (Haugen 1950, 223), drawing on attempts to characterize languages' "lexical borrowability" (see Van Hout and Muysken 1994). For other, non-European language pairs in other scripts, it may also be necessary to characterize the role of transliteration in lexical interference. Ultimately, language status is simply one variable that may influence translation strategies, and additional research will be necessary to examine its influence across time periods, text types, and languages.

6. Syntactic interference

6.1. Chapter introduction

In comparison to lexical interference, interference on the syntactic level may take much subtler forms. Syntax may be characterized as "the rules that determine the sentence structure of a particular language" (Costa-Jussà and Farrús 2014, 12). Syntactic interference is thus highly dependent upon the degree of similarity between the sets of rules governing SL and TL syntaxes, which of course varies substantially among different language pairs.

Translation scholars have speculated that language status influences the extent to which translations' syntactic features either conform to TL conventions (normalization) or subvert them in order to accommodate those of their source texts (interference), regardless of structural compatibility between SLs and TLs. Baker (1996, 183) posits that translations from high-status SLs are less likely to normalize. She asserts normalization to be "most evident in the use of typical grammatical structures" and emphasizes punctuation in particular as reflecting this universal translation phenomenon (ibid.). Despite his scant discussion of the syntactic manifestations of interference, Toury (2012, 312) notes that "the distance between languages" appears to have no "automatic bearing" on interference in translated texts. Venuti (1995, 122-123) also makes sparing references to the imposition of SL syntactic conventions on target texts, alluding for instance to the "syntactic inversions" in Francis Newman's 1851 English translation of Horace's Latin.

The relative subtlety of syntactic interference in translation perhaps indicates that this form of cross-linguistic influence is more aptly hosted in neighboring fields such as contact linguistics, whose methodologies are more inclined toward gradual, historically contextualized language change. Contact linguistics boasts a much richer and longer history than translation theory with regard to syntactic interference (see Aikhenvald 2007). Kroch (2001, 716) cites language contact as one of the primary drivers of syntactic change. One pertinent historical example is offered by Hickey (2010, 18-19), who emphasizes the heavy influence of English syntax on modern Irish, whose word order now more closely resembles that of English despite the major typological dissimilarity between the two languages. Characterizing Irish syntax as "permeable" in this respect, he contends that this historical case exemplifies the tendency for a language's marginalized position to leave it vulnerable to being "infiltrated syntactically by a co-existing dominant language" (Hickey 2010, 19-20). The syntactic influence of the "super-dominant English" on Irish is still observable in "subtle" yet "infiltrating" ways (ibid., 20), better captured via a broad analysis of gradual contact between the two languages than synchronic comparisons of translated Irish with non-translated Irish.

Research on bilingualism conceptualizes interference in a similar manner, often referring to the effects of syntactic or structural priming on bilingual speakers. Maier et al. (2017) find that, when asked to translate utterances between German and English, bilinguals (untrained as translators) show a clear tendency to preserve the word order of the source segment, indicating evidence of cross-linguistic structural priming. Chen et al. (2013) similarly find that cross-linguistic syntactic priming between Chinese and English manifests in terms of word order. These studies demonstrate the framing of cross-linguistic syntactic interference as a psycholinguistic phenomenon.

Broadly speaking, neighboring fields' investigations of syntactic influence between languages commonly frame word order as a fundamental element of syntax. Word order may be considered as a metafunction of syntax, representing the "marking of syntactic functions" (Teich 2003, 54-55). Moreover, words may be further assigned to their superordinate parts-of-speech (POS) categories for the purpose of cross-linguistic comparisons of syntactic features. POS categories are therefore primary constituent elements of syntax; the syntactic structure of any language may be largely defined by the ordered dependencies of these constituent elements, or rather, the permissible – or simply probabilistic – orders in which POS categories appear in relation to one another. By conceptualizing syntax as the aggregate of commonly used POS sequences, a given language's syntax may be characterized by the RF distribution of its POS n-grams. In this manner, it is possible to compare the syntactic composition of two languages by comparing their POS n-gram RF distributions. Certain POS n-grams may have very similar rates of usage across two structurally similar languages, such as French and Italian or other languages with common historical roots. Other POS n-grams may occur only in one language or the other. Comparing the POS n-gram relative frequency (RF) distributions of SLs, TLs, and SL>TL translated texts may therefore indicate the extent to which SLs cause syntactic interference in translated texts. As such, this study's operationalization of syntactic interference/normalization in translation is based on these comparisons.

6.2. Related works

To date, corpus-based research on syntactic interference in translation has overwhelmingly focused on the reproduction of specific SL syntactic features in structurally compatible TLs. It is now commonly accepted in corpus-based translation studies that SL interference is likely when SLs and TLs have common syntactic features (De Sutter and Vermeire 2019, 13). In a straightforward example, Maia (1998) examines the frequency of first-person pronouns in original and translated English and Portuguese, finding clear evidence that the structural rigidity of English manifests in translated Portuguese despite the latter's greater flexibility in word order.

Having developed more complex methodologies, the literature on syntactic interference in translation now offers greater depth in the contrastive typologies of SLs and TLs. Teich (2003) offers what is perhaps the most in-depth empirical study of syntactic interference in translation to date, applying both parallel and comparable corpus methodology to the analysis of scientific texts translated between English and German. As a necessary precursor to the study, she conducts a multi-faceted contrastive analysis of the typological differences between the languages, first identifying then comparing the frequencies of common syntactic properties between comparable SL and TL texts, then using these comparisons to predict the various syntactic features' frequencies in translated texts (ibid., 220-222). POS categories form the basis of analysis for the majority of the syntactic features under examination, with POS tagging being the only automatic annotation tool used due to its simplicity and high reliability (ibid., 167). Differences between the predicted and observed frequencies of syntactic features are then characterized and quantified as SL interference or TL normalization depending on whether they are closer to their corresponding frequencies in the comparable SL and TL subcorpora (ibid., 181). Her study finds no consistent patterns of interference and normalization among the selected syntactic features, regardless of the translation direction (Teich 2003, 208). However, the results also show that, overall, German translations deviate more from the syntactic conventions of comparable German texts than English translations from comparable English texts, which she attributes to structural differences between the two languages (ibid., 218). Best demonstrated in this study, systematically accounting for structural differences between SLs and TLs is a defining feature of much research on syntactic interference and normalization in translation.

Investigating syntactic interference in literary translations between German and Dutch, De Sutter and Van de Velde (2008) use a registerially-controlled, mixed comparable and (bidirectionally) parallel corpus to compare frequencies of the relative placements of prepositional phrases. Noting that most corpus-based research merely identifies features distinguishing translations from non-translations (i.e., translation universals), they aim to disentangle structural from non-structural influences ("language-internal and language-external factors") in order to uncover "differences in the underlying cognitive-functional system that determines linguistic choices" (ibid., 2-3). Similar to Teich (2003), their methodology detects a statistically significant difference in the number of prepositional phrases used in original Dutch and original German, formulating their predictions for the frequencies of this phenomenon in translations between the two languages accordingly (De Sutter and Van de Velde 2008, 9). The study determines, for instance, that "translated Dutch moves away from the syntactic preferences of original Dutch" and "tak[es] a position in between original Dutch and (original) German" (ibid., 13). In this regard, German>Dutch translations are shown to exhibit syntactic interference, while translations in the opposite direction exhibit syntactic normalization (ibid., 31-32). These findings seem to implicitly confirm the current study's predicted association between SL status and syntactic interference, as the German SL would be assumed to be higher in status compared to Dutch, given the former language's more widespread use.

Combining product- and process-oriented research methods, Hansen-Schirra (2011) also applies both parallel and comparable corpus methodology to characterize syntactic interference and normalization in translations of scientific texts in the English-German language pair. She identifies a variety of common syntactic features to be compared between translations (in both directions) and the SL and TL comparable corpora, using POS categories as the basis of many of these features (Hansen-Schirra 2011, 155). As with Teich (2003), her study determines a mixture ("hybridization") of interference and normalization among the selected features (Hansen-Schirra 2011, 155). In one example, it is found that English texts use first-person pronouns more often than German texts, and that German translations of English texts therefore use more first-person pronouns than their original German counterparts (ibid., 152). Although the study does not directly examine language status, it posits the dominant position of English source texts in scientific discourse as a potential reason why original German scientific texts now exhibit a higher frequency of first-person pronouns compared to older texts (ibid., 154). While the studies mentioned in this section exemplify methodologies that take into account the structural differences between specific languages, they are perhaps not amenable to the aims of the study presented here, which requires techniques that may be more readily applied to diverse language pairs. In this vein, some studies have centered on cross-linguistic comparisons of POS n-grams, as these sequences constitute "abstract syntactic structures devoid of content" that prove relatively stable when examined across different languages (Lembersky et al. 2012, 809).

Hadley (2023) devises two methods of comparing the syntactic compositions of direct and indirect translations; both methods involve calculating differences between the RFs of POS n-grams in translated texts and their corresponding RFs in their source texts and a wide range of reference corpora. His case study compares POS n-gram RFs between the English novel *Oliver Twist*, its direct French translation, and its indirect Spanish translation (translated from the French version) with those of (non-)translated (non-)fiction in all three languages (ibid., 106). His first method of operationalizing syntactic similarity entails determining which of the reference corpora contains the most POS n-grams whose RF is closest to its corresponding value in the translation (ibid., 106). A foreseeable issue with this method is that it effectively weights all POS n-gram comparisons equally regardless of their frequencies in the translation. For instance, if the RFs of a translation's five most common POS n-grams are closest to their corresponding values in reference corpus A, and the RFs of the translation's five rarest POS n-grams are closest to their corresponding values in reference corpus B, the method would assess these data points as equal indicators of syntactic conformity. The second method entails determining which of the reference corpora has the lowest aggregate differential between its values corresponding to the POS n-gram RFs in the translation (Hadley 2023, 106). While this calculation does factor in the magnitudes of POS n-gram RF differentials, Hadley (ibid., 108) cautions that aggregating these POS n-gram RF differentials does not take into account their direction or distribution in each reference corpus, creating the possibility for them to "balance one another out when all the differences are considered in aggregate."

This approach does not explicitly account for structural differences between languages. Hadley (2023, 132-133) points out that the languages in his analysis are closely related, but anticipates that "substantial difference in the grammatical structures of the languages" would not undermine his devised methodology. In this manner, the methodology is ostensibly replicable across languages while maintaining the comparability of results. However, the framing offered here departs significantly from the previously described studies, which are eminently concerned with properly accounting for the syntactic (in)compatibility between SLs and TLs. According to Teich (2003, 220), it is "impossible to say what SL shining through [read: interference] or TL normalization mean without making reference to contrastive knowledge about the language systems of which the texts under investigation are instantiations." The tension between these methodological polarities – i.e., maximizing or minimizing the use of explicit contrastive linguistic knowledge in measuring syntactic interference – perhaps indicates that an ideal methodology represents some balance of language-agnostic and language pair-specific characteristics, depending on the study's aims and context.

Syntactic features with significantly different frequencies in translations relative to comparable TL texts also provide fruitful grounds for more granular analyses. Using comparable and parallel corpus methodology, Chlumská (2018) examines the behavior of three prominent POS 4-grams whose increased frequencies distinguish Czech literary translations (translated from English) from original Czech literature, demonstrating the manner in which these POS sequences reflect syntactic interference from their English source texts. Based on comparisons with corresponding (word-based) n-grams, she concludes that the distinct behavior of the selected POS n-grams in Czech literary translations reflects a mixture of translation universals (explicitation, normalization, and interference), and does not seem to be attributable to any one factor in particular (Chlumská 2018, 115).

Other works have established fundamental linguistic differences between translation and non-translated texts on the basis of aggregate POS n-gram distributions. Volansky et al. (2015) apply supervised machine learning techniques to test 32 potential linguistic classifiers distinguishing translated texts from original TL texts, sorting each classifier into categories of hypothesized features of translation: simplification, explicitation, normalization, interference, or miscellaneous. Posited as potential classifiers of interference, distributions of POS 1-grams, 2-grams, and 3-grams are indeed found to be highly accurate (≥90%) predictors of a text's translation status (ibid., 110). On the whole, their findings indicate that interference is the "most robust phenomenon typifying translations" (ibid., 111). Lembersky et al. (2012, 822) also report substantial differences in POS n-gram distributions between English translations and original English texts. Translations' distinct POS n-gram distributions are in part why language models built on translated texts enhance statistical MT more than language models built on original TL texts (ibid., 809).

As the first data-driven (i.e., corpus-based) approach to automatic translation, statistical MT (SMT) struggled considerably with word order (see Bisazza and Federico 2016). Word-order issues may be partially remedied by replacing tokens with POS tags and leveraging information from POS sequences. Popović and Ney (2006) leverage POS tags to improve the word reordering of phrase-based SMT, offering different mechanisms for language pairs requiring short- or long-range reorderings. Syntactic information has proven vital to MT evaluation, as well. The most historically popular automatic evaluation metric, BLEU (Papineni et al. 2002), is widely criticized for its inability to reflect the suitability of MT output syntax (see Castilho et al. 2018, 18). Popović and Ney (2011, 665) demonstrate that POS tags constitute vital linguistic information that may enhance automatic MT evaluation, including the identification and analysis of diverse errors. Naturally, language pairs involving divergent syntactic structures such as German and English are more likely to produce word order errors for SMT (ibid., 681). The development of neural MT (NMT) in the mid-2010s transformed the entire apparatus of automatic translation strategies, and in turn reconfigured the typical syntactic features characterizing MT output. As noted by Bentivogli et al. (2018, 2): “Word reordering is the strongest aspect of NMT compared to [other MT] systems.” Subsequent research compares the syntactic features of various MT systems.

Toral (2019) measures various linguistic features projected to distinguish post-edited MT output from exclusively human-produced translations, finding the former to exhibit a higher degree of SL interference in terms of its POS sequences. Using UDPipe to replace all tokens with automatically generated POS tags, the study measures the level of interference for each translation (post-edited MT or human-produced) as the translation’s perplexity with respect to an SL language model minus its perplexity with respect to a TL language model (ibid., 6). Perplexity is a logarithmic function that denotes the appropriateness of a language model in predicting a given text, where higher values express a higher degree of unpredictability. In this manner, a high (positive) result for the calculated difference in perplexities indicates that a translation’s POS sequences are more similar to those of the TL, while a low (negative) result indicates that they are more similar to those of the SL. As expected, raw and post-edited MT exhibit the highest and second-highest degrees of interference (i.e., lowest perplexity difference scores), and statistical MT output exhibits more interference than neural MT output (ibid., 7). Although these findings are fairly consistent across the selected language pairs, the study’s methodology is tailored for comparisons of interference in various MT architectures’ output instead of between language pairs. Furthermore, the use of perplexity to evaluate interference requires additional machine learning techniques and does not easily lend itself to microstructural analyses.

Thus, it appears that no translation studies research to date has examined the effects of language status and syntactic interference/normalization as a primary focus or for a wide range of languages. What is needed is a language-universal methodology for characterizing the levels of syntactic interference or normalization in translation, such that these measurements may still be directly compared across the typologically distinct language pairs included in this project. Although Teich’s (2003, 224) methodology is hypothetically language-independent, its requirements for highly involved contrastive SL-TL analyses and the availability of both parallel and comparable corpora make this approach unrealistic for the current study’s purposes, as this project includes only comparable corpora and translation corpora for a much wider variety of language pairs.

For these reasons, it is necessary to develop a new methodology for operationalizing syntactic interference and normalization in translation.

6.3. Methodology

6.3.1. Calculating SINC

This study introduces a novel, language-agnostic operationalization of syntactic interference and normalization in translation – a metric termed the **syntactic interference/normalization coefficient (SINC)**. This metric encapsulates a comparison of the RF distributions of POS n-grams across a comparable SL subcorpus, a comparable TL subcorpus, and a given translated text. Comparisons between these three RF distributions may be used to triangulate the degree to which the syntactic features of translations conform to typical SL syntactic features or typical TL syntactic features.

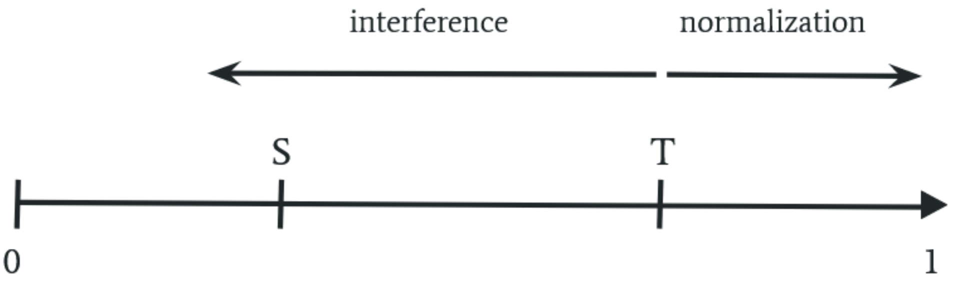
It is worth reiterating that one of the foundational aims of corpus-based translation research is the methodical and empirical study of *translational behavior* – the typical features that distinguish translated texts from non-translated texts. As demonstrated in the preceding section, differences between POS n-gram RF distributions of translated texts and those of comparable non-translated texts may be counted among these distinguishing features. The process of identifying macro-level differences between the syntactic compositions of translated and non-translated texts is best demonstrated by considering this comparison for an individual POS n-gram.

6.3.1.1. Calculating SINC for individual POS n-grams

Because any translated text is necessarily linguistically situated in the TL, and is therefore subject to the structural constraints of the TL syntax, the RF of any given POS n-gram in a comparable TL subcorpus may serve as the baseline or *expected value* of the RF of that same POS n-gram in the translated text. The extent to which the observed value in the translated text diverges from this expected value constitutes translational behavior insofar as it reveals whether a particular syntactic construction – a sequence of POS categories – is used in the translation under examination relatively more or less often than in comparable TL texts. According to Van Oost et al. (2016, 3), normalization is when a translation conforms to the linguistic conventions of comparable TL texts, while *over*-normalization is characterized by the "over-use of typical patterns of the target language." For the purposes of this study, "normalization" is conceptualized as the latter, i.e., the *exaggeration* of typical TL patterns. Using this conceptual framework, syntactic interference/normalization in translation may be characterized by further comparison to the POS n-gram's corresponding RF in the SL comparable corpus.

Recalling that interference constitutes the influence of SL features on a translated text, while normalization constitutes the exaggeration of typical TL features in the translated text, the comparison of the RFs of any one particular POS n-gram across the three relevant textual bodies (comparable SL subcorpus, comparable TL subcorpus, and the translated text itself) may be visualized along a simple line graph. Figure 16 below compares a given POS n-gram's hypothetical RFs across SL and TL reference corpora, with the RF values exaggerated for the sake of visibility:

Figure 16: Comparison of a POS n-gram's RFs in comparable SL and TL subcorpora



S = RF of the given POS n-gram in the comparable SL subcorpus
T = RF of the given POS n-gram in the comparable TL subcorpus

The arrows in Figure 16 represent the potential divergence of the (unpictured) observed value – i.e., the POS n-gram's RF in the translated text (hereafter "X") – from the expected value (T). Regardless of the relative positionings of S and T, the observed value's (X) divergence from the expected value (T) may be conceptualized as *movement* either toward or away from S, constituting interference and normalization, respectively. Here, X's movement toward S naturally represents interference, whereas its movement away from S represents normalization – an exaggeration of the direction of T relative to S[13].

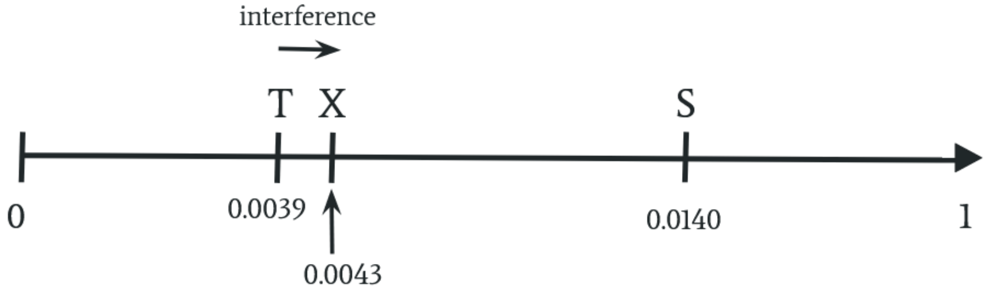
However, the difference between the observed value (X) and the expected value (T) alone cannot be taken as a decisive metric of interference/normalization that may be compared across diverse language pairs, as the distance between these two values is surely constrained in some manner by the syntactic similarities between the SL and the TL. While the values X and T are inherently comparable given that both are situated in the same grammatical system, S is derived from an entirely different grammatical system, and is thus subject to an entirely different set of structural constraints. Because the present study's operationalization of syntactic interference must account for all possible language pairs in the multilingual corpus, the applied metric must be able to control for the structural differences between various syntaxes. What is needed in order for this calculation to be language-universal is a means of expressing the distance between X and T relative to the "natural" difference between a given POS n-gram's compatibility with the SL and TL syntaxes.

A comparison between a given POS n-gram's RFs in the SL and TL comparable subcorpora may fulfill this function. Conceptually, the distance between S and T embodies the degree to which a POS n-gram differs in its habitual usage within the SL and TL grammatical systems, respectively. For example, if the POS 3-gram *det-adj-noun* has a relatively high value for S, it indicates that this sequence is a fairly common grammatical convention in the SL. If this same POS 3-gram has a value for T that is significantly closer to zero, it means that the sequence *det-adj-noun* is highly unusual and therefore less structurally compatible with the TL grammatical system. The greater the distance between S and T, the more the SL and TL syntaxes differ in their structural amenability to the POS n-gram in question. The usage of *det-adj-noun* in a translation, then, must be contextualized relative to this distance. This scenario is demonstrated in Figure 17, which illustrates the RF of the POS 3-gram *det-adj-noun* in

[13] As made clearer later on, the value of T is not necessarily greater than the value of S; therefore, it is important to emphasize that interference is always characterized as the movement of X in the direction of S, regardless of the relative values (or positionings) of S and T.

the English>Croatian translation *Oliver Twist* relative to the comparable SL and TL subcorpora:

Figure 17: SINC(gram) visualization for det-adj-noun *in EN>HR* Oliver Twist



S = RF of the POS 3-gram *det-adj-noun* in the comparable English subcorpus

T = RF of the POS 3-gram *det-adj-noun* in the comparable Croatian subcorpus

X = RF of the POS 3-gram *det-adj-noun* in the EN>HR translation *Oliver Twist*

Visualized in the manner above, it is evident that the translation's increased usage of *det-adj-noun* relative to original Croatian texts has moved toward that of the comparable English texts, reflecting interference. However, this movement appears very slight when compared with the natural difference between English and Croatian in the usage of POS sequence. Calculating the degree of syntactic interference or normalization embodied by a particular POS *n*-gram's comparative usage in a translated text therefore requires a formula that logically connects these two pertinent distances or comparisons: X – T and S – T.

For any given POS n-gram in a translation in any given language pair, the degree of syntactic interference/normalization its usage reflects may be calculated as the magnitude of the distance between the expected and observed values (|X – T|) expressed as a proportion of the distance between the differences in its usage between the comparable SL and TL subcorpora (|S – T|). This calculation results in a standardized coefficient which may be compared across different POS n-grams in different language pairs.

However, in cases where the denominator (|S – T|) happens to be an extremely small value (relatively speaking), and the numerator (|X – T|) happens to be an extremely large value, this calculation produces a massive coefficient that disproportionately skews the progressively larger-scale SINC calculations (described later on) that characterize a translated text's composite or aggregate level of syntactic interference/normalization. Such cases were noticeably frequent in preliminary tests of the SINC methodology. It was determined that these extreme values did not portray meaningful indicators of syntactic interference or normalization, and instead reflected situations in which the RF of a POS n-gram in translation deviated significantly from its corresponding values in both the comparable SL and TL subcorpora, with their disproportionate impacts on higher-level SINC obscuring the effects of more pertinent POS n-gram comparisons. In fact, when this basic calculation amounts to a value outside the range between -1 and $+1$, it perhaps indicates that comparisons encoded in this calculation exceed the boundaries of the concept intended to be operationalized – namely, the extent to which a translation's syntax deviates from TL conventions *relative to SL conventions* (i.e., syntactic interference/normalization). This logic is also applied by Teich (2003, 210), who notes that it is impossible to characterize syntactic features in translation as interference (or "shining through" as she described it) or normalization when "there is no difference between the original [SL and TL] texts in the first place." A more theoretically sound method of determining syntactic interference/normalization in translation therefore prevents the operationalized metrics of these concepts from exceeding the magnitude of the observable differences between the syntactic structures of SLs and TLs. To this end, the devised formula limits the values of raw SINC scores for individual POS n-grams to 1 by defining the equation's denominator as the *maximum value* between the two comparisons ($|X - T|$ and $|S - T|$), such that, in the event that the POS n-gram's RF in the translation deviates more from that its of corresponding value in the TL more than the SL value deviates from the TL value (i.e., $|X - T| > |S - T|$), the formula will produce a maximum output of 1, since the numerator will simply be equal to the denominator.

Synthesizing all of these conditions, the following formula (conditioned for consistent positive/negative representations of interference and normalization, respectively) has been developed to calculate the *raw* SINC score of a given POS n-gram in a translated text (raw $\text{SINC}_{\text{gram}}$):

$$raw\ SINC_{gram} = \frac{|X - T|}{max(|S - T|, |X - T|)}$$

S = RF of the given POS n-gram in the comparable SL subcorpus

T = RF of the given POS n-gram in the comparable TL subcorpus

X = RF of the given POS n-gram in the translated text

However, because the variables S, T, and X may assume different relative values and positionings across different POS n-grams, the incidental positive/negative signs of the values in the numerator ($|X - T|$) and denominator ($|S - T|$) will interact unpredictably. Therefore, in order to characterize interference and normalization consistently, it is necessary to first calculate the absolute values of the two distances in the formula, and only thereafter assign a positive or a negative value to the overall coefficient.

Regardless of the relative order of S, T, and X for any particular POS n-gram, interference is always assigned a positive value for the sake of consistency, meaning that SINC is positive:

syntactic interference: $SINC_{gram} > 0$

Conversely, normalization is always assigned a negative value:

syntactic normalization: $SINC_{gram} < 0$

As such, the SINC formula to include a constant (C) that variably equals –1 or 1 depending on the following conditions:

If X deviates from T in the direction of S (i.e., syntactic interference): [i.e., if S < X < T, T < X < S, X < S < T, or T < S < X]

$$C = 1$$

If X deviates from T away from S (i.e., syntactic normalization): [if X < T < S, or S < T < X]

$$C = -1$$

Therefore, the *true* SINC score of a given POS n-gram in a translated text (true $\text{SINC}_{\text{gram}}$[14]) is calculated as follows:

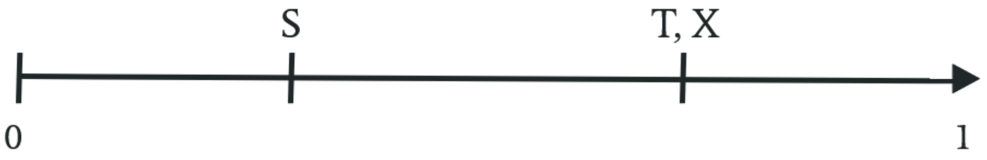
$$true\ SINC_{gram} = C \times \frac{|X - T|}{max(|S - T|, |X - T|)}$$

Practically speaking, it may be unlikely that any $\text{SINC}_{\text{gram}}$ score will amount to exactly –1, 0, or 1, but it will help to visualize each of these scenarios in order to illustrate the logic and interpretations of the formula.

In the unlikely event that $\text{SINC}_{\text{gram}} = 0$, as shown in Figure 18, there is neither interference nor normalization for the POS n-gram in question, as the observed value X precisely matches the expected value T (i.e., the numerator equals zero).

[14] Hereafter, all references to “$\text{SINC}_{\text{gram}}$ scores” refer to *true* $\text{SINC}_{\text{gram}}$ scores.

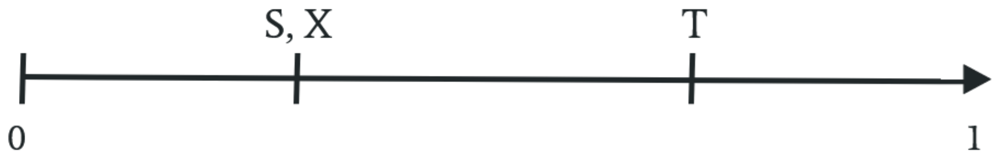
Figure 18: SINC(gram) example no. 1



Regardless of the distance between S and T, the fact that X is equivalent to T results in a SINC$_{\text{gram}}$ score of 0.

As illustrated in Figure 19, when SINC$_{\text{gram}}$ is equal to +1, the RF of the POS n-gram in the translated text (X) is equivalent to its counterpart in the comparable SL subcorpus (S), so that the distance expressed in the numerator ($|X - T|$) is perfectly proportional to that of the denominator ($|S - T|$).

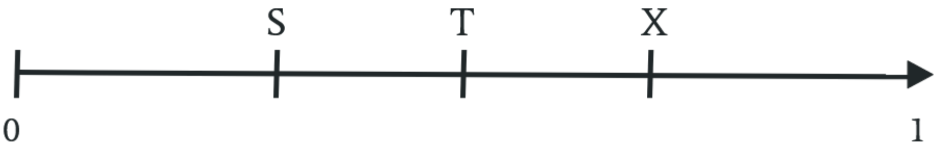
Figure 19: SINC(gram) example no. 2



In a certain sense, a SINC$_{\text{gram}}$ of +1 may therefore be conceptualized as "full interference", in that the usage of a given POS n-gram in the translation fully aligns with its corresponding usage in the SL. (Note that, although the observed value X is lower than the expected value T, it has moved in the direction of S, which constitutes interference and thus warrants a positive value for the true SINC$_{\text{gram}}$ score.) Similarly, in the event that the POS n-gram occurs in the comparable TL subcorpus ($T > 0$) but not in the comparable SL subcorpus ($S = 0$), and its RF in the translation is subsequently lower than in the original TL texts ($X < T$), it may be interpreted to support Tirkkonen-Condit's (2004) unique items hypothesis, which posits that a TL item – for example, a syntactic structure – that does not have a counterpart in the SL (or possibly an infrequently-occurring SL counterpart) will tend to be underrepresented in translation.

As another demonstration, when $\text{SINC}_{\text{gram}}$ is equal to –1, as shown in Figure 20, the interpretation is less straightforward. In this case, the distances in the numerator and denominator are once again perfectly proportional, except X has moved in the opposite direction from S, extending beyond T.

Figure 20: SINC(gram) example no. 3



A $\text{SINC}_{\text{gram}}$ score of –1 ("full normalization") thus denotes an *exaggeration* of the TL syntactic conventions that is exactly proportional to the natural difference between the SL and TL syntaxes in the use of the POS n-gram in question, as represented by S – T. (Note that, although the observed value X is higher than the expected value T, it has moved in the opposite direction of S, which constitutes normalization and thus warrants a negative value for the true $\text{SINC}_{\text{gram}}$ score.)

In this manner, the usage of any given POS *n*-gram in translation may be categorized and measured as an indicator of syntactic interference or normalization. It should be emphasized that the interpretations of $\text{SINC}_{\text{gram}}$ scores amounting to +1 or –1 ("full interference" or "full normalization") are strictly applicable individual POS n-grams, and not for the overall syntactic composition of a translated text. In order to characterize a translated text's overall degree of syntactic interference or normalization, it is necessary to aggregate $\text{SINC}_{\text{gram}}$ scores into progressively higher levels.

6.3.1.2. Calculating SINC for various n-gram sizes

The methodology has thus far abstained from defining the exact POS n-gram sizes under examination, although this determination is crucial to the methodology. Hadley (2023, 116) demonstrates that "higher n-gram numbers effectively reduce comparability" between translations and comparable texts, and subsequently chooses 4-grams as the maximum n-gram length for his methodology. BLEU is also typically calculated up to 4-grams (see Papineni et al. 2002; Callison-Burch et al. 2006), as co-occurrences of n-grams in the candidate and reference translations also tend to decrease as the n-gram length increases. Similarly seeking to maximize the comparability of POS sequences, this study therefore calculates SINC$_{\text{gram}}$ scores on the basis of POS n-grams up to 4-grams. Having set the maximum POS n-gram length to 4-grams, it is now necessary to define a method for calculating a single summary statistic to convey the composite effect (i.e., syntactic interference or normalization) of all SINC$_{\text{gram}}$ scores of a given *n*-gram length in a translated text. This value is encapsulated in SINC-*n*, where *n* represents a particular *n*-gram length.

The **SINC-*n* score** denotes the degree to which all POS grams of *n* length in a given translation reflect an aggregate level of interference or normalization. It is not rationalizable to calculate SINC-*n* as a simple arithmetic mean of all SINC$_{\text{gram}}$ scores of a given *n*-gram length, as this method would fail to account for the fact that certain POS n-grams occur much more frequently than others in a translated text. The most frequently occurring POS n-grams should therefore factor more heavily into the overall characterization of a translated text's distinguishing syntactic features.

For this reason, it is more justifiable to calculate SINC-*n* as the weighted arithmetic mean of the SINC values of all POS n-grams in the translated text, where each SINC$_{\text{gram}}$ score is weighted according to the POS n-gram's frequency in the translated text:

For each POS n-gram length (with *m* total POS n-grams):

$$SINC\text{-}n = \frac{\sum_{i=1}^{m} Z_i \, X_i}{\sum_{i=1}^{m} X_i}$$

Z = the SINC$_{\text{gram}}$ score for each POS *n*-gram of the length specified in SINC-*n*
X = the POS n-gram's RF in the translation

Using the formula above, for example, the SINC-2 score of the English>Croatian translation *Oliver Twist* amounts to +0.069, which is interpreted as interference. On the contrary, if SINC-*n* were to be calculated using a simple average, this example would yield a SINC-2 score of –0.037, which would be interpreted as normalization. It is apparent that this methodological choice has significant consequences for the current study. As stated earlier, the major advantage of using a weighted average over a simple arithmetic average is that it prioritizes the most frequent POS n-grams in the translation. In the aforementioned English>Croatian translation, the POS 2-gram *noun-punct* occurs 10,050 times (RF = 0.058) while *adj-propn* occurs 136 times; the formula above rightfully factors the more frequent POS 2-gram more heavily into the SINC-2 score.

By compiling individual SINC$_{\text{gram}}$ scores for POS sequences of *n*-gram length into composite SINC-*n* scores, it is possible to devise a comprehensive measure of syntactic interference/normalization encapsulating all POS n-grams in an entire translated text.

6.3.1.3. Calculating a text's composite SINC score

The **SINC$_{\text{text}}$ score** denotes the degree to which a given translation's SINC-*n* scores – and, by extension, all of its SINC$_{\text{gram}}$ scores – reflect an aggregate level of interference or normalization. As in the calculation of SINC$_{\text{gram}}$ scores, the positive and negative signs of SINC-*n* scores may interact unpredictably if aggregated by multiplication, as would be the case in harmonizing the SINC-*n* scores by taking a geometric mean such as with BLEU. Because the variance in the text's POS n-grams' RFs has already been accounted for in the SINC-*n* score calculations, the composite SINC$_{\text{text}}$ score for a translated text may be calculated as the simple arithmetic average:

$$SINC_{\text{text}} = \frac{\sum_{n=1}^{4} SINC\text{-}n}{4}$$

Thus, comparisons of the RFs of POS n-grams in translated texts with their corresponding values in comparable SL and TL subcorpora may be ultimately combined into a single metric for the overall degree of interference/normalization exhibited by a translation's syntactic composition. The following section details the process of implementing this methodology, including the specific digital tools used.

6.3.2. Implementing SINC

The language-agnostic tagger spacy-udpipe[15] was used to generate universal POS tags for all texts in all languages. The tool is based on resources developed as part of the Universal Dependencies project[16], which is a massive, crowdsourced initiative to develop digital tools for "crosslinguistically consistent morphosyntactic annotation" for a vast range of typologically diverse languages (De Marneffe et al. 2021, 255). In careful consideration of theoretical developments in linguistic typologies, the Universal Dependencies team puts forth a set of 17 universal POS tags capable of categorizing any word in any language (ibid., 260-261). Notably, punctuation marks are assigned the POS tag *punct*; this study includes all POS n-grams involving *punct* tags, given Baker's (1996, 183) assertion of the strong manifestations of syntactic interference in this dimension.

All texts in the corpus were converted from raw tokens into their POS tags. The AntConc corpus processing software was used to derive POS n-grams of up to 4-grams from the POS-only versions of the comparable subcorpora and translated texts. Using spreadsheets, POS n-gram RFs in each translated text were then compared to their counterparts in the comparable SL and TL subcorpora. $\text{SINC}_{\text{gram}}$, SINC-*n*, and $\text{SINC}_{\text{text}}$ scores were calculated accordingly. (See Worksheet 4 for an example of the SINC calculations performed for a sample translation.)

[15] https://spacy.io/universe/project/spacy-udpipe/
[16] https://universaldependencies.org/

6.3.3. Hypothesis testing

The two complementary subhypotheses constituting the project's primary data analysis (see Section 4.6.2.) are adapted to this study as follows:

Subhypothesis I:

> As SL status increases relative to the TL status, translations are expected to exhibit an *increasing* degree of syntactic interference. Therefore, as the TL remains constant and the SL increases, it is expected that there is a *positive association* between SL status and $\mathrm{SINC}_{\mathrm{text}}$.

Subhypothesis II:

> As TL status increases relative to the SL status, translations are expected to exhibit a *decreasing* degree of syntactic interference. Therefore, while the SL remains constant and the TL increases, it is expected that there is a *negative association* between TL status and $\mathrm{SINC}_{\mathrm{text}}$.

6.4. Results

Initially, a general overview of the data is provided by way of basic summary statistics. The study then conducts the fixed TL and fixed SL analyses in order to test the study's two subhypotheses. Finally, this section ranks texts according to their $\text{SINC}_{\text{text}}$ scores.

6.4.1. Summary statistics

The distributions of $\text{SINC}_{\text{text}}$ and SINC-*n* scores offer preliminary indications of the devised methodology's effectiveness in controlling for variation in the syntactic similarity between SLs and TLs. SINC scores should, in theory, represent comparable values across translations of typologically diverse language pairs, thereby enabling valid, language-agnostic comparisons of the degrees of syntactic interference/normalization in translation. Table 17 below provides summary statistics on the distributions of $\text{SINC}_{\text{text}}$ and SINC-*n* scores across all 122 translations in the corpus:

Table 17: Summary statistics for SINC(text) and SINC-n scores

	$\text{SINC}_{\text{text}}$	SINC-1	SINC-2	SINC-3	SINC-4
mean	+0.025	+0.105	+0.007	–0.025	+0.015
median	+0.037	+0.098	–0.002	–0.028	–0.130
st. dev.	+0.140	+0.249	+0.110	+0.087	+0.415
min.	–0.266	–0.484	–0.308	–0.239	–0.403
max.	+0.391	+0.820	+0.291	+0.169	+0.948
range	+0.657	+1.304	+0.598	+0.407	+1.351

The distribution of $\text{SINC}_{\text{text}}$ scores is the most crucial subset of the data in Table 17, as it constitutes the primary unit of the study's operationalization of syntactic interference/normalization. The similarity between the mean (+0.025) and median (+0.037) $\text{SINC}_{\text{text}}$ scores indicates a highly symmetrical distribution. Furthermore, in consideration of the study's theoretical underpinnings, the closeness of these values to zero indicates that the mean and median translations bear close syntactic resemblance to comparable texts, exhibiting only very slight levels of syntactic normalization.

The translation subcorpus (n = 122) is comprised of a nearly even split between translations from comparatively lower-status languages into higher-status languages (n = 64; 52.46%) and translations from comparatively higher-status languages into lower-status languages (n = 58; 47.54%). A natural extension of the study's hypothesis is the anticipation that there is likewise a roughly even split between the number of translations exhibiting an overall syntactic interference ($\text{SINC}_{\text{text}} > 0$) and those exhibiting an overall syntactic normalization ($\text{SINC}_{\text{text}} < 0$). Since SINC scores are calculated for both individual texts as well as n-grams of different sizes, and may be characterized as either interference or normalization depending on whether they are positive or negative values, the tendency for SINC scores of any level to skew heavily in one direction could reveal unintended bias in the methodology. For example, if all 122 texts' SINC-4 scores amount to negative values (normalization), it could be the case that calculating SINC-4 by using the RF distributions of POS 4-grams across the relevant subcorpora invariably results in negative values, perhaps due to the much higher number of possible POS 4-grams in texts and subcorpora compared to n-grams of smaller sizes. It is thus useful to calculate the total number of SINC scores at various levels across the 122 translated texts constituting interference and normalization, as shown in Table 18:

Table 18: Distribution of SINC(text) / SINC-n scores - interference/normalization

	$\text{SINC}_{\text{text}}$	SINC-1	SINC-2	SINC-3	SINC-4
interference (> 0)	65 (53.28%)	80 (65.57%)	61 (50%)	50 (40.98%)	27 (22.13%)
normalization (< 0)	57 (46.72%)	42 (34.43%)	61 (50%)	72 (59.02%)	95 (77.87%)

Table 18 shows that the distributions of SINC-1 and SINC-3 scores are somewhat comparable, while the distribution of SINC-2 scores is perfectly balanced. However, the skew of SINC-4 scores toward normalization (95/122; 77.87%) is apparently causing the $\text{SINC}_{\text{text}}$ scores' slight skew toward normalization, since $\text{SINC}_{\text{text}}$ represents the arithmetic mean of a translated text's SINC-*n* scores. Still, the imbalance between texts exhibiting interference (65/122; 53.28%) and normalization (57/122; 46.72%) does not flag any glaring methodological issues, as it represents a feasible outcome of the study.

6.4.2. Fixed TL and fixed SL analyses

Tables 19 and 20 present the results of the fixed TL and fixed SL analyses, displaying Kendall's tau value, p-value, and population size for each fixed TL or fixed SL analysis as the complementary language (SL and TL, respectively) increases in status. If detected, statistically significant findings (p ≤ .05) confirming the expected association between the study's two variables are highlighted in gray, and statistically significant findings contradicting the expected association are italicized.

Table 19: Fixed TL analysis for syntactic interference/normalization

TL (fixed)	n (texts)	Kendall's tau* (increasing SL status)	p value
English	38	– *.234*	*.029*
French	20	– .084	.320
German	23	.575	< .001
Italian	10	.112	.344
Swedish	10	.000	.500
Croatian	15	– .045	.416
Irish	6	–	–

*Hypothesized positive association.

Table 20: Fixed SL analysis for syntactic interference/normalization

SL (fixed)	n (texts)	Kendall's tau** (increasing TL status)	p value
English	24	*.708*	< *.001*
French	30	.066	.320
German	24	.195	.108
Italian	12	.131	.294
Swedish	23	*.354*	*.016*
Croatian	7	.053	.437
Irish	2	–	–

**Hypothesized negative association.

6.4.3. Scatter plots for fixed target languages

In the scatter plots in Figures 21-27, each point represents a text in a subcorpus composed of all translations into the specified TL. SLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish TL subcorpus, its scatter plot is included here.

Figure 21: SINC(text) scores for all translations into English

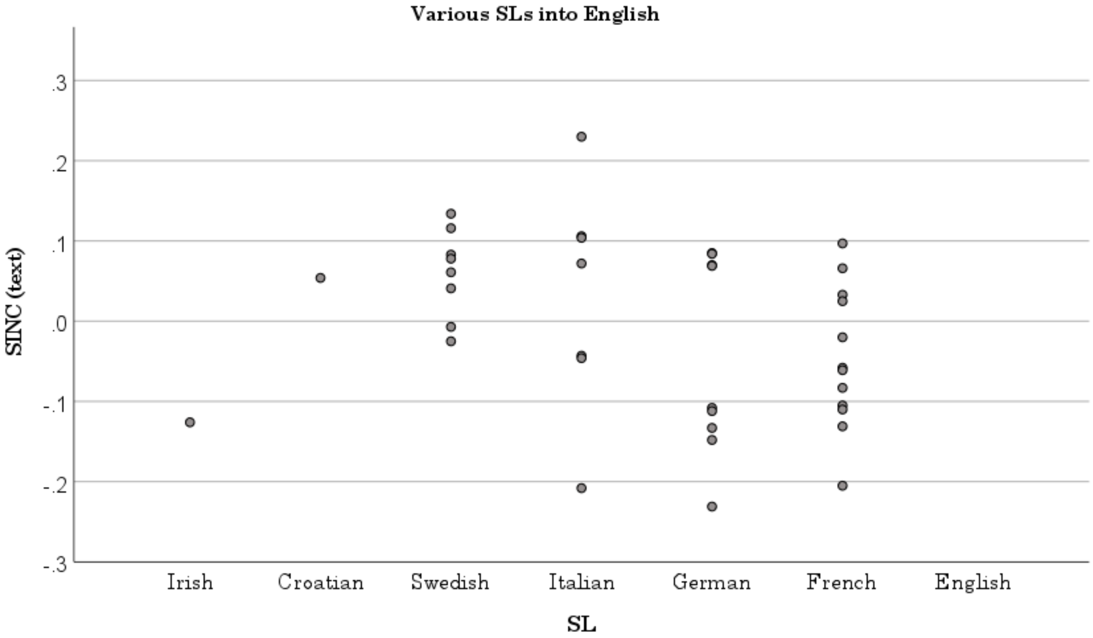


Figure 22: SINC(text) scores for all translations into French

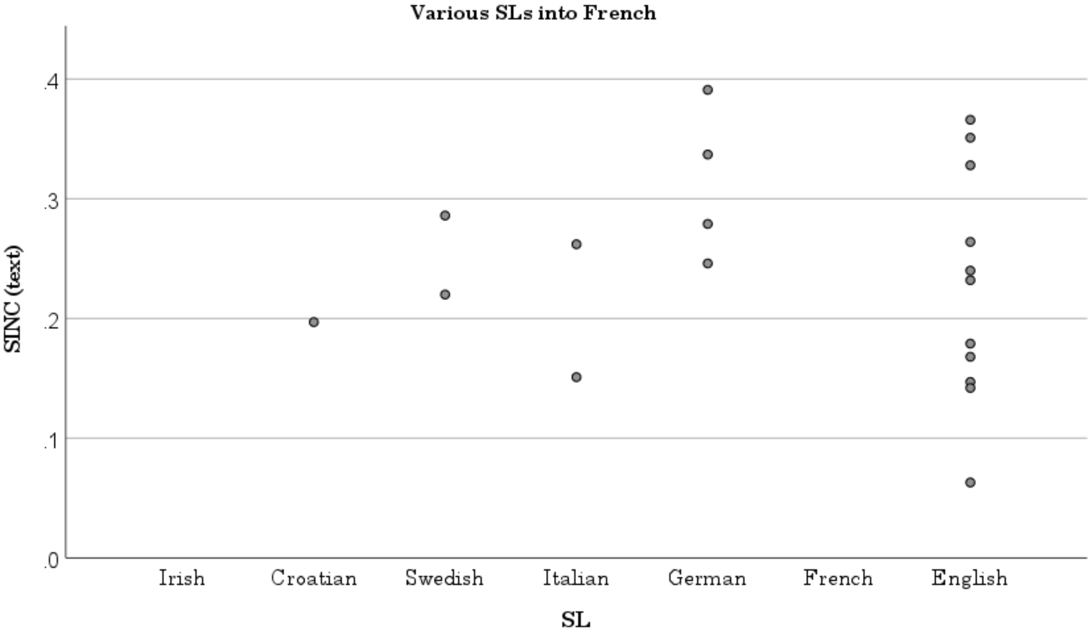


Figure 23: SINC(text) scores for all translations into German

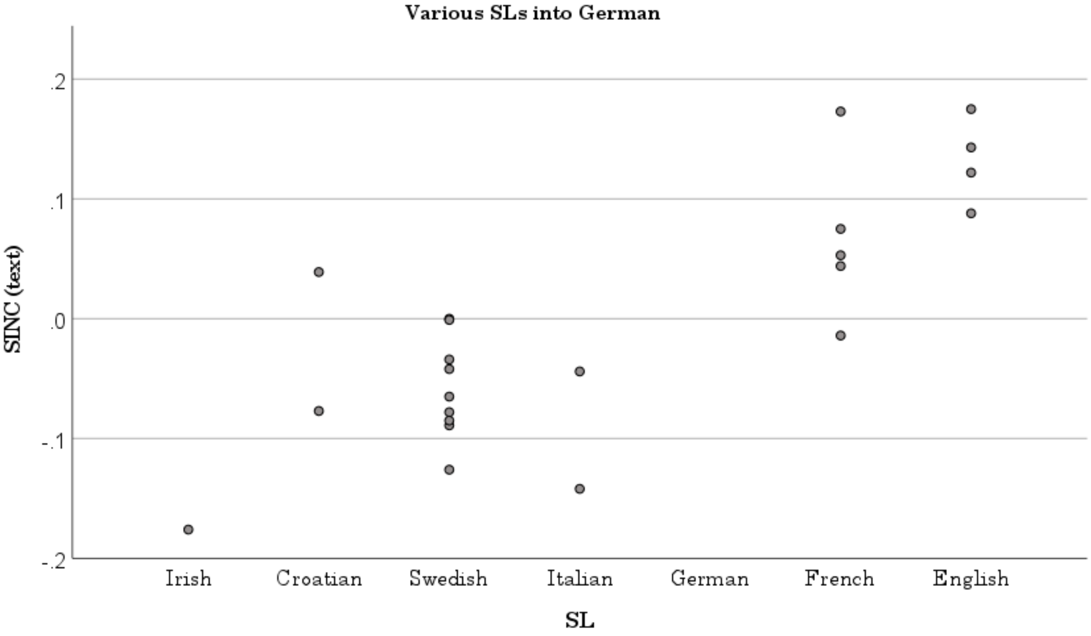


Figure 24: SINC(text) scores for all translations into Italian

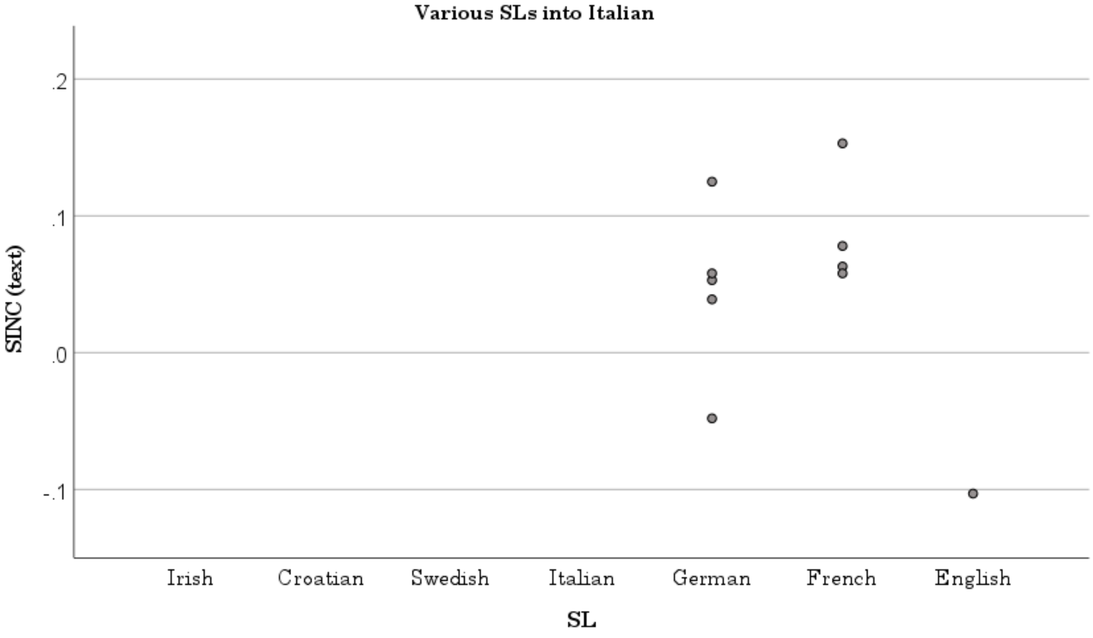


Figure 25: SINC(text) scores for all translations into Swedish

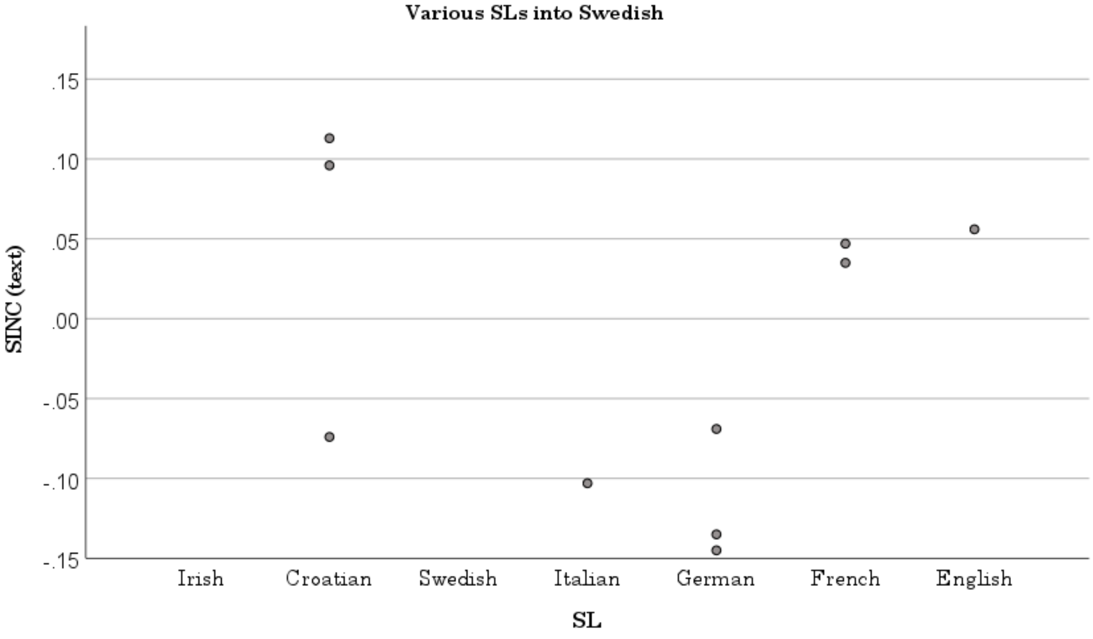


Figure 26: SINC(text) scores for all translations into Croatian

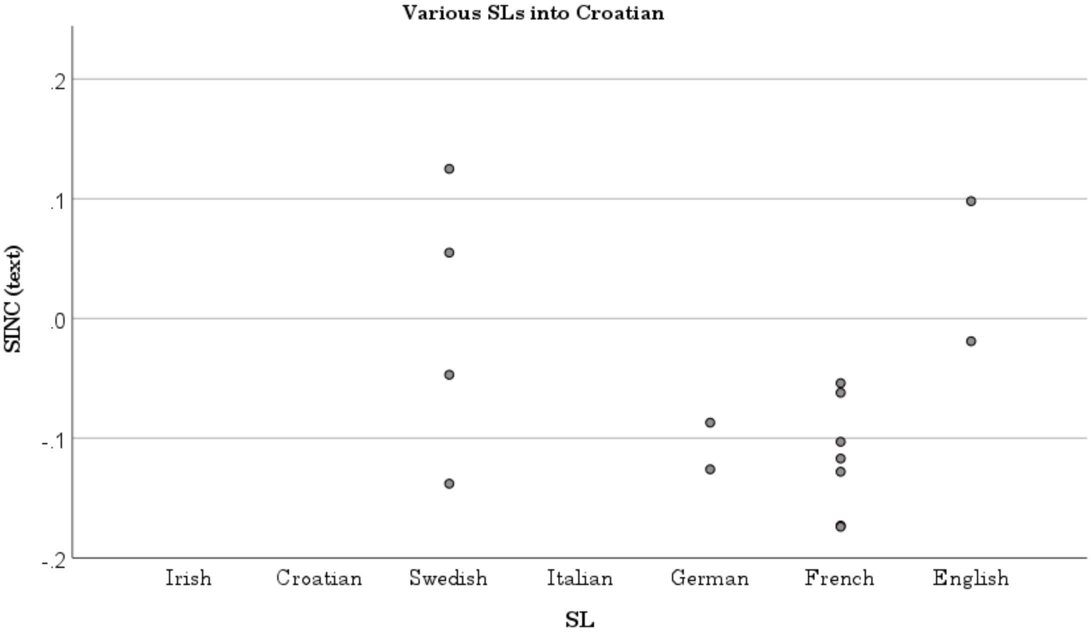
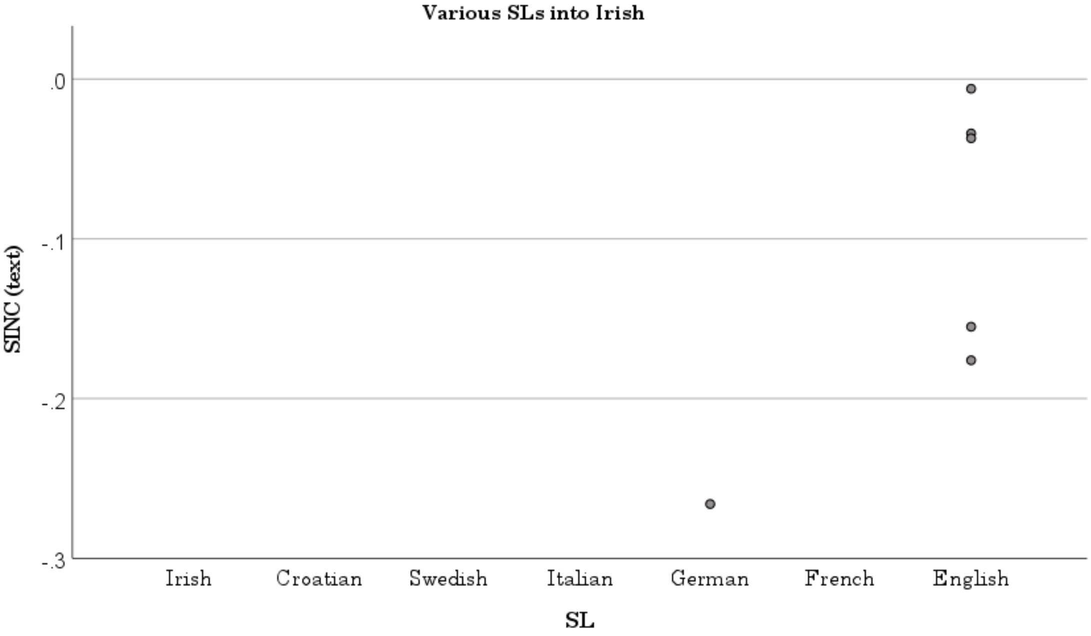


Figure 27: SINC(text) scores for all translations into Irish



6.4.4. Scatter plots for fixed source languages

In the scatter plots in Figures 28-34, each point represents a text in a subcorpus composed of all translations from the given SL. TLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish SL subcorpus, its scatter plot is included here.

Figure 28: SINC(text) scores for all translations from English

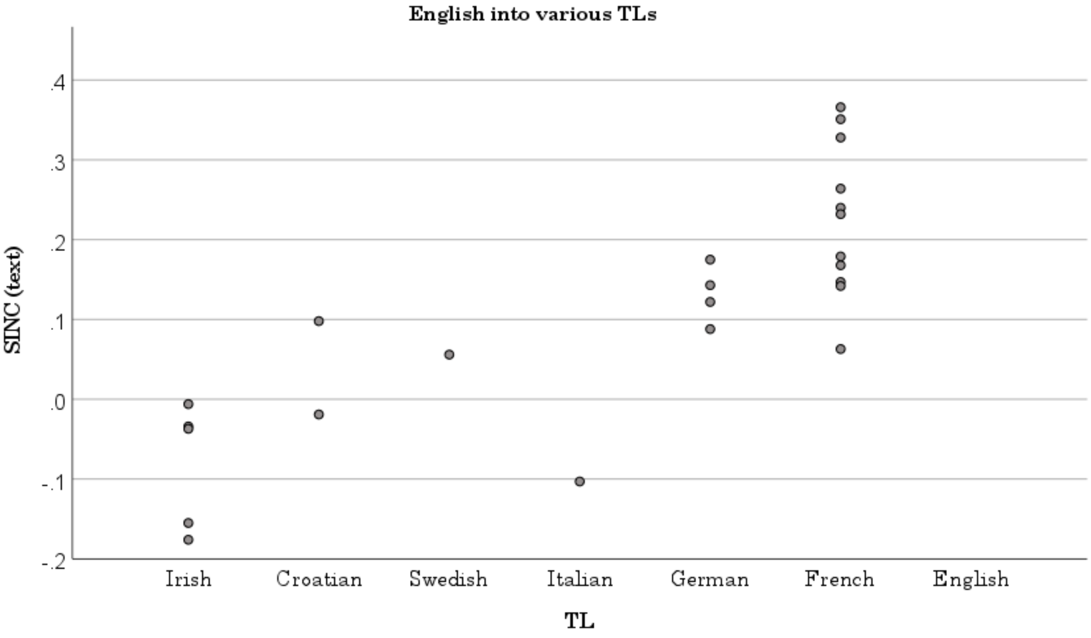


Figure 29: SINC(text) scores for all translations from French

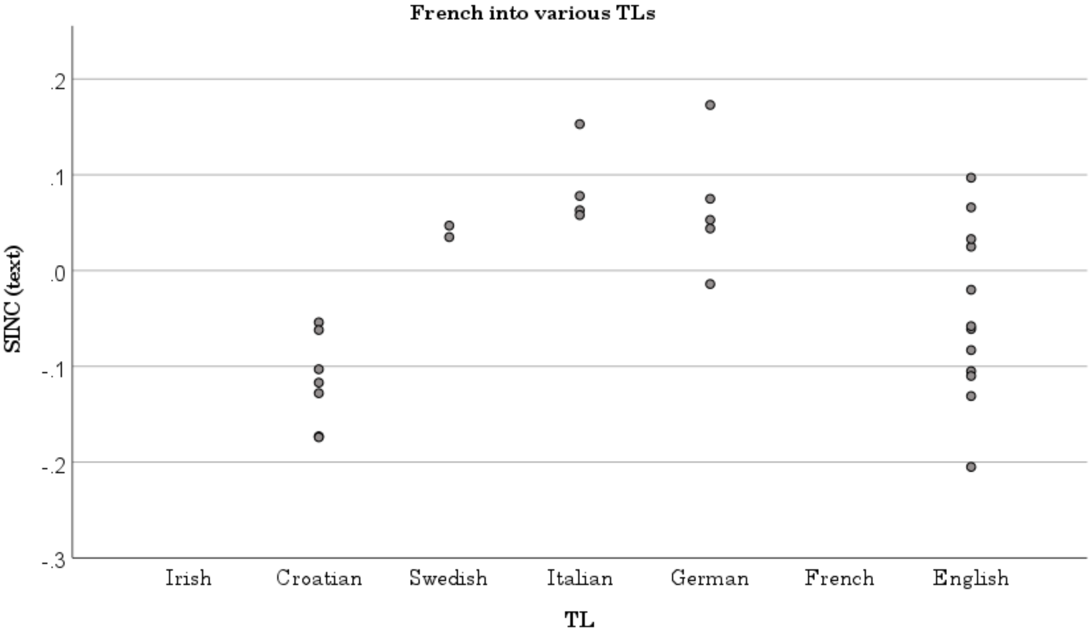


Figure 30: SINCtext scores for all translations from German

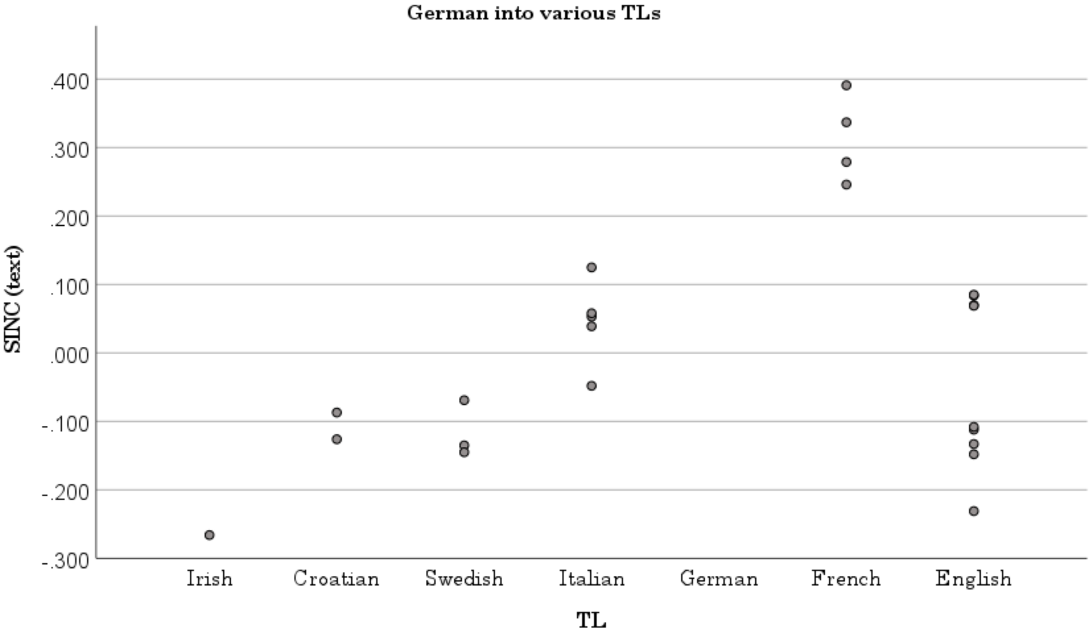


Figure 31: SINC(text) scores for all translations from Italian

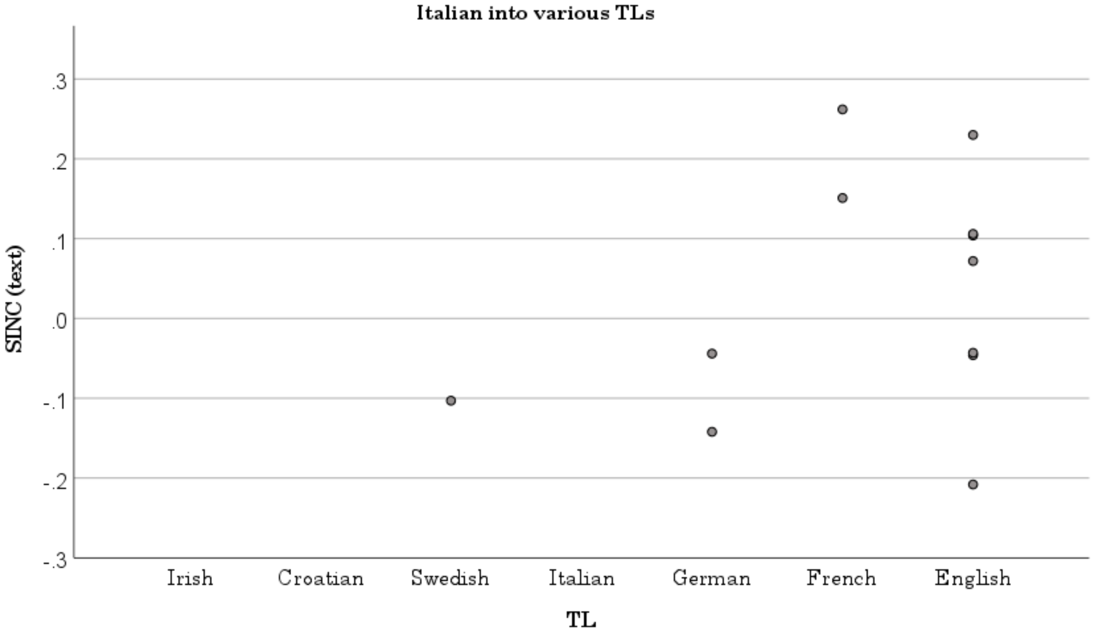


Figure 32: SINC(text) scores for all translations from Swedish

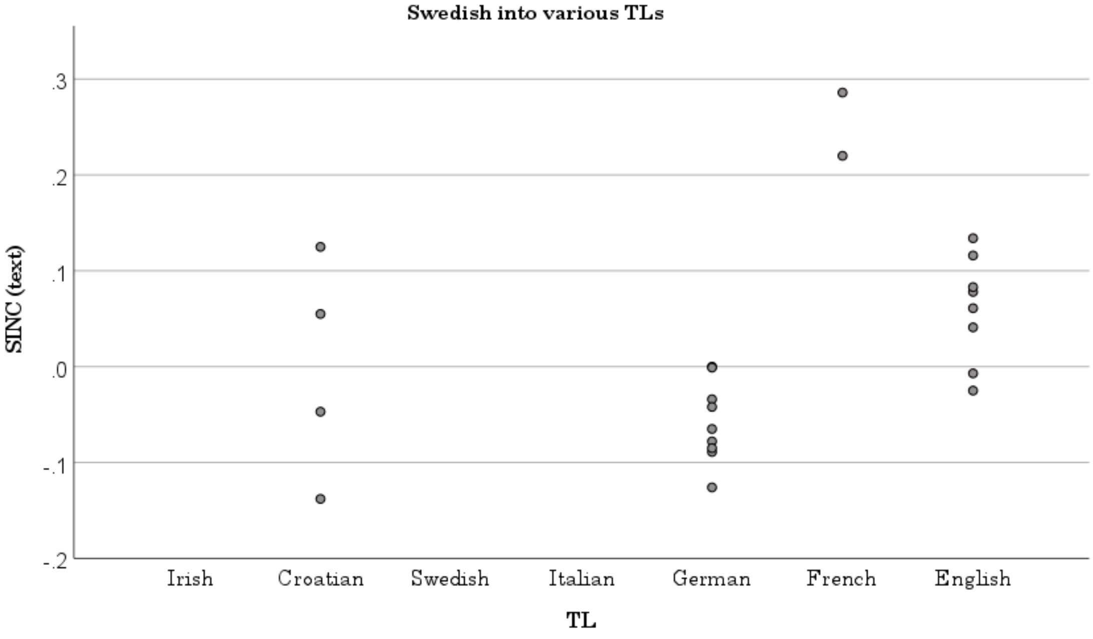


Figure 33: SINC(text) scores for all translations from Croatian

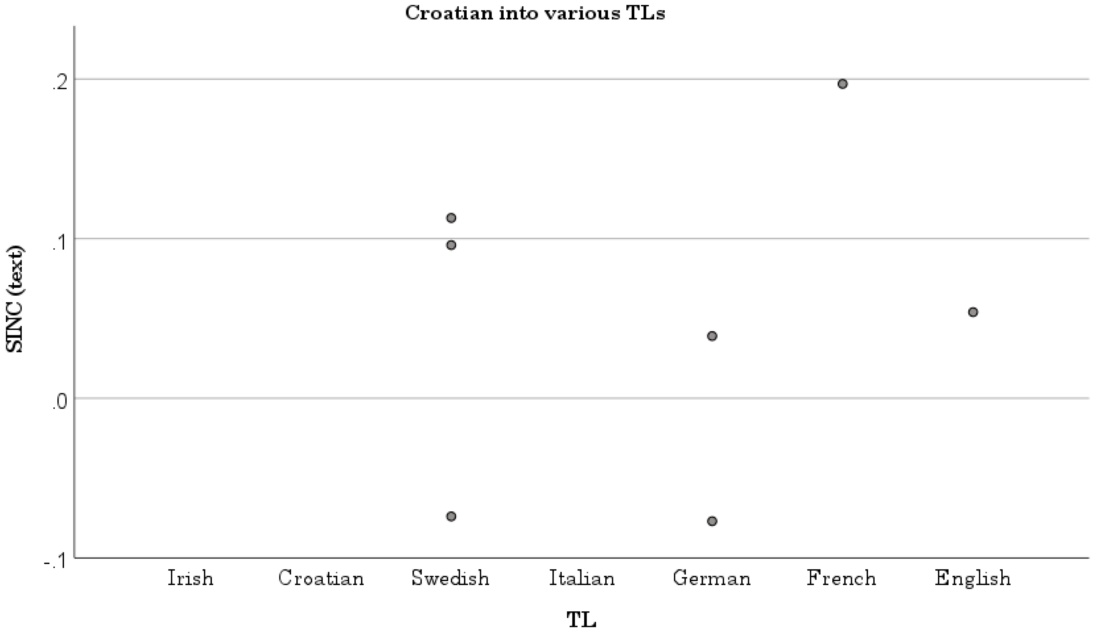
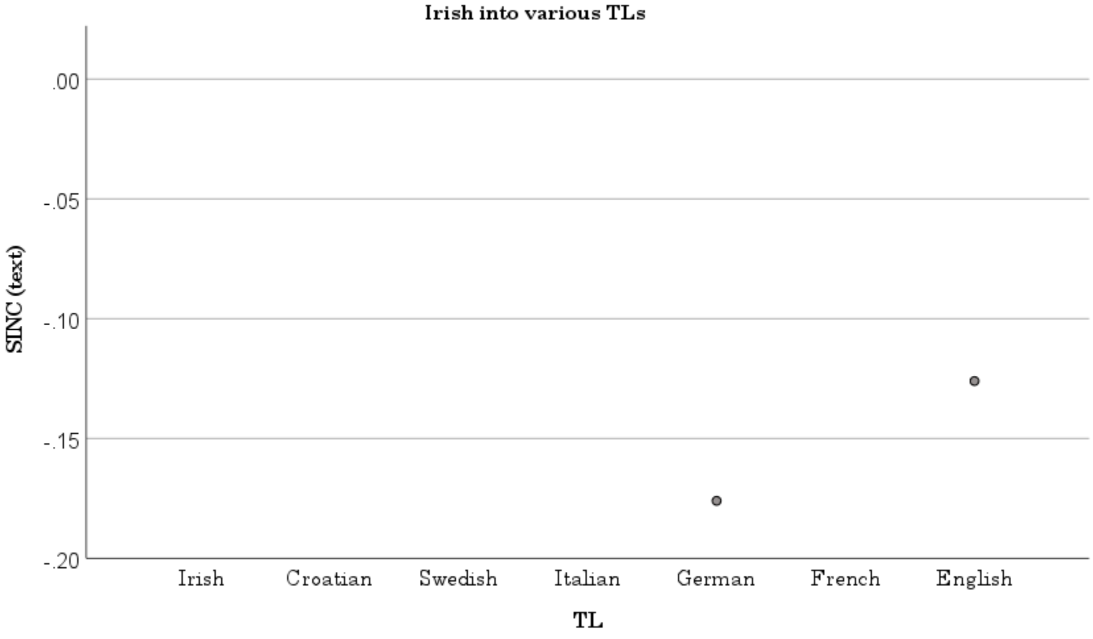


Figure 34: SINC(text) scores for all translations from Irish



6.4.5. Translated texts ranked by SINC$_{text}$ score

Table 21: All translated texts ranked by SINC (text) score

Rank	Translation	SL	TL	SINC (text)
1	La Mère de Dieu	German	French	+0.391
2	Le mort vivant	English	French	+0.366
3	Le grillon du foyer	English	French	+0.351
4	La Pantoufle de Sapho	German	French	+0.337
5	Le portrait de Dorian Gray	English	French	+0.328
6	La légende de Gösta Berling	Swedish	French	+0.286
7	La débâcle impériale - Juan Fernandez	German	French	+0.279
8	Un amant	English	French	+0.264
9	Feu Mathias Pascal	Italian	French	+0.262
10	Tristan	German	French	+0.246
11	Le crime de Lord Arthur Savile	English	French	+0.240
12	Les chasseurs de chevelures	English	French	+0.232
13	The Triumph of Death	Italian	English	+0.230
14	Au bord de la vaste mer	Swedish	French	+0.220
15	Enterrement à Thérésienbourg	Croatian	French	+0.197
16	Les trois hommes en Allemagne	English	French	+0.179
17	Der Amateursozialist	English	German	+0.175
18	Salambo	French	German	+0.173
19	Le magasin d'antiquités, Tome II	English	French	+0.168
20	I tre moschettieri, vol. I	French	Italian	+0.153
21	Une femme	Italian	French	+0.151
22	Dans l'abîme	English	French	+0.147
23	Das Bildnis des Dorian Gray	English	German	+0.143
24	Le magasin d'antiquités, Tome I	English	French	+0.142
25	The Story of Gösta Berling	Swedish	English	+0.134
26	Ja i moj sin	Swedish	Croatian	+0.125
27	Silvia ossia - La povera signorina - La gioventù provetta	German	Italian	+0.125
28	Zwei Städte	English	German	+0.122

Rank	Translation	SL	TL	SINC (text)
29	Invisible Links	Swedish	English	+0.116
30	Begravning i Teresienburg och andra noveller	Croatian	Swedish	+0.113
31	The Desire of Life	Italian	English	+0.106
32	A woman at bay	Italian	English	+0.104
33	Sretni vladar - Slika Doriana G	English	Croatian	+0.098
34	Mother of Pearl	French	English	+0.097
35	Återkomsten	Croatian	Swedish	+0.096
36	Der Weihnachtsabend	English	German	+0.088
37	The Chief Justice	German	English	+0.085
38	Royal Highness	German	English	+0.084
39	The Home; Or, Life in Sweden	Swedish	English	+0.083
40	I tre moschettieri, vol. III	French	Italian	+0.078
41	Christ Legends	Swedish	English	+0.078
42	Die Frau von dreißig Jahren	French	German	+0.075
43	The Patriot	Italian	English	+0.072
44	Blanche - The Maid of Lille	German	English	+0.070
45	Hermann	German	English	+0.069
46	A Tale of Brittany	French	English	+0.066
47	I tre moschettieri, vol. II	French	Italian	+0.063
48	La guerre des mondes	English	French	+0.063
49	Strife and Peace	Swedish	English	+0.061
50	I tre moschettieri, vol. IV	French	Italian	+0.058
51	Siddharta	German	Italian	+0.058
52	David Copperfield	English	Swedish	+0.056
53	Legende o Kristu	Swedish	Croatian	+0.055
54	On the Edge of Reason	Croatian	English	+0.054
55	Il Messaggio dell'Imperatore	German	Italian	+0.053
56	Bouvard und Pécuchet	French	German	+0.053
57	Vicomte de Bragelonne	French	Swedish	+0.047
58	Gegen den Strich	French	German	+0.044
59	Downstream	Swedish	English	+0.041
60	Hände	Croatian	German	+0.039

Rank	Translation	SL	TL	SINC (text)
61	I Buddenbrook	German	Italian	+0.039
62	Den Hemlighetsfulla ön	French	Swedish	+0.035
63	Under Sentence of Death; Or, a Criminal's Last Hours	French	English	+0.033
64	The Dream	French	English	+0.025
65	Das Buch vom Brüderchen	Swedish	German	+0.000
66	Unsichtbare Bande	Swedish	German	–0.001
67	Cnoc na nGabha II	English	Irish	–0.006
68	Married	Swedish	English	–0.007
69	Bübü vom Montparnasse	French	German	–0.014
70	Oliver Twist	English	Croatian	–0.019
71	Round the World in Eighty Days	French	English	–0.020
72	The Miracles of Antichrist	Swedish	English	–0.025
73	Christuslegenden	Swedish	German	–0.034
74	Cnoc na nGabha I	English	Irish	–0.034
75	Cnoc na nGabha III	English	Irish	–0.037
76	Ein Stück Lebensgeschichte und andere Erzählungen	Swedish	German	–0.042
77	The Intruder	Italian	English	–0.043
78	Ich und Er	Italian	German	–0.044
79	The House by the Medlar-Tree	Italian	English	–0.046
80	Pakleni Stroj	Swedish	Croatian	–0.047
81	La Morte a Venezia - Tristano - Tonio Kroger	German	Italian	–0.048
82	Germinal	French	Croatian	–0.054
83	Very Woman	French	English	–0.058
84	The Countess of Rudolstadt	French	English	–0.061
85	Izabrane novele - Guy de Maupassant	French	Croatian	–0.062
86	Gösta Berling: Erzählungen aus dem alten Wermland	Swedish	German	–0.065
87	Huset Buddenbrook	German	Swedish	–0.069
88	Händer	Croatian	Swedish	–0.074
89	Die Rückkehr des Filip Latinovicz	Croatian	German	–0.077
90	Der Sohn einer Magd	Swedish	German	–0.078
91	The Romance of a Poor Young Man	French	English	–0.083
92	Pastor Hallin	Swedish	German	–0.085

Rank	Translation	SL	TL	SINC (text)
93	Proces	German	Croatian	–0.087
94	Die Inselbauern; oder, Die Leute auf Hemsö	Swedish	German	–0.089
95	Il fantasma di Canterville e il delitto di Lord Savile	English	Italian	–0.103
96	20.000 milja pod morem	French	Croatian	–0.103
97	En kvinnas liv	Italian	Swedish	–0.103
98	The Devil's Pool	French	English	–0.105
99	The Merchant of Berlin	German	English	–0.108
100	Twenty Years After	French	English	–0.110
101	Frederick the Great and His Family	German	English	–0.112
102	Thérèse Raquin	French	Croatian	–0.117
103	Die Gotischen Zimmer	Swedish	German	–0.126
104	Preobrazaj	German	Croatian	–0.126
105	The Dirty Dust	Irish	English	–0.126
106	Gospođa Bovary	French	Croatian	–0.128
107	A Virgin Heart	French	English	–0.131
108	Gertrude's Marriage	German	English	–0.133
109	Amerika	German	Swedish	–0.135
110	Gösta Berling (HR)	Swedish	Croatian	–0.138
111	Una donna - Geschichte einer Frau	Italian	German	–0.142
112	Slottet	German	Swedish	–0.145
113	The Wish	German	English	–0.148
114	Blátha Bealtaine	English	Irish	–0.155
115	Put oko svijeta u 80 dana	French	Croatian	–0.173
116	Put u srediste zemlje	French	Croatian	–0.174
117	Dracula	English	Irish	–0.176
118	Cré na Cille	Irish	German	–0.176
119	A Cardinal Sin	French	English	–0.205
120	Farewell Love	Italian	English	–0.208
121	Joseph II. and His Court	German	English	–0.231
122	Eachtra Pheadair Schlemihl	German	Irish	–0.266

6.5. Discussion

6.5.1. Fixed TL analysis

Among the results for the fixed TL analysis, the hypothesized positive association between SL status and $\mathrm{SINC}_{\mathrm{text}}$ is observed in just one fixed TL subcorpus:

> Fixed TL subcorpora:
> German ($\tau = .575$; moderate positive association; $p < .001$)

In the fixed German TL subcorpus ($n = 23$), the text exhibiting the highest degree of syntactic interference is the English>German translation *Der Amateursozialist* ($\mathrm{SINC}_{\mathrm{text}} = +.226$), whose language pair includes the highest-status SL in the corpus, while the next seven highest $\mathrm{SINC}_{\mathrm{text}}$ scores in the fixed German TL subcorpus are translations from either English or French, the two high-status languages. Out of all German translations with a comparatively lower-status SL, only one (Croatian>German *Hände*) registers a positive $\mathrm{SINC}_{\mathrm{text}}$ score (+.039), indicating syntactic interference. The rest of the translations from comparatively lower-status languages register as syntactic normalization (i.e., $\mathrm{SINC}_{\mathrm{text}} < 0$), as predicted. In the fixed German TL subcorpus ($n = 23$), the text exhibiting the highest degree of syntactic normalization is the Irish>German translation *Cré na Cille* ($\mathrm{SINC}_{\mathrm{text}} = –0.176$), which is also the text with the lowest-status SL in the corpus. Thus, nearly all data points in the fixed German TL subcorpus neatly align with the outcomes predicted by subhypothesis I.

However, the only other statistically significant correlation detected among the fixed TL subcorpora is in the fixed English TL subcorpus, which contradicts the hypothesized positive association:

> Fixed TL subcorpora:
> English ($\tau = – .260$; weak negative association; $p = .018$)

In the fixed English TL subcorpus (n = 38), there is a perfectly even split between the number of texts exhibiting an overall degree of syntactic interference and those exhibiting an overall degree of syntactic normalization. The top five texts displaying the highest degrees of syntactic interference are translated from Italian (medium-status) or Swedish (low-status), while 12 out of the top 14 (85.71%) texts exhibiting the highest degrees of syntactic normalization are translated from French or German – the two highest-status SLs in the subcorpus. These results of the fixed English TL analysis unambiguously contradict the prediction that higher degrees of syntactic interference are associated with higher-status SLs.

Because only two statistically significant correlations are detected among the fixed TL subcorpora, with one confirming and one contradicting the hypothesized positive association, the results of this study's fixed TL analysis fail to provide evidence supporting subhypothesis I.

6.5.2. Fixed SL analysis

Among the results for the fixed SL analysis, the hypothesized negative association between TL status and $\text{SINC}_{\text{text}}$ is not detected in any of the fixed SL subcorpora. On the contrary, two statistically significant positive associations are detected:

Fixed SL subcorpora:
English ($\tau = .708$; strong negative association; $p < .001$)
Swedish ($\tau = .354$; weak to moderate negative association; $p = .016$)

In the fixed English SL subcorpus, 15 of the 17 (88.24%) translations exhibiting composite syntactic interference are translated into French or German, the two highest-status TLs in the subcorpus. This trend contradicts subhypothesis II, as English would be expected to induce less interference in translations into the two languages closest to it in status compared to translations into lower-status languages. Moreover, five of the seven (71.43%) translations exhibiting the highest degrees of syntactic normalization are English>Irish translations. This finding is highly unexpected, as translations from the highest-status SL into the lowest-status TL would be expected to reflect the highest degrees of syntactic interference in the corpus. While it is possible that Irish translators of English texts counteracted the power dynamics between these languages by exaggerating typical Irish syntactic structures, it is more likely that the clustering of English>Irish translations in these results indicates a shortcoming of the SINC methodology, as discussed further in Section 6.6.

In the fixed Swedish SL subcorpus, six of the 11 (54.55%) translations exhibiting composite syntactic interference are translated into English, while two (2/11, 18.18%) are the lone French translations in the subcorpus. This trend contradicts subhypothesis II, as the comparatively lower status of Swedish would be predicted to induce an overall effect of syntactic normalization when translated into the high-status TLs. Two of the translations of Swedish source texts exhibiting syntactic interference are Croatian translations; these findings may be attributed to the comparatively higher (yet only slightly) status of the SL, however the subcorpus' other two Croatian translations exhibit syntactic normalization, thereby negating the emergence of a potential pattern for the Swedish>Croatian language pair.

The expectation that translations from Swedish into the subcorpus' comparatively higher-status languages – English, French, and German – demonstrate syntactic normalization holds true for 10 of the 19 (52.63%) translations into these three TLs, including eight of the nine (88.89%) German translations. As already indicated, however, only two of the eight (25%) English translations and neither of the French translations exhibit syntactic normalization. It is unclear why the subcorpus of translations from Swedish exhibits this trend. Because the only statistically significant findings in the fixed SL analysis contradict the expected outcome, the results of the fixed SL analysis fail to produce evidence supporting subhypothesis II. Taken together, the results of the fixed TL and SL analyses do not support the study's hypothesis that syntactic interference tends to increase as SL status increases relative to TL status. Having failed to confirm the hypothesis via the fixed TL and SL analyses, the study now examines the data on a text-by-text basis in order to ascertain other noteworthy trends between SL status and syntactic interference/normalization in translation.

6.5.3. Ranked-text analysis

As shown in Table 21, ranking all 122 translations in the corpus according to $\mathrm{SINC}_{\mathrm{text}}$ does not reveal any consistent relationship between SL status and syntactic interference. However, it is rather striking that 15 of the top 16 (93.75%) texts with the highest degrees of syntactic interference ($\mathrm{SINC}_{\mathrm{text}} > 0$) are translations into French, with the other being a translation into English – the other high-status TL. While all possible SLs are represented among these 16 translations (with the exception of Irish), nearly half (7/16; 43.75%) are English>French translations. It is not apparent why French translations – or English>French translations, in particular – exhibit such high degrees of syntactic interference. Translations exhibiting the highest degrees of syntactic normalization ($\mathrm{SINC}_{\mathrm{text}} < 0$) do not offer much clarification either. While four out of the ten (40%) lowest $\mathrm{SINC}_{\mathrm{text}}$ scores are translations into English, the highest-status TL, five out of the ten (50%) are translations into Croatian (a low-status TL) or Irish (a very low-status TL), and the same amount (5/10; 50%) are translations from high-status SLs.

It is worth noting that Irish, the lowest-status language, is present in the language pairs of four of the ten (40%) most syntactically normalized translations in the corpus, despite only eight of the 122 total translations (6.56%) being either from or into Irish. Contrary to initial predictions, three of these highly normalized translations are translations into Irish, with the lowest $\mathrm{SINC}_{\mathrm{text}}$ score in the entire corpus achieved by the lone German>Irish translation (*Eachtra Pheadair Schlemihl*). Given the concerted efforts to revitalize the Irish language, it is possible that Irish translators' attitudes toward their language led them to prioritize and even exaggerate TL syntactic conventions, deliberately undermining their higher-status SLs. Despite their direct contradiction of the study's hypothesis, these results offer possible insights into the complex interplay between language status and language prestige and its effect on translation strategies. The historical relationship between English and Irish best illustrates this complexity, which will be explored further in the conclusion to the thesis.

An overview of the total number of translations exhibiting syntactic interference and syntactic normalization in comparison to their language pairs' status differentials further discredits the hypothesis (see Worksheet 5.3.). Of the 122 translations, 65 texts

(53.28%) exhibit an overall degree of syntactic interference ($\mathrm{SINC}_{\mathrm{text}} > 0$), with 33 of these translations (33/65, 50.77%) having a higher-status SL relative to their TL and aligning with the hypothesized association. On the contrary, a nearly equal number of translations exhibiting syntactic interference (32/65, 49.23%) have a comparatively lower-status SL and contradict the expected results. There are 57 translations (57/122, 46.72%) exhibiting an overall degree of syntactic normalization ($\mathrm{SINC}_{\mathrm{text}} < 0$), with 32 translations (32/57, 56.14%) having a higher-status TL relative to their SL and aligning with the hypothesized association. Once again, a comparable number of translations exhibiting syntactic normalization (25/57, 43.86%) have a comparatively lower-status TL and contradict the predicted outcome. From the ranked-text analysis, it is further apparent that the results do not support the hypothesized associations between 1) comparative SL status and syntactic interference, or 2) comparative TL status and syntactic normalization.

6.5.4. Microstructural analysis

In order to further dissect the findings with respect to the relationship – or lack thereof – between the variables as well as the newly developed SINC methodology, it is useful to identify specific POS n-grams for closer examination by considering known structural differences between the various SLs and TLs.

The second-lowest $\mathrm{SINC}_{\mathrm{text}}$ score is captured by the German>English translation *Joseph II. and His Court* ($\mathrm{SINC}_{\mathrm{text}} = –0.231$). As a translation from the comparatively lower-status German into English, the highest-status language, this high degree of syntactic normalization is generally aligned with the study's hypothesis. Consideration of prominent structural differences between the conventional word orders of German and English proves useful in identifying and interpreting noteworthy data points on the microstructural level (i.e., individual $\mathrm{SINC}_{\mathrm{gram}}$ scores). For instance, in German sentences or clauses with two verbs, such as the modal verb (*können*) and the main verb (*lesen*) in the following fabricated example, the second verb must be placed at the end of the sentence:

Wir *können* das Buch *lesen*.
[we] [can] [the] [book] [read]

In English, the modal verb and the action verb are positioned consecutively:

We *can read* the book.

Predictably, a comparison of the RF distributions of POS 2-grams for the English and German comparable subcorpora reflects this fundamental word-order difference. The most commonly occurring POS 2-gram in the comparable English subcorpus is *noun-punct* (RF = .0569), which perhaps indicates that nouns are the most common POS to end a sentence or clause in English. On the contrary, in the comparable German subcorpus, *noun-punct* is the fourth-most common POS 2-gram and has a substantially lower RF (.0379) than in the comparable English subcorpus, whereas *verb-punct* is the second-most common POS 2-gram (RF = .0402). The finding that *verb-punct* occurs more frequently than *noun-punct* in German may thus indicate that two-verb sentences and clauses are rather common in German. Conversely, the more frequent usage of *noun-punct* is a syntactic feature distinguishing original English texts from original German texts.

The RF of the POS 2-gram *noun-punct* in *Joseph II. and His Court* is even higher (X = .0693) than its corresponding value in the comparable TL subcorpus (T = .0569), moving in the opposite direction of the corresponding value in the comparable SL subcorpus (S = .0379). Thus the exaggerated use of the POS 2-gram *noun-punct* in *Joseph II. and His Court* relative to comparable English texts constitutes syntactic normalization on a microstructural level ($\mathrm{SINC}_{\mathrm{text}}$ = – 0.231). Given the status differential between German and English, the hypothesis anticipates that similar outcomes characterize the most frequent POS n-grams – and, in turn, the most heavily weighted $\mathrm{SINC}_{\mathrm{gram}}$ calculations – for the text in question, such that its aggregate $\mathrm{SINC}_{\mathrm{text}}$ score amounts to a negative value, which turns out to be the case.

A microstructural analysis of POS n-grams in the English>German translation *Der Weihnachtsabend* – translated in the opposite direction – provides two complementary data points. The POS 2-grams *noun-punct* and *verb-punct* in the translation reflect slight degrees of syntactic interference (*noun-punct* SINC$_{\text{gram}}$ = +0.173; *verb-punct* SINC$_{\text{gram}}$ = +0.035). These findings suggest that English syntactic conventions slightly influence the use of these two POS 2-grams in the translation. Overall, the text exhibits a marginal degree of syntactic interference (*Der Weihnachtsabend* SINC$_{\text{text}}$ = + 0.088), aligning it with the expected outcomes of translations from English, a high-status language, into German, a medium-status language.

Two closely related languages in the corpus – French and Italian – also merit close consideration. Non-null subjecthood (i.e., the mandatory placement of a subject before a verb) differentiates French from the rest of the Romance language family, which otherwise exhibits a high degree of structural uniformity (D'Alessandro 2021, 311-312). This structural difference is evident in the SINC$_{\text{gram}}$ scores for the POS 2-gram *pron-verb*: in all four French>Italian translations (all different volumes of *I tre moschettieri*), the use of preverbal pronouns reflects interference (*pron-verb* SINC$_{\text{gram}}$ > 0). It appears that the mandatory use of preverbal pronouns in French leads to an increase in the use of optional preverbal pronouns in French>Italian translations compared to original Italian texts. It is quite possible that the status differential between these languages has strengthened the influence of this structural priming. Consider the following sentence taken from the French>Italian translation *I tre moschettieri, vol. I*:

> Io diffonderò la parola, mio caro, siate tranquillo.
> [I] [will spread] [the] [word], [my] [dear], [be] [calm].

The pronoun *io* is optional in this sentence, and while it is possible that the translator simply chose to include it for the sake of emphasis, this choice may also be attributable to syntactic interference from the French source text. Of course, it is impossible to make any such determinations for individual examples, but in the aggregate, the SINC$_{\text{gram}}$ scores clearly indicate that the French>Italian translations reflect syntactic interference in their usage of the POS 2-gram *pron-verb*. Though the SINC$_{\text{gram}}$ scores for preverbal pronouns across all French>Italian translations seem relatively meager (*pron-verb*

$\text{SINC}_{\text{gram}} < +0.3$ for all four texts), they are weighted heavily in SINC-*n* scores due to their comparatively high RFs in the translations.

Swedish also has a rare (and, in the present corpus, unique) structural feature worth examining. Like most of the other languages in this multilingual corpus, English places definite articles before nouns (e.g., "the dog"). In lieu of a definite article corresponding to "the", Swedish adds suffixes to the ends of nouns, such as *-en* ("the") to *hund* ("dog"):

Hund*en* jagar boll*en*.
[the dog] [chases] [the ball]

This three-word sentence would then be converted into the POS 3-gram *noun-verb-noun*. Although both of the nouns in the sentence above possess definiteness, their definite articles constitute suffixes rather than distinct word forms, so the POS-tagger does not register any determiners. Still, Swedish has indefinite articles (*en hund* or "a dog") and other definite articles (*den hunden* or "that dog") that are standalone word forms placed before nouns. This basic syntactic distinction between English and Swedish creates a natural expectation that the RF of, for example, the POS 3-gram *det-noun-verb* is lower in the comparable Swedish corpus than in the comparable English corpus, as Swedish is slightly less structurally amenable to prenominal determiners than English.

Bearing in mind this structural difference, the POS 3-gram *det-noun-verb* in the English>Swedish translation *David Copperfield* provides noteworthy results. As expected, the POS 3-gram's RF is higher in the comparable English subcorpus ($S = 0.0033$) than in the comparable Swedish subcorpus ($T = 0.0021$). Given this difference, the study's hypothesized association produces the expectation that the higher-status SL would induce the increased use of this syntactic construction in translation relative to its typical use in comparable Swedish texts. However, the opposite occurs: the RF of the POS 3-gram *det-noun-verb* in the English>Swedish translation *David Copperfield* is even lower ($X = 0.0018$) than its corresponding value in the comparable Swedish subcorpus, meaning that this POS sequence reflects syntactic normalization (*det-noun-*

verb $\text{SINC}_{\text{gram}}$ = –0.2861). On the whole, however, the translation's overall composition reflects syntactic interference ($\text{SINC}_{\text{text}}$ score = +.056).

There are myriad other possibilities for a microstructural analysis of the results, given the multiplicity of POS n-grams and $\text{SINC}_{\text{gram}}$ calculations across all 122 translations in the corpus. Although no POS 1-grams or POS 4-grams were analyzed in this section, knowledge of the selected languages' structural necessities or tendencies may benefit the interpretation of these granular results as well. Nonetheless, the methodology has been designed so as to draw meaningful conclusions based on aggregated results (i.e., $\text{SINC}_{\text{text}}$ scores). To the extent that the newly devised SINC methodology adequately reflects translated texts' composite levels of syntactic interference and normalization, the results of the study do not provide evidence supporting the hypothesis that there is a positive association between comparative SL status and syntactic interference.

6.6. Limitations

There is an inevitable tension between this study's two overarching aims: 1) to produce and interpret data with respect to the study's hypothesis, and 2) to assess the validity of the newly devised methodology. Although the SINC formula perhaps proves useful in isolating the effects of specific SL syntactic conventions on translations, the novelty of its characterization of translations' composite syntactic makeups – in the form of its comparisons of various POS n-gram RF distributions – as interference or normalization requires further scrutiny, and is thus the primary focus of this section.

Perhaps the most glaring indication that the SINC methodology requires additional modification is the observable tendency for the $\text{SINC}_{\text{text}}$ scores of translations in a given language pair to cluster in the study's various analyses. This trend is rather apparent in the ranked-text analysis (see Table 21), where the top 16 texts are French translations, with 43.75% of them being English>French translations. Moreover, consecutive rankings throughout the list of all 122 translations often reflect the same language pair, and translations in the same language pair generally appear to be ranked near each other. The clustering effect is also evident in, for instance, the scatter plot in Figure 32, which shows that there is very little overlap in the $\text{SINC}_{\text{text}}$ scores for Swedish>German, Swedish>French, and Swedish>English translations. At first glance, these patterns seem to suggest that translations in the same language pair are highly uniform in their levels of syntactic interference/normalization. However, given that there is no consistent association between SL status and syntactic interference in this study, it is more likely that this clustering is attributable to the SINC methodology's inability to fully control for structural differences between languages in its attempt to compare syntactic inference across translations in diverse language pairs.

The limitation of $\text{SINC}_{\text{gram}}$ scores between –1 and +1 invites additional critique from a theoretical perspective. Even if the mechanism for preempting extreme values is warranted, it may be argued that coefficients beyond this narrow range are still relevant to the appraisal of syntactic interference/normalization in translation. A more nuanced statistical approach may devise a strategy for incorporating such values while still negating the possibility of outliers skewing higher-level SINC calculations. For instance, the range of possible $\text{SINC}_{\text{gram}}$ scores may be expanded on the basis of scores' standard deviations or distributions more generally.

POS tagging errors may also skew results, especially given the anticipated differences in accuracy between the project's selected languages. In particular, antiquated spelling conventions in older texts may prove unrecognizable to the spacy-udpipe annotator. A review of the English>Swedish translation *David Copperfield* exemplifies this problem; for example, the archaic spelling of the Swedish preposition *av* ("*af*" in the text) is frequently tagged incorrectly as a noun or a proper noun. This issue likely also arises with dialect variations such as in Irish, whose three major dialects – Connacht, Munster, and Ulster – differ not only in orthography but also in syntax. Intuitively speaking, the more POS n-grams are incorrectly identified, the more these errors dilute (i.e., decrease) the RFs of correctly identified POS n-grams. Therefore, it may be the case that SINC scores of translations with many more POS n-gram identification errors – perhaps due to certain conditions like when a text or language has a notably lower POS tagging accuracy – are uniformly affected in one or the other direction (i.e., interference or normalization), depending on the language pair and text. If comparable subcorpora and translations have significantly different levels of incorrectly identified POS n-grams, or if their incorrectly identified POS n-grams are distributed in radically different yet internally consistent ways (e.g., frequently occuring proper nouns as misidentified as adjectives), SINC scores may be highly unreliable. Despite the proclaimed universality of the POS tagset, differences in tokenization and potentially also in POS identification may seriously undermine cross-lingual comparisons of POS n-grams; for instance, *l'homme* in French and *l'amore* in Italian are both recognized as one token, despite the presence of the definite article, which is recognized as a separate token and tagged as a determiner (*det*) in other languages such as English and German. Relatedly, although the Universal Dependencies POS tag set is explicitly designed for the purpose of enabling theoretically valid comparisons across syntactically diverse languages, not all 17 POS classes are used for all languages (De Marneffe et al. 2021, 261). The POS tag *x* – used as a miscellaneous tag for, e.g., foreign words – occurs in the comparable French subcorpus but not in the comparable Swedish corpus for some unknown reason.

The SINC methodology may be considered robust in the sense that it involves comparisons for all POS n-grams in a translated text instead of cherry-picking language-specific syntactic features. Still, syntax cannot be fully represented in terms of POS sequences. POS n-grams may be considered a "(shallow) syntactic structure" (Lembersky et al. 2012, 822), and the universal POS tags used in this study are intentionally "coarse-grained" (De Marneffe et al. 2021, 261). As discussed in the literature review, the results reinforce the notion that the universality of a translation feature's operationalization is perhaps inversely proportional to the depth of analysis it permits. Teich (2003, 226-227) refers to POS tagging as a "fairly reliable technique of linguistic annotation" yet a "rather shallow kind of annotation", but still affirms that this type of annotation allows for "the extraction of instances of particular syntactic patterns." Limitations in POS n-gram sizes (up to 4-grams) may also preclude the analysis of SL interference on long-range word reorderings, which constitute vital syntactic information particularly for languages such as German (Popović and Ney 2011, 682). Moreover, as noted by Chlumská (2018, 107), previous research indicates that typological differences between languages may prevent direct cross-lingual comparisons between POS n-grams of the same length. A more fine-grained measurement of syntactic interference in translation might leverage machine learning techniques to examine more linguistically and computationally complex syntactic dependencies, as reflected by Sidorov's (2019) "syntactic n-grams" concept. In fact, he strongly insists that POS n-grams reflect not syntactic information but *morphosyntactic* information (ibid., 48).

The universal POS tag set in this study discounts the additional semantic and syntactic information encoded in the MT evaluation measures provided by Popović and Ney (2011, 30), such as verb tense forms and cases. The study's lack of a systematized method of accounting for each text's tense may thus skew results. For instance, some of the selected languages rely more heavily on auxiliary verbs for conjugations in the past tense, thereby artificially inflating the frequencies of POS n-grams involving auxiliary verbs. Likewise, the lack of a systematic method of categorizing and proportionately sampling texts from different narrative modes (e.g., first-person or third-person narratives) may also skew the results. Focusing on first-person pronouns, for example, Maia (1998, 1) selects texts "contain[ing] a large number of natural monologues and dialogues" in order to facilitate cross-lingual comparability.

Lastly, a significant methodological limitation is the reliance on comparisons between translations and comparable SL subcorpora instead of their specific source texts. Although this limitation applies to all three studies included in this thesis, it may be considered particularly pertinent to measurements of syntactic interference. This study is the only one of the three studies contained in this thesis whose operationalization makes direct comparisons between features in translated texts and SL texts. As Chlumská (2018, 106) contends, by only using comparable corpora and excluding comparisons between translations and their respective source texts, it may not be possible to surmise the explanatory power of any variables beyond a mere description of results. Teich (2003, 10) also argues forcefully that parallel corpora are absolutely necessary for the characterization of syntactic interference.

6.7. Chapter conclusion

The study presented in this chapter did not produce evidence supporting the hypothesis that SL status is positively associated with syntactic interference in translation, or its inverse prediction that TL status is positively associated with syntactic normalization in translation. Syntactic interference/normalization was operationalized using a novel methodology based on the comparison of RF distributions of POS n-grams between translations and comparable SL and TL corpora. There were several statistically significant findings among the fixed TL and fixed SL analyses. As predicted in subhypothesis I, there was a positive association ($\tau = .299$; $p = .032$) between SL status and syntactic interference for translations from various SLs into German. On the contrary, SL status was negatively associated with syntactic interference for English translations ($\tau = –.260$; $p = .018$). Moreover, the only statistically significant associations detected in the fixed SL analysis contradicted subhypothesis II: there were positive associations between TL status and syntactic interference in the fixed English SL subcorpus ($\tau = .708$; $p < .001$) and the fixed Swedish SL subcorpus ($\tau = .354$; $p = .016$). These results should be interpreted with caution, however, in light of the limitations arising from the exploratory nature of the study's novel methodology.

7. Paratextual foreignization

7.1. Chapter introduction

Having measured the interplay between language status and selected linguistic markers of SL influence on the lexical and syntactic composition of target texts, the project now turns toward the potential impact of comparative SL status on translations' paratextual features. In a fundamental sense, the paratextual dimension is different from the previous two. Lexical and syntactic manifestations of SL influence may be conceptualized as strictly linguistic features of translation found in the body of the text itself, whereas paratextual features constitute categorically distinct features of translations, as explored later on.

Before proceeding, it is necessary to establish a definition of paratexts that is workable for translation studies research. The majority of theoretical and empirical inquiries into paratexts – whether oriented toward translations or other text types – build upon the work of Genette (1997). Batchelor (2018, 12) summarizes Genette's framework of paratexts, which conceptualizes a paratext as "any element which conveys comment on the text, or presents the text to readers, or influences how the text is received." Among the key characteristics of a paratext are its creator (its sender) and whom it was written for (its addressee). There is, of course, a wide variety of potential senders and addressees, all of whom approach the text with different motivations and expectations. The placement of translations and translators within this framework is thus a key theoretical concern in establishing a research paradigm for the use of paratexts in translation.

Whereas Genette primarily characterizes translations as merely paratexts of their respective source texts, Batchelor's (2018, 142) framework distinguishes translations as standalone texts in their own right, containing their own sets of paratextual elements and relations. She defines a paratext as "a consciously crafted threshold for a text which has the potential to influence the way(s) in which the text is received," where a text is "any written or spoken words forming a connected piece of work" (ibid., 142). This broad definition encapsulates such diverse phenomena as in-text footnotes, translators' prefaces, interpreters' body language, literary critics' book reviews, as well as posters and trailers for newly released films – all of which are considered valid as objects of study in research on paratexts. Within the present study, of course, the paratexts of interest are those appearing in the corpus' various translated texts.

How then do such paratexts relate to the dichotomy of SL- and TL-oriented translation strategies, which underpins the overarching project's main research question? Batchelor (2018, 32-33) notes the strong link between translation scholars' inquiries into paratexts and Venuti's (1995) concept of translator visibility, as will be further demonstrated in the following section. As observed in the introduction to the thesis, Venuti's dichotomy between foreignization and domestication hinges on the notion of translator visibility, which involves both the linguistic composition of the translation itself as well as any features that are ancillary to the main text yet contained in the same volume. The latter of these categories may naturally be subsumed under Batchelor's definition of paratexts, as these paratexts "influence the way the text is received" by making overt the translator's role in facilitating the text's production, thus raising the visibility of the text's status as a translation.

In this manner, translator-produced paratexts instantiate SL influence on the target text via their mere indication of the text's status as a translation, drawing attention to its relational property to a source text. Thus, the paratextual dimension of SL influence on translations is contingent on the perception of a text's relational property – i.e., its fundamental relation to a source text and hence status as a translation (see Toury's [2012] Source-Text Postulate). This characteristic stands in contrast to the purely linguistically-oriented traces of SL influence located within the body of the text itself. Features such as footnotes, for example, "do not belong to the main argument and are therefore placed outside the main verbal sequence" (Buts and Jones 2021, 311). The present study therefore borrows Venuti's framing and conceptualizes SL influence on translations' paratextual features as *foreignization* rather than *interference*, distinguishing on theoretical grounds the present study from those comprising the previous two chapters.

7.2. Related works

As noted by Batchelor (2018, 34), theoretical links between the use of paratexts and translator visibility are recurrent in the literature and sometimes considered in relation to "intercultural power dynamics", with many works taking a decidedly prescriptive stance. Yuste Frías (2012, 132) advocates for paratextual spaces as primary sites of translator visibility. McRae (2006, 39-41) and Podlevskikh Carlström (2022, 64) supplement their empirical findings with emphatic calls for greater translator visibility via paratexts, presuming target audiences' general disdain for these features as impediments to their use. Hermans (2007, 23-24) extols paratexts' potential to increase the visibility of translated texts' status as such, thus unambiguously distinguishing them from non-translated texts. In light of these forceful and polemic calls, paratexts in translation may be considered highly illustrative markers of the general inclinations toward translator visibility in different cultural and historical contexts (see Coldiron 2012). It is therefore possible that paratextual features of translations quite plainly reflect the power imbalances between languages and cultures that Venuti suggests govern translator visibility. Most case studies of paratexts in translation, however, approach such interlingual and intercultural power dynamics only indirectly, focusing instead on the ways in which paratexts reflect translators' ideological stances.

In one of the most widely cited works on paratexts in translation, Tahir-Gürçağlar (2002, 44) aims to further historical translation research's objective of "explor[ing] the socio-cultural contexts in which translated texts are produced and received" by investigating the ways in which paratexts reflect a target culture's definition of and norms surrounding the practice of translation. Though not overtly linked, much of her approach aligns closely with Venuti's concept of translator visibility. She examines a pair of Turkish literary translations commissioned by private and government-operated publishers in the mid-20th century, differentiating the ways in which their paratexts reflect various perspectives on the appropriateness of translator visibility in light of the publishers' respective positionings within the Turkish political system (ibid., 47). Toledano Buendía (2013) discusses the two primary functions of translator's notes – explanatory and commentary – via an examination of Spanish translations of English fiction from the 18th and 19th centuries. She concludes that, regardless of the specific function of the translator's note, this type of paratext serves to raise the visibility of the text's status as a translation (Toledano Buendía 2013, 161). A plethora of other research centers on close readings of paratexts in literary translations that aim to uncover translators' ideological motivations (see Martin 2006; Alvstad 2012; Pellatt 2013). However, such research is necessarily limited to small selections of texts, making it difficult to categorize cross-cultural and cross-lingual patterns of translator paratexts in the aggregate.

Large-scale, systematic research on the use of paratexts in translation is relatively scant. Dimitriu (2009, 195) examines translators' prefaces in more than 65 fiction and non-fiction Romanian translations in an attempt to categorize their primary aims and functions. Regardless of function, these paratexts reveal valuable insights into translators' methods and strategies, and may therefore "help build bridges between the theory and the practice of translation" while also serving as "palpable proofs of the translators' visibility" (ibid., 230). Using a similar framing, McRae (2006) compiles hundreds of literary translations in a variety of major languages, finding that only one-fifth contain translator prefaces. Indeed, the existing literature finds translators' prefaces to be generally uncommon, indicating a pervasive lack of translator visibility in this regard (Bilodeau 2019, 66).

Paloposki (2010) assembles a corpus of nearly 100 books translated from a variety of genres and SLs into Finnish around the beginning of the 20th century. The study primarily looks for patterns of footnote usage across individual books or translators (ibid., 98). The only trends related to SLs that Paloposki (ibid., 99) identifies are found in translations from Swedish and German, which contain notably fewer footnotes. The author initially speculates that this may be attributable to a perceived closeness between the source and target cultures, resulting in less pressure to use footnotes to provide additional context for readers. However, upon closer inspection, it seems that this trend is more likely attributable to the conventions of "folk literature" in translation, which typically entail "the levelling out of cultural specificity" (ibid., 99).

Podlevskikh Carlström (2022) tallies and analyzes different types of translator-attributed paratexts – including translator prefaces as well as in-text notes and/or commentary – in over 80 Swedish translations of post-Soviet Russian novels. She argues that footnotes increase translator visibility in texts regardless of whether they are explicitly signed by the translator or implicitly attributable to the translator (Podlevskikh Carlström 2022, 57-58). This systematic study speculates which factors affect the variability of paratextual translator visibility. Considering "high-brow" literature to be more prestigious than "popular" literature, Podlevskikh Carlström (ibid., 50) hypothesizes that translators' use of paratexts is heavily dependent on the prestige of the source text. However, the results of the study do not demonstrate a clear difference in the general use of paratexts – and thus translator visibility – among source texts differing in prestige (ibid., 63). The one type of paratext for which there is an unambiguous pattern is the footnote (translator's note) – although the author does not speculate as to why this particular paratext may be more commonly found in high-brow literature than in popular fiction (ibid., 63).

Footnotes in particular are considered one of the most visually imposing types of paratext deployed in translation. Newmark (1988, 92) cautions against the use of footnotes in translation, citing their ostensibly disruptive effect on the reading experience. That footnotes are "extremely visible" compared to other paratexts is perhaps why their appropriateness as a translation strategy is frequently contested (Paloposki 2010, 88). While much work on paratexts in translation focuses on translator prefaces, it may be reasoned that prefaces are more easily ignored or overlooked by readers given their placement outside the main text. Given their more prominent placement with respect to the main text, translator footnotes may therefore serve as an ideal marker of translator visibility on the paratextual level, as further depicted in the study's methodology outlined in the following section.

Regardless of the specific type, paratexts have been frequently tied to political ideologies of senders with respect to authors of source texts or source cultures in specific contexts. Absent from the literature, however, is a systematic investigation of the frequency of translator paratexts across translations representing a range of language pairs. Notably, McRae's (2006) large-scale exploration of the presence of translator prefaces in translated literature does include translations from 29 SLs into English. Still, the tendencies of English-language translators of various SLs regarding the use of these paratexts are ancillary to her classification of common patterns among these paratexts' perceived functions and contents.

Despite frequent references to paratexts as instantiations of Venuti's translator visibility, translation scholars have yet to directly examine translator visibility in the paratextual dimension as a function of SL/TL power imbalances – arguably the core concept beneath Venuti's assertion of low translator visibility in English translations. Moreover, it is difficult to make even tentative claims about the relationship between language status and paratextual foreignization, as previous studies generally fail to uncover consistent patterns of translator paratexts for their respective language pairs or TLs. Against this background, this chapter conducts an empirical investigation of the potential correlation between comparative SL status and paratextual foreignization in translation.

7.3. Methodology

7.3.1. Defining translator footnotes (TFNs)

The previous section suggested that footnotes in particular serve as a suitable type of paratext to represent paratextual foreignization in translation. The diverse paratexts covered by Batchelor's (2018, 172) maximally inclusive definition of the term necessarily differ in their likelihood of being observed by the reader in the first place. Given their close proximity to the main text itself, footnotes and endnotes rank among the most visible (in the literal sense) types of paratexts (Pellatt 2013, 2). These similar paratextual elements therefore serve as prominent indicators of translators' visibility.

As noted by Paloposki (2010, 94), however, it is not always apparent whether a footnote may be attributable to the translator or the author of the source text. Podlevskikh Carlström (2022, 47) likewise emphasizes the necessity of distinguishing peritexts "that belong to the source text and those that were created for the translation." Pym (2004, 70-73) refers to this conundrum as "first-person displacement" – the "I" in paratexts is typically understood to refer to the source text's author unless otherwise noted. To the extent that footnotes – as well as their close relatives, endnotes – are attributable, whether explicitly or implicitly, to a translator rather than to the source text's author, their mere presence in the target text constitutes evidence of foreignization. Despite slight differences in location, footnotes and endnotes serve nearly identical functions, rendering them categorically similar according to Batchelor's (2018, 142) framing. This study therefore applies the term *translator footnotes* (TFNs) to refer to both footnotes and endnotes attributable to translators; the process of identifying TFNs in the corpus is described in the next section.

7.3.2. Identifying TFNs

It was first necessary to identify all TFN candidates – that is, all footnotes and endnotes in each target text. The corpus' pre-digitized translations, primarily drawn from Project Gutenberg, contain a variety of typographical features to mark footnotes and endnotes, including asterisks or, more frequently, numbers enclosed in brackets. As such, these most common identifying typographical features were searched for to determine the paratextual consistency of each text. Once identified, this pattern was used as a search query to extract all footnotes from the text in question. This process was repeated for all translations in the corpus. All footnotes were compiled into a single table providing each paratext's contents as well as its text's SL and TL (see Worksheet 6).

A further step was required in order to distinguish between footnotes and endnotes attributable to the source text's author (or other senders) vs. the translator and thus confirm their status as TFNs. This step involved reviewing the contents of the paratext and, if necessary, its textual referent in the body of the translation. In a number of cases, footnotes were explicitly attributed to the translator or author by way of their initials or a plain statement of attribution. For instance, many footnotes in the Croatian translations were signed with the tag *prev.* to indicate *prevoditelj* ("translator"). In many other cases, however, the paratext's attribution was not indicated and therefore necessarily inferred. If, for example, the paratext constituted an explanation of the cultural context behind a specific SL term (which may or may not have been left untranslated in the target text), it was assumed to have been generated in the course of translation. Other cases were more ambiguous. In translations of historical novels such as *Fredrick the Great and His Family* and *The Merchant of Berlin* (both German>English), many footnotes commented on the historical accuracy of certain events of dialogue portrayed in the text, with some even going so far as to cite historical scholarship. These paratexts were attributed to authors, as it is feasible that they originated in the source texts themselves.

Over the course of reviewing footnotes' contents, it was observed that many footnotes were linked to what the previous chapter on lexical interference referred to as code switches – extended passages (typically songs or poems) reproduced in a language other than the TL. However, other related yet notably distinct phenomena were also observed. In the translation corpus, footnotes and their in-text referents embodied a variety of language combinations. For instance, consider the following three possible scenarios in a hypothetical French>English translation:

1) a footnote contains an English rendering of a French passage that is reproduced in the body of the target text exactly as found in the (French) source text

2) a footnote contains an exact reproduction of the corresponding French passage in the source text that is rendered into English in the body of the target text

3) a footnote contains an English rendering of a Latin passage that is reproduced in the body of the target text exactly as found in the (French) source text, which may or may not contain a footnote translating the passage into French

The first of the scenarios above perhaps reflects a straightforward instantiation of paratextual foreignization: the lexical interference reflected by the target text's SL code switch operates in tandem with the corresponding TL rendering in the associated footnote to draw direct attention (i.e., visibility) to the text's status as an SL>TL translation. In the second scenario, the placements of the SL and TL contents are reversed, yet the paratext and its referent similarly reveal a translational relationship between the target and source texts situated in their respective languages, thus raising the translator's visibility via the deliberate placement of the corresponding passage from the source text.

The third scenario, however, stands in contrast to the other two. Although the footnote does contain an English translation of the associated passage, the associated

Latin passage in the main body of the target text does not match the language of the source text (French). It may be reasoned that footnotes of this nature are not unambiguously attributable to the translator, as they fall outside the strict dichotomy of SL- and TL-oriented influence concerned in the project. Footnotes containing TL translations of passages in languages other than target texts' SLs are therefore judged not to be TFNs. This decision is vital to one text in particular. In the present corpus, the Croatian>Swedish translation *Återkomsten* is distinctly multilingual: dialogue in the source text is frequently written in languages other than the SL (e.g., French or German). The Swedish translation reproduces these code switches in these same languages, providing TL renderings via footnotes.

A number of TFN candidates pointed toward other possible senders beyond merely the author and translator. As part of the Croatian Portala e-lektire initiative, the digitized texts (available in .pdf and/or .epub file formats) were enhanced with additional educational paratexts, including classroom discussion questions, glossaries, and footnotes providing additional historical context (e-Lektire 2024). Therefore, footnotes in these texts are possibly attributable to the author of the source text, the original translator, or the annotator(s) repurposing the translation for educational purposes. Though the paratexts were not uniformly labeled across the e-lektire texts, some were explicitly attributed to the original translator, and others were explicitly attributed to the author of the source text. The originator of the rest of the footnotes was then inferred based on the paratext's content as well as the labeling pattern(s) exhibited throughout the text in question. In many cases, it was difficult to infer whether a footnote may be attributed to the original translator or the e-lektire annotator – a limitation which will be revisited later on.

Once the absolute frequency of TFNs was calculated for each translated text, the study calculated and adjusted TFN RFs to each relevant subcorpus. For instance, when examining levels of paratextual foreignization for all translations into Swedish, the TFN RF for each text was calculated relative to the total number of tokens in the subcorpus comprising all translations into Swedish (TFN $\text{RF}_{>\text{TL}}$, where TL = Swedish).

7.3.1. Hypothesis testing

The two complementary subhypotheses constituting the project's primary data analysis (see Section 4.6.2.) are adapted to this study as follows:

Subhypothesis I:

> As SL status increases relative to the TL status, translations are expected to exhibit an *increasing* degree of paratextual foreignization. Therefore, as the TL remains constant and the SL increases, it is expected that there is a *positive association* between SL status and TFN RF.

Subhypothesis II:

> As TL status increases relative to the SL status, translations are expected to exhibit a *decreasing* degree of paratextual foreignization. Therefore, while the SL remains constant and the TL increases, it is expected that there is a *negative association* between TL status and TFN RF.

No statistical test is performed for Irish source texts into the various TLs, as there are only two texts within this category.

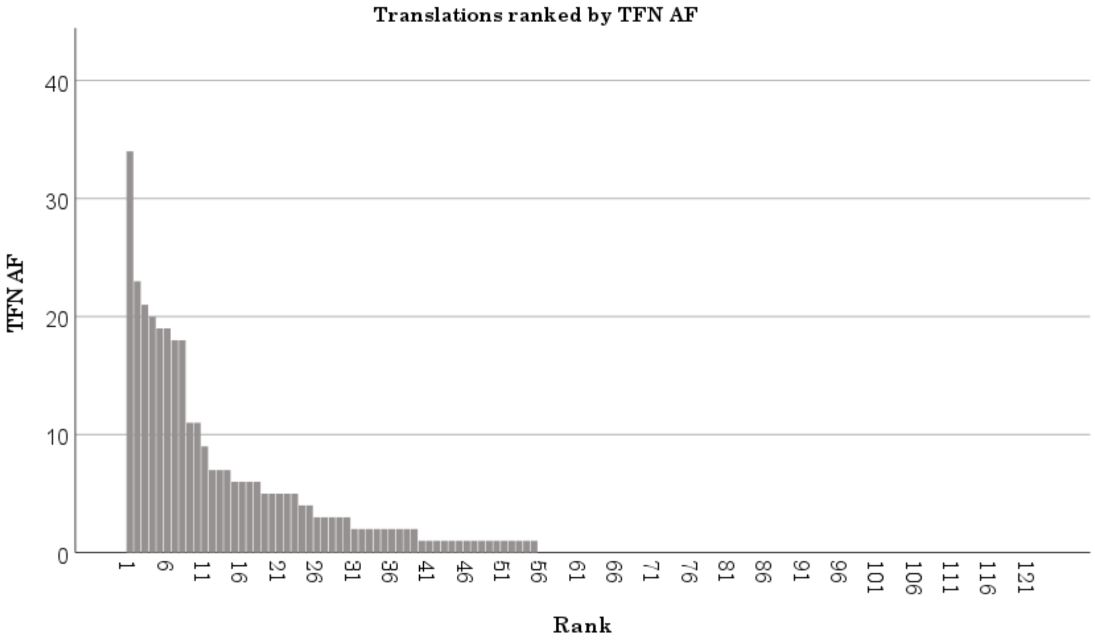
7.4. Results

Firstly, a general overview of the data is provided by way of basic summary statistics. The study then conducts the fixed TL and fixed SL analyses in order to test the two subhypotheses. Lastly, TFN RFs are presented as they correspond to each status pair (SP), language pair, and individual text, relativized to each subcorpus in question. All RFs presented are normalized per 100,000 tokens.

7.4.1. Summary statistics

In total, 330 TFNs are identified in the multilingual translation corpus. The list of all TFNs – along with their corresponding texts, SLs, and TLs – is provided separately (see Worksheet 6.1.). Figure 35 below presents a histogram displaying each translated text ranked according to their TFN AF.

Figure 35: Texts ranked by TFN AF



Exactly ten translations (8.20%) contain ten or more TFNs (AF ≥ 10), and 16 (13.11%) contain just a single TFN (AF = 1). The TFN AF of the highest-ranking text (34) is significantly higher than that of the second-highest text (23).

Notably, more than half of all translations (67/122; 54.92%) contain zero TFNs, meaning that appropriate measures of central tendency for the data set (i.e., median and mode) are equal to zero.

7.4.2. Fixed TL and fixed SL analyses

Tables 22 and 23 present the results of the fixed TL and fixed SL analyses, displaying Kendall's tau value, p-value, and population size for each fixed TL or fixed SL analysis as the variable language increases in status. If detected, statistically significant findings (p ≤ .05) confirming the expected association between the study's two variables are highlighted in gray, and statistically significant findings contradicting the expected association are italicized.

Table 22: Fixed TL analysis for paratextual foreignization

TL (fixed)	n (texts)	Kendall's tau* (increasing SL status)	p value
English	38	.071	.297
French	20	.163	.199
German	23	– .124	.248
Italian	10	.143	.323
Swedish	10	.181	.268
Croatian	15	.026	.454
Irish	6	–	–

*Hypothesized positive association.

Table 23: Fixed SL analysis for paratextual foreignization

SL (fixed)	n (texts)	Kendall's tau** (increasing TL status)	p value
English	24	.259	.070
French	30	.222	.074
German	24	.144	.209
Italian	12	.235	.181
Swedish	23	.152	.203
Croatian	7	– .250	.245
Irish	2	–	–

**Hypothesized negative association.

7.4.3. Scatter plots for fixed target languages

In the scatter plots in Figures 36-42, each point represents a text in a subcorpus composed of all translations into the specified TL. For each text, TFN RF is calculated as the frequency of TLWs relative to the total number of tokens in the subcorpus consisting of all translations into the fixed TL. SLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish TL subcorpus, its scatter plot is included here.

Figure 36: TFN RFs for all translations into English

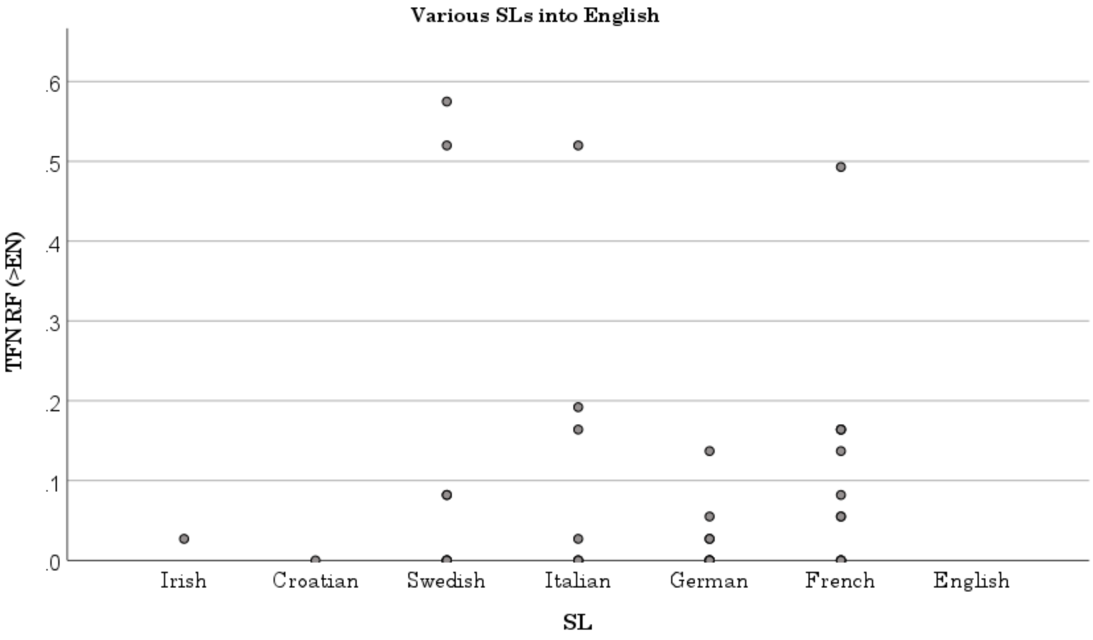


Figure 37: TFN RFs for all translations into French

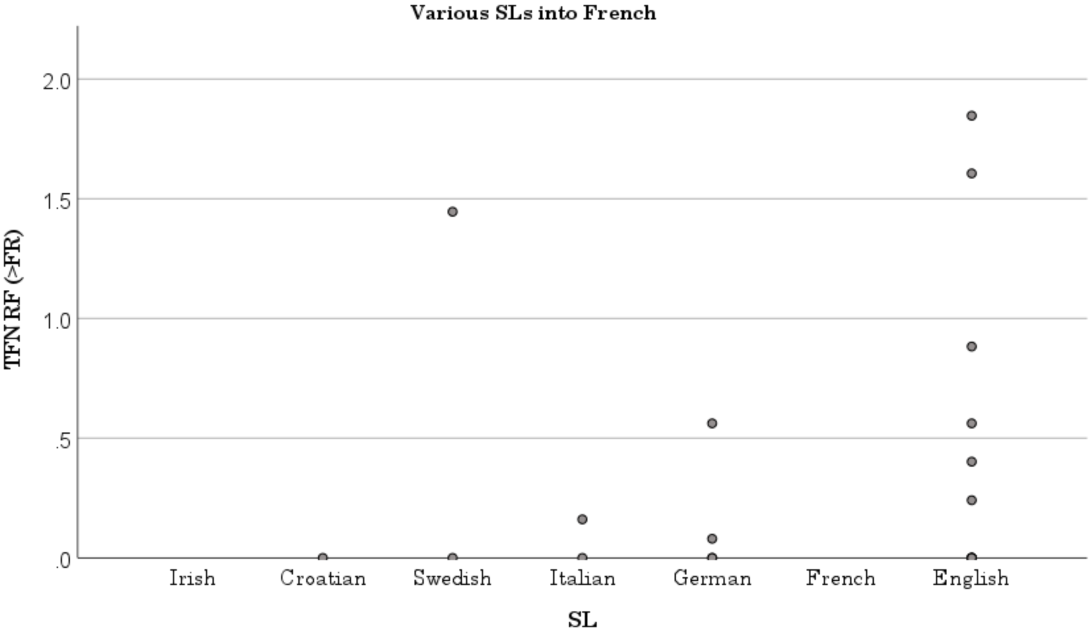


Figure 38: TFN RFs for all translations into German

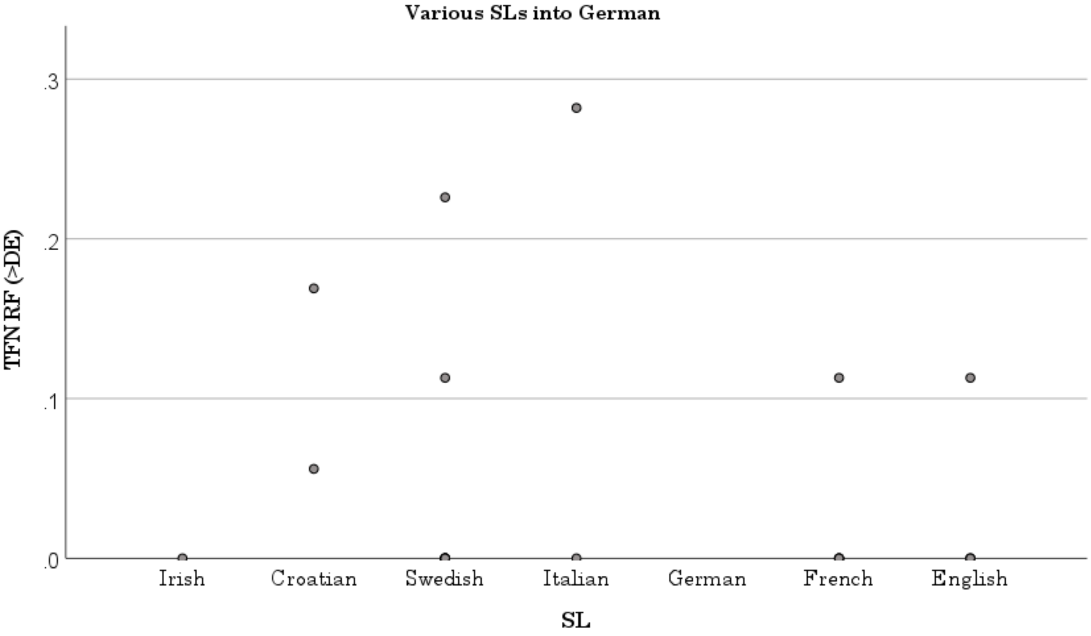


Figure 39: TFN RFs for all translations into Italian

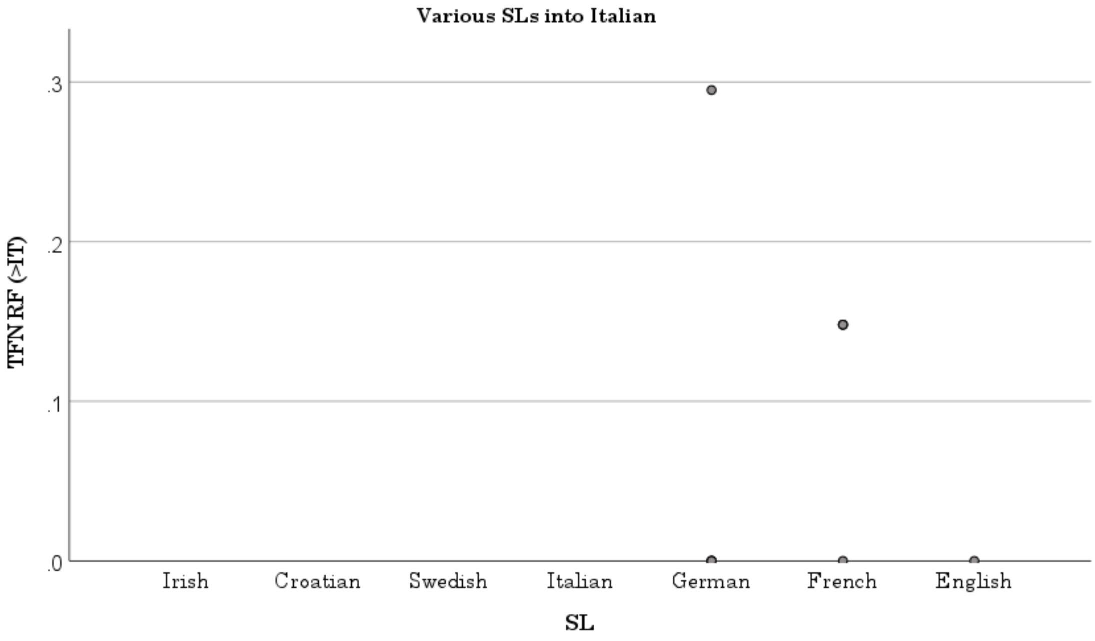


Figure 40: TFN RFs for all translations into Swedish

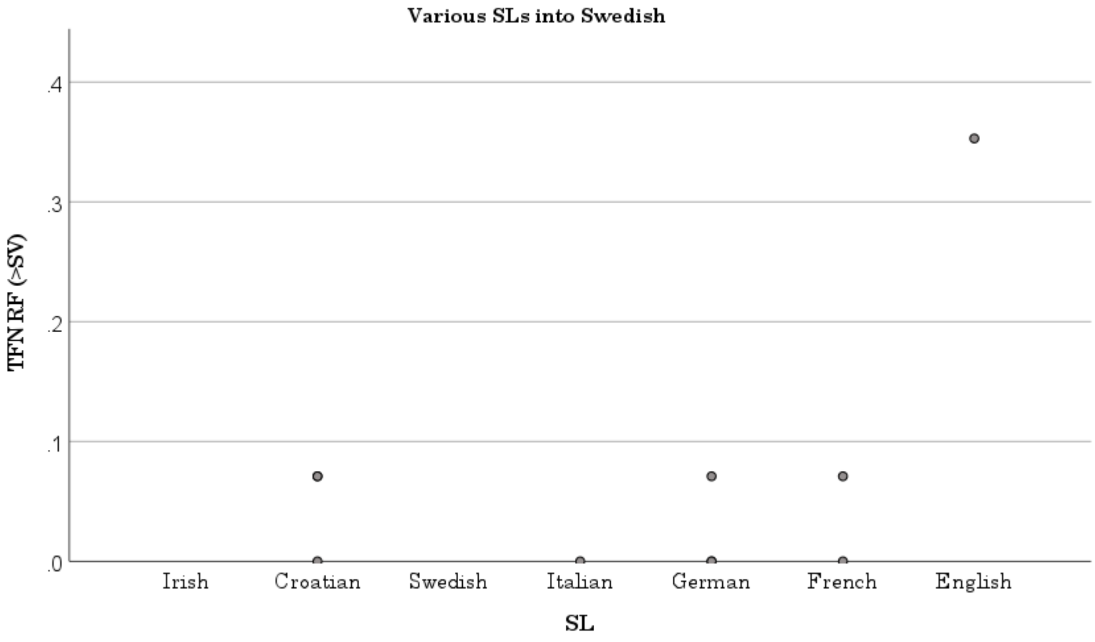


Figure 41: TFN RFs for all translations into Croatian

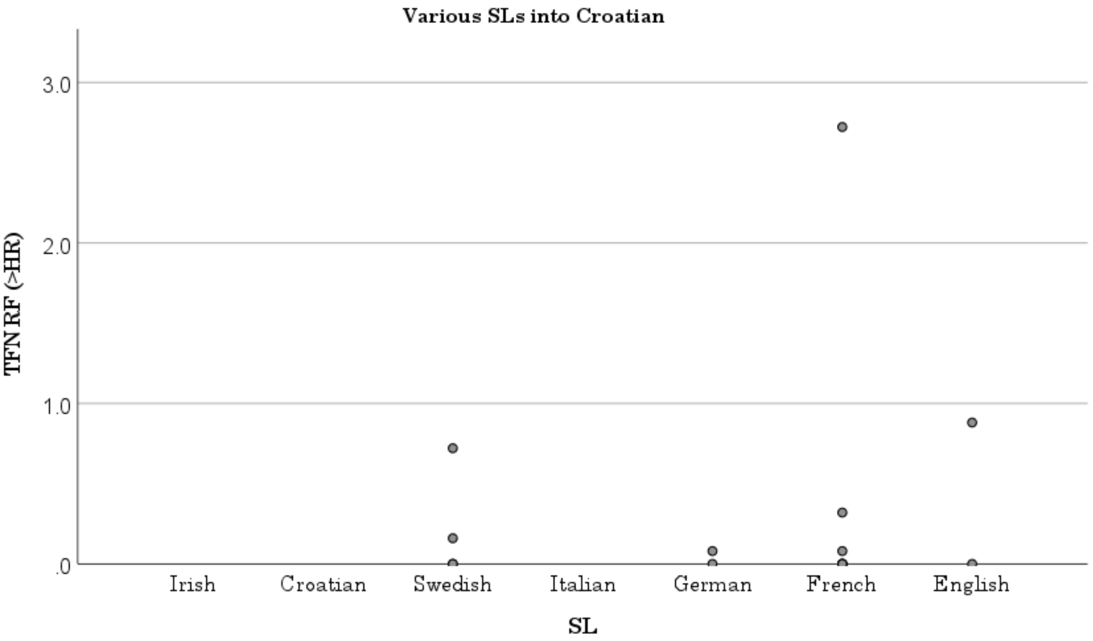
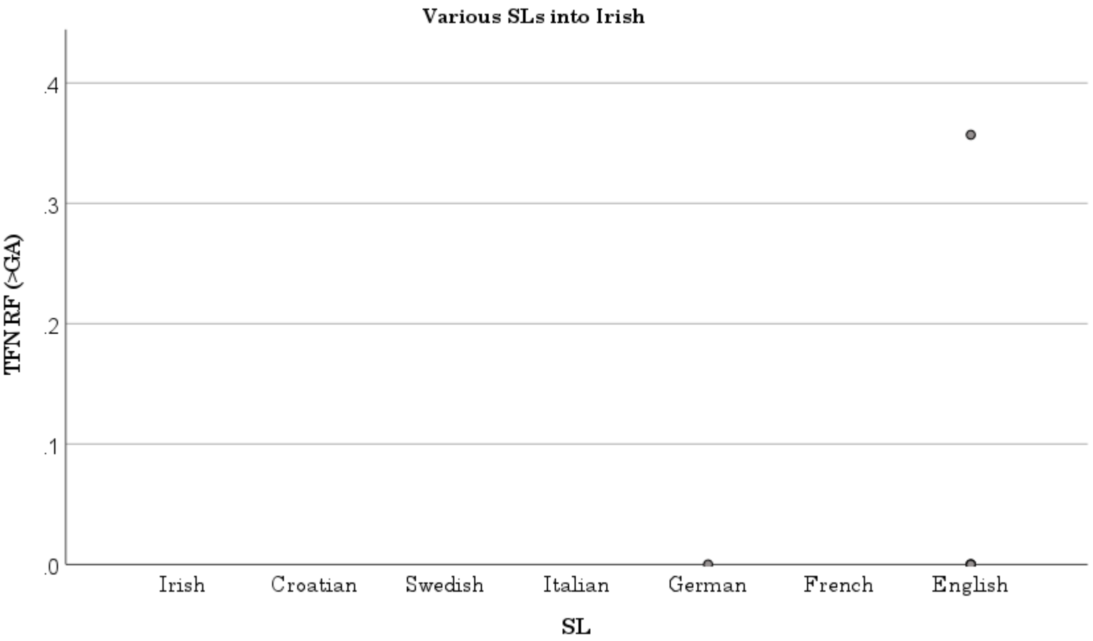


Figure 42: TFN RFs for all translations into Irish



7.4.4. Scatter plots for fixed source languages

In the scatter plots in Figures 43-49, each point represents a text in a subcorpus composed of all translations from the given SL. For each text, TLW RF is calculated as the frequency of TLWs relative to the total number of tokens in the subcorpus consisting of all translations from the fixed SL. TLs are presented from lowest status to highest status (left to right) along the x-axis. Although Kendall's tau is not calculated for the fixed Irish SL subcorpus, its scatter plot is included here.

Figure 43: TFN RFs for all translations from English

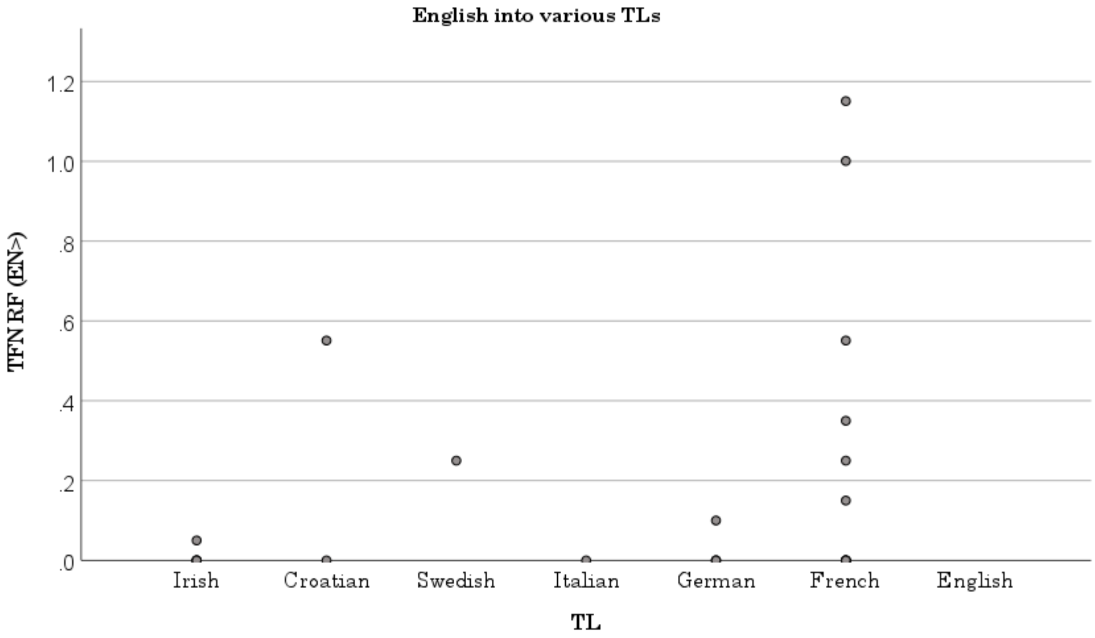


Figure 44: TFN RFs for all translations from French

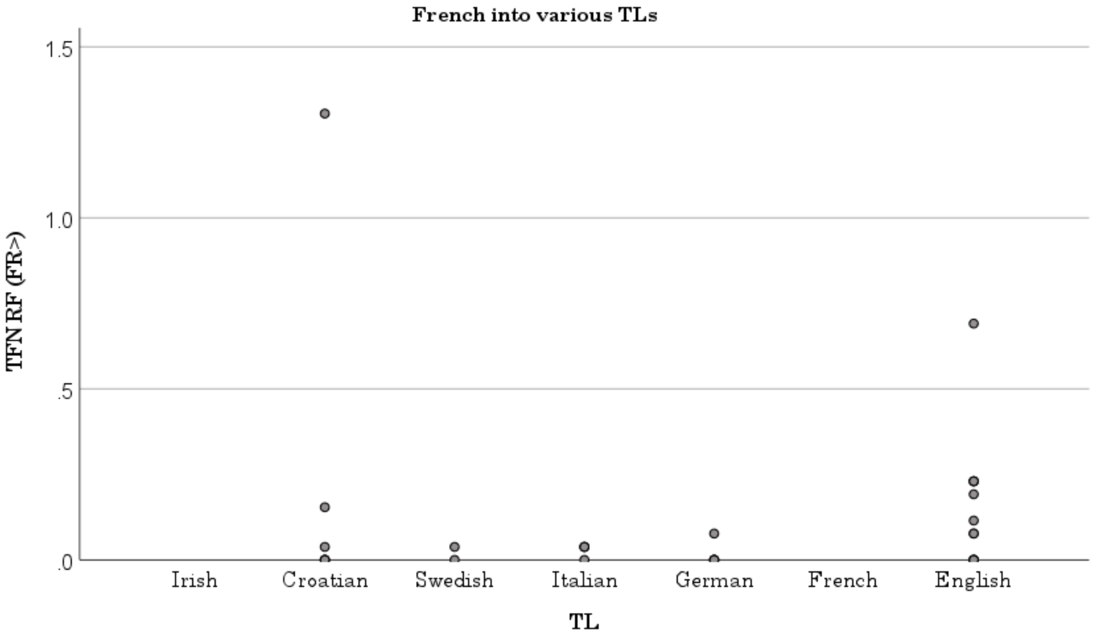


Figure 45: TFN RFs for all translations from German

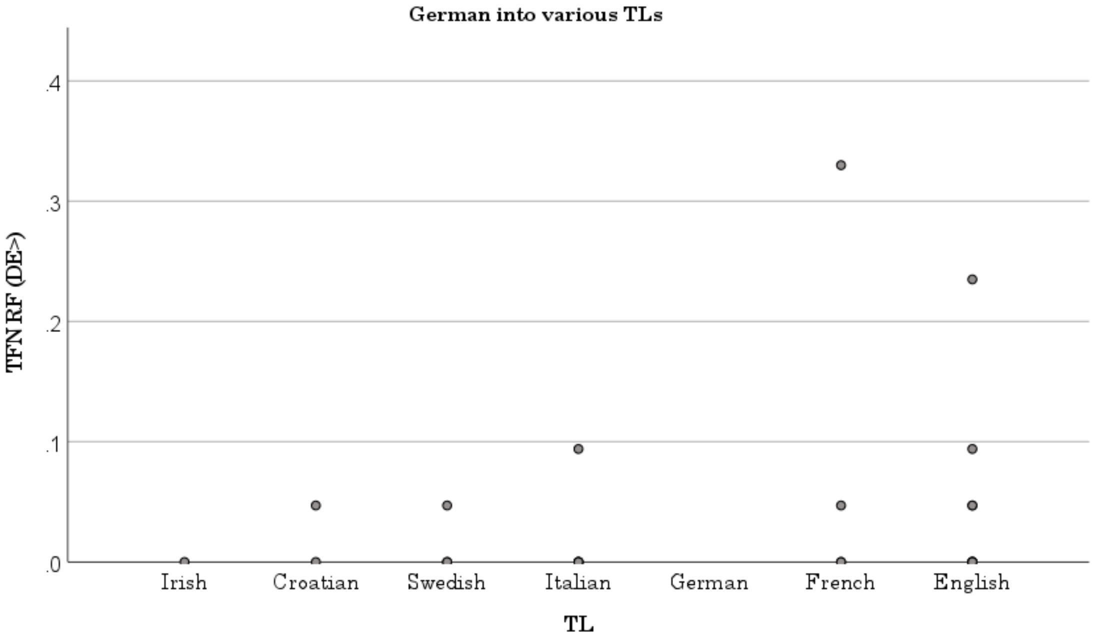


Figure 46: TFN RFs for all translations from Italian

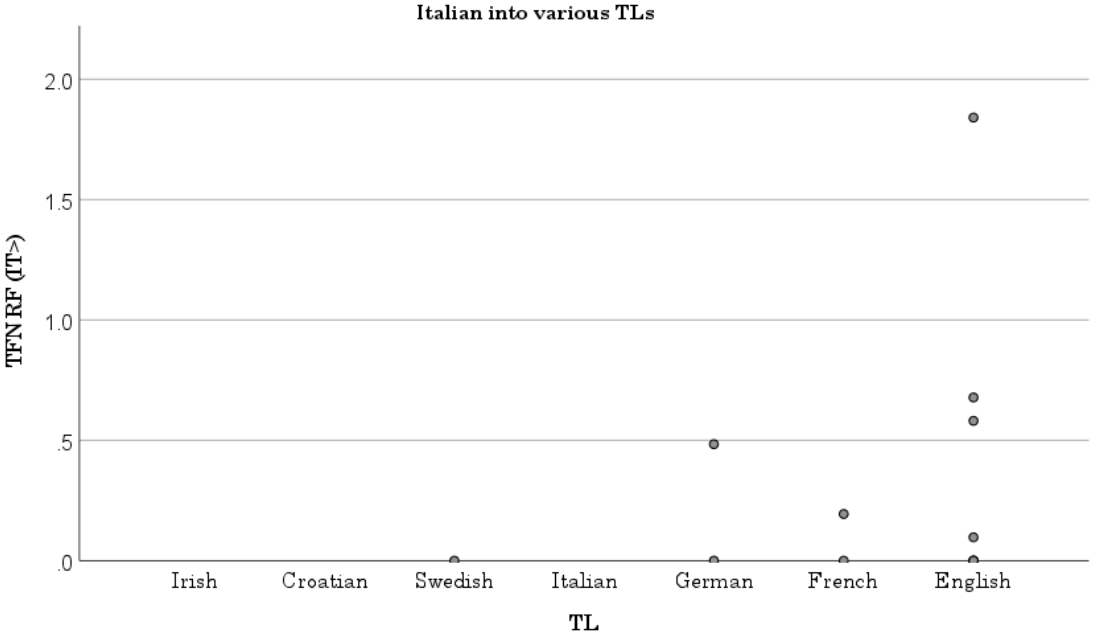


Figure 47: TFN RFs for all translations from Swedish

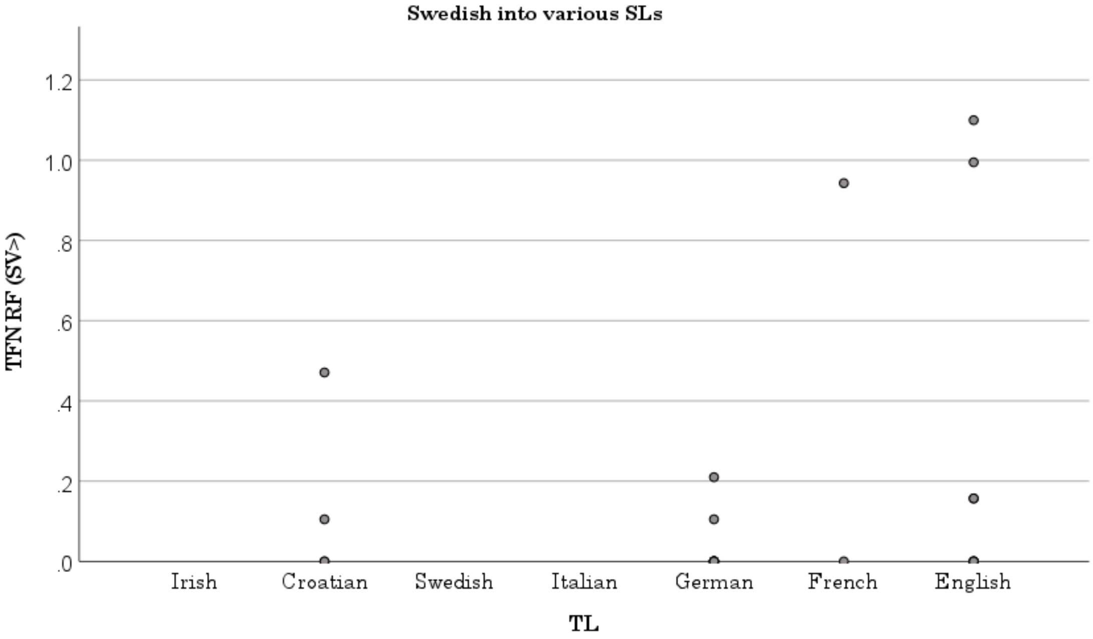


Figure 48: TFN RFs for all translations from Croatian

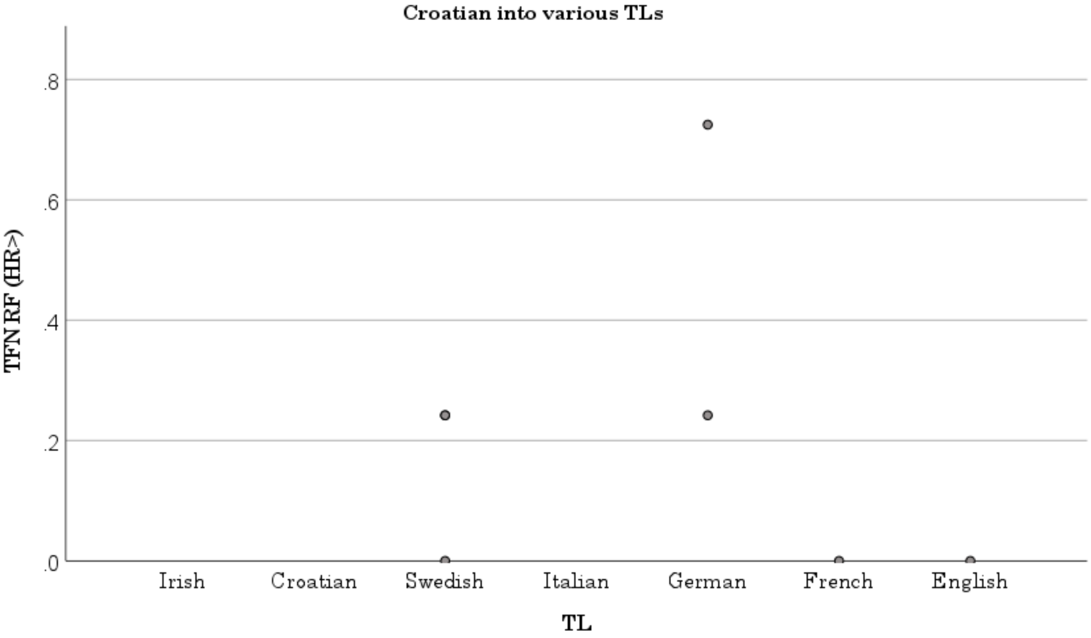
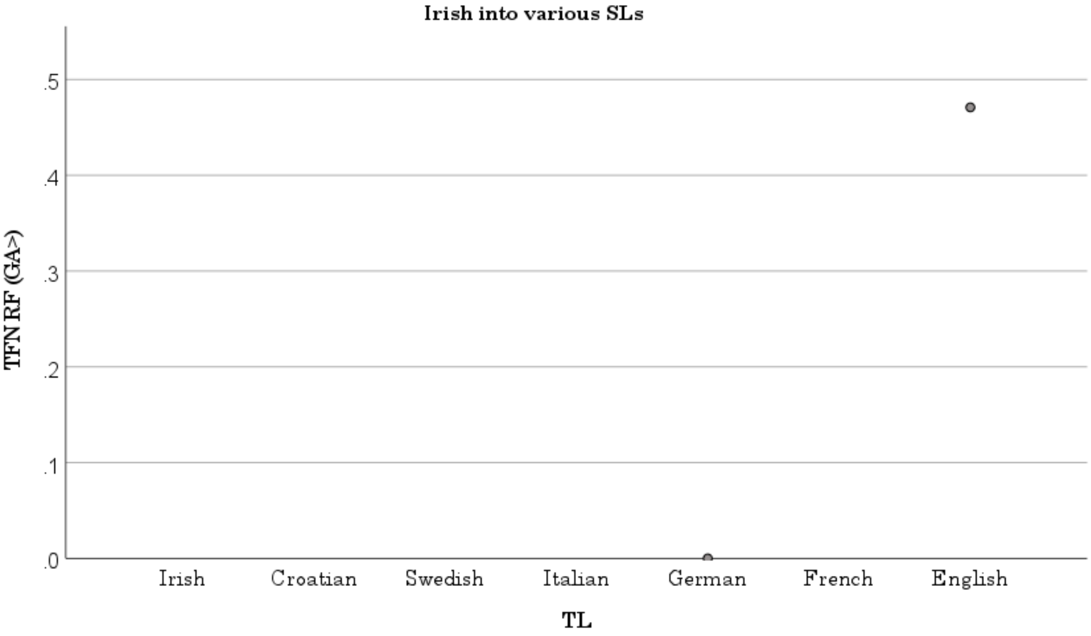


Figure 49: TFN RFs for all translations from Irish



7.4.5. Status pairs (SPs) ranked by TFN RF

In Table 24 below, English and French are categorized as high-status languages, German and Italian are categorized as medium-status languages, and Swedish and Croatian are categorized as low-status languages. Irish is categorized as an outlying very low-status language; given the small population of texts translated into and from Irish, it is excluded from the SPs. TFN RF is calculated relative to the subcorpus formed around each respective SP.

Table 24: Status pairs ranked by TFN RF

Rank	Status pair (SP)	TFN AF	TFN RF (/SP subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
1	low>high	64	6.558	975,890	12
2	high>high	117	6.353	1,841,750	23
3	high>low	56	3.476	1,611,201	12
4	medium>high	52	2.635	1,973,481	22
5	low>low	13	2.534	512,991	7
6	low>medium	10	1.199	834,307	11
7	medium>medium	7	1.131	619,171	7
8	high>medium	7	0.785	891,875	14
9	medium>low	2	0.370	540,658	6

7.4.6. Language pairs ranked by TFN RF

Table 25: Language pairs ranked by TFN RF

Rank	SL	TL	TFN AF	TFN RF (/lang pair subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
1	Swedish	French	18	13.049	137,937	2
2	English	French	69	8.407	820,785	11
3	German	French	8	6.490	123,276	4
4	Swedish	English	46	6.018	764,406	8
5	French	Croatian	39	6.345	614,696	7
6	Italian	English	33	5.175	637,707	7
7	English	Croatian	11	5.064	217,207	2
8	French	English	48	4.701	1,020,965	12
9	Swedish	Croatian	11	3.375	325,891	4
10	Italian	German	5	2.843	175,872	2
11	Croatian	German	4	2.608	153,361	2
12	English	Swedish	5	1.421	351,759	1
13	French	Italian	3	1.406	213,402	4
14	Italian	French	2	1.392	143,639	2
15	German	Croatian	1	1.103	90,671	2
16	Croatian	Swedish	2	1.069	187,100	3
17	Irish	English	1	0.923	108,293	1
18	Swedish	German	6	0.881	680,946	9
19	German	English	9	0.842	1,068,859	9
20	French	German	2	0.608	328,726	5
21	English	German	2	0.608	328,784	4
22	German	Swedish	1	0.267	375,134	3
23	French	Swedish	1	0.234	427,539	2
24	German	Italian	2	0.451	443,299	5
25	English	Irish	1	0.387	258,499	5
26	Irish	German	0	0.000	104,145	1

Rank	SL	TL	TFN AF	TFN RF (/lang pair subcorpus)	Subcorpus size (tokens)	Subcorpus size (texts)
27	English	Italian	0	0.000	20,963	1
28	Italian	Swedish	0	0.000	74,853	1
29	Croatian	English	0	0.000	54,070	1
30	Croatian	French	0	0.000	19,477	1
31	German	Irish	0	0.000	21,943	1

7.4.7. Translated texts ranked by TFN RF

Table 26: All translated texts ranked by TFN RF

Rank	Translation	SL	TL	TFN AF	TFN RF (/text)	Text size (tokens)
1	Le crime de Lord Arthur Savile	English	French	23	66.307	34,687
2	Strife and Peace	Swedish	English	19	38.081	49,893
3	Mother of Pearl	French	English	18	34.964	51,481
4	Gospođa Bovary	French	Croatian	34	33.520	101,432
5	Au bord de la vaste mer	Swedish	French	18	27.100	66,420
6	Les chasseurs de chevelures	English	French	20	15.961	125,302
7	Hermann	German	English	5	15.823	31,600
8	La Mère de Dieu	German	French	7	15.608	44,849
9	The Patriot	Italian	English	19	15.523	122,396
10	The Home; Or, Life in Sweden	Swedish	English	21	13.859	151,531
11	The Devil's Pool	French	English	5	13.110	38,138
12	Farewell Love	Italian	English	7	10.859	64,463
13	Le magasin d'antiquités, Tome I	English	French	11	9.241	119,036
14	Le portrait de Dorian Gray	English	French	7	9.101	76,917
15	Gösta Berling (HR)	Swedish	Croatian	9	8.208	109,648
16	A Tale of Brittany	French	English	6	8.070	74,348
17	The House by the Medlar-Tree	Italian	English	6	7.709	77,833

Rank	Translation	SL	TL	TFN AF	TFN RF (/text)	Text size (tokens)
18	Oliver Twist	English	Croatian	11	7.681	143,217
19	Una donna - Geschichte einer Frau	Italian	German	5	7.463	67,001
20	Der Weihnachtsabend	English	German	2	7.421	26,951
21	Tristan	German	French	1	7.360	13,587
22	Izabrane novele - Guy de Maupassant	French	Croatian	4	5.496	72,774
23	Le mort vivant	English	French	3	4.674	64,180
24	Invisible Links	Swedish	English	3	4.312	69,578
25	Le magasin d'antiquités, Tome II	English	French	5	4.219	118,510
26	Der Sohn einer Magd	Swedish	German	4	3.807	105,072
27	The Romance of a Poor Young Man	French	English	2	3.782	52,887
28	Hände	Croatian	German	3	3.699	81,096
29	Under Sentence of Death; Or, a Criminal's Last Hours	French	English	3	3.692	81,246
30	Round the World in Eighty Days	French	English	2	3.497	57,188
31	The Countess of Rudolstadt	French	English	6	3.230	185,779
32	La Morte a Venezia - Tristano - Tonio Kroger	German	Italian	2	3.159	63,315
33	Die Frau von dreißig Jahren	French	German	2	2.923	68,430
34	Feu Mathias Pascal	Italian	French	2	2.586	77,341
35	Twenty Years After	French	English	6	2.478	242,102
36	Cnoc na nGabha I	English	Irish	1	2.363	42,314
37	The Story of Gösta Berling	Swedish	English	3	2.320	129,283
38	Pakleni Stroj	Swedish	Croatian	2	1.942	102,973
39	Begravning i Teresienburg och andra noveller	Croatian	Swedish	1	1.934	51,695
40	I tre moschettieri, vol. IV	French	Italian	1	1.881	53,155
41	I tre moschettieri, vol. II	French	Italian	1	1.770	56,502
42	I tre moschettieri, vol. I	French	Italian	1	1.723	58,036
43	Händer	Croatian	Swedish	1	1.508	66,329
44	David Copperfield	English	Swedish	5	1.421	351,759
45	Gösta Berling: Erzählungen aus dem alten Wermland	Swedish	German	2	1.419	140,916

Rank	Translation	SL	TL	TFN AF	TFN RF (/text)	Text size (tokens)
46	Die Rückkehr des Filip Latinovicz	Croatian	German	1	1.384	72,265
47	Proces	German	Croatian	1	1.379	72,526
48	The Dirty Dust	Irish	English	1	0.923	108,293
49	The Triumph of Death	Italian	English	1	0.886	112,818
50	Royal Highness	German	English	1	0.850	117,601
51	The Merchant of Berlin	German	English	1	0.830	120,423
52	Germinal	French	Croatian	1	0.684	146,156
53	Joseph II. and His Court	German	English	2	0.553	361,720
54	Huset Buddenbrook	German	Swedish	1	0.504	198,365
55	Vicomte de Bragelonne	French	Swedish	1	0.414	241,263
56	Sretni vladar - Slika Doriana G	English	Croatian	0	0.000	73,990
57	20.000 milja pod morem	French	Croatian	0	0.000	118,137
58	Put oko svijeta u 80 dana	French	Croatian	0	0.000	57,721
59	Thérèse Raquin	French	Croatian	0	0.000	57,641
60	Put u srediste zemlje	French	Croatian	0	0.000	60,835
61	Preobrazaj	German	Croatian	0	0.000	18,145
62	Ja i moj sin	Swedish	Croatian	0	0.000	65,633
63	Legende o Kristu	Swedish	Croatian	0	0.000	47,637
64	On the Edge of Reason	Croatian	English	0	0.000	54,070
65	A Cardinal Sin	French	English	0	0.000	43,518
66	A Virgin Heart	French	English	0	0.000	38,586
67	The Dream	French	English	0	0.000	88,841
68	Very Woman	French	English	0	0.000	66,851
69	Blanche - The Maid of Lille	German	English	0	0.000	8,085
70	Frederick the Great and His Family	German	English	0	0.000	254,743
71	Gertrude's Marriage	German	English	0	0.000	55,575
72	The Chief Justice	German	English	0	0.000	66,469
73	The Wish	German	English	0	0.000	52,643
74	A woman at bay	Italian	English	0	0.000	77,499

Rank	Translation	SL	TL	TFN AF	TFN RF (/text)	Text size (tokens)
75	The Desire of Life	Italian	English	0	0.000	95,242
76	The Intruder	Italian	English	0	0.000	87,456
77	Christ Legends	Swedish	English	0	0.000	52,893
78	Downstream	Swedish	English	0	0.000	126,573
79	Married	Swedish	English	0	0.000	82,956
80	The Miracles of Antichrist	Swedish	English	0	0.000	101,699
81	Enterrement à Thérésienbourg	Croatian	French	0	0.000	19,477
82	Dans l'abîme	English	French	0	0.000	13,762
83	La guerre des mondes	English	French	0	0.000	63,893
84	Le grillon du foyer	English	French	0	0.000	30,907
85	Les trois hommes en Allemagne	English	French	0	0.000	67,320
86	Un amant	English	French	0	0.000	106,271
87	La débâcle impériale - Juan Fernandez	German	French	0	0.000	57,572
88	La Pantoufle de Sapho	German	French	0	0.000	7,268
89	Une femme	Italian	French	0	0.000	66,298
90	La légende de Gösta Berling	Swedish	French	0	0.000	71,517
91	Das Bildnis des Dorian Gray	English	German	0	0.000	78,142
92	Der Amateursozialist	English	German	0	0.000	89,649
93	Zwei Städte	English	German	0	0.000	134,042
94	Bouvard und Pécuchet	French	German	0	0.000	91,599
95	Bübü vom Montparnasse	French	German	0	0.000	28,200
96	Gegen den Strich	French	German	0	0.000	42,481
97	Salambo	French	German	0	0.000	98,016
98	Cré na Cille	Irish	German	0	0.000	104,145
99	Ich und Er	Italian	German	0	0.000	108,871
100	Christuslegenden	Swedish	German	0	0.000	51,513
101	Das Buch vom Brüderchen	Swedish	German	0	0.000	49,716
102	Die Gotischen Zimmer	Swedish	German	0	0.000	77,690
103	Die Inselbauern; oder, Die Leute auf Hemsö	Swedish	German	0	0.000	46,220

Rank	Translation	SL	TL	TFN AF	TFN RF (/text)	Text size (tokens)
104	Ein Stück Lebensgeschichte und andere Erzählungen	Swedish	German	0	0.000	63,132
105	Pastor Hallin	Swedish	German	0	0.000	65,595
106	Unsichtbare Bande	Swedish	German	0	0.000	81,092
107	Blátha Bealtaine	English	Irish	0	0.000	8,154
108	Cnoc na nGabha II	English	Irish	0	0.000	61,401
109	Cnoc na nGabha III	English	Irish	0	0.000	34,586
110	Dracula	English	Irish	0	0.000	112,044
111	Eachtra Pheadair Schlemihl	German	Irish	0	0.000	21,943
112	Il fantasma di Canterville e il delitto di Lord Savile	English	Italian	0	0.000	20,963
113	I tre moschettieri, vol. III	French	Italian	0	0.000	45,709
114	I Buddenbrook	German	Italian	0	0.000	220,210
115	Il Messaggio dell'Imperatore	German	Italian	0	0.000	102,523
116	Siddharta	German	Italian	0	0.000	34,832
117	Silvia ossia - La povera signorina - La gioventù provetta	German	Italian	0	0.000	22,419
118	Återkomsten	Croatian	Swedish	0	0.000	69,076
119	Den Hemlighetsfulla ön	French	Swedish	0	0.000	186,276
120	Amerika	German	Swedish	0	0.000	86,239
121	Slottet	German	Swedish	0	0.000	90,530
122	En kvinnas liv	Italian	Swedish	0	0.000	74,853

7.5. Discussion

7.5.1. Fixed TL and fixed SL analyses

Examining all translations into each TL and comparing the TFN RF$_{S>TL}$ as the status of the SL increases, there are no statistically significant associations between comparative SL status and TFN RF. Likewise, there are no statistically significant associations between the variables when examining all translations from each SL and varying the TL. The scatter plots in Figures 36-49 do not reveal any particularly striking patterns, except for perhaps those of the fixed French TL subcorpus (Figure 37) and the fixed English SL subcorpus (Figure 43), in which translations between the two high-status languages reflect some of the highest degrees of paratextual foreignization in each respective subcorpus. The study's secondary analyses are necessary to scrutinize this potential trend.

7.5.2. Status pair (SP) analysis

Organizing the data according to SP, as shown in Table 24, reveals some potentially noteworthy results. That the highest-ranking SP is low>high (TFN RF$_{SP}$ = 6.558) resolutely contradicts the study's hypothesis, as translations from comparatively lower-status SLs are expected to exhibit the *lowest* levels of paratextual foreignization. Slightly lower is the high>high SP (TFN RF$_{SP}$ = 6.353), followed by the third-ranking high>low SP (TFN RF$_{SP}$ = 3.476). The presence of high-status languages among the top three SPs seemingly corroborates the trend uncovered in Chapter 5's study on lexical interference, where language pairs involving high-status languages (whether SL or TL) exhibit the highest degrees of SL influence.

7.5.3. Language pair analysis

Adjusting TFN RFs according to the specific language pair subcorpora and ranking all 31 language pairs accordingly also fails to produce evidence supporting the study's hypothesis. Among the top ten language pairs, two of the four are combinations of the high-status languages themselves: English>French (ranked second; TFN RF$_{\text{LP}}$ = 8.407) and French>English (ranked eighth; TFN RF$_{\text{LP}}$ = 4.701). Though Swedish>French translations contain the highest traces of paratextual foreignization (TFN RF$_{\text{LP}}$ = 13.049), the small population size (n = 2) dilutes the significance of this finding. The high rankings of French>Croatian (ranked fifth; TFN RF$_{\text{LP}}$ = 6.345; n = 7) and English>Croatian (ranked seventh; TFN RF$_{\text{LP}}$ = 5.064; n = 2) seemingly confirm that the translations from these high-status languages into the lower-status Croatian exhibit greater levels of paratextual foreignization. However, revisiting the scatterplot of translations of various SLs into Croatian reveals that the high TFN RF$_{\text{LP}}$ for French>Croatian may be attributable to a single outlier (*Gospođa Bovary*), as more than half (4/7; 57.14%) of the seven French>Croatian translations contain no TFNs at all. The small population size of the English>Croatian subcorpus (n = 2) further precludes the emergence of any definite conclusions regarding Croatian translations of English and French source texts. The rest of the top ten language pairs are combinations in which the SL is comparatively lower in status than the TL, or in which the SL and TL are in the same SP and thus comparable in status.

Surprisingly, French is the TL in the top three language pairs, and six of the top ten language pairs have either English or French as their TL, while high-status languages serve as the SL in only four of the top ten language pairs. Based on this limited view of the data, it may appear slightly more likely that translations *into* comparatively high-status exhibit higher levels of paratextual foreignization, which directly contradicts the study's central hypothesis. Moreover, the observation that the English>French language pair captures the second ranking and the French>English language pair captures the eighth ranking once again seems to suggest that language pairs involving high-status languages as SLs, TLs, or both exhibit more SL influence, here in the form of paratextual foreignization. However, a more granular analysis of the data stymies the emergence of any consistently discernable pattern.

7.5.4. Ranked-text analysis

It is noteworthy perhaps that the top 14 texts ranked according to TFN RF$_{\text{text}}$ include at least one high-status language in their language pair, with the overwhelming majority having a high-status TL. In fact, all 14 of these texts are translations into English or French, except for the French>Croatian translation *Gospođa Bovary*. Checking these data against the SP analysis, it becomes clear that just two translations from the high-status languages into Croatian – *Oliver Twist* and *Gospođa Bovary* – are accounting for the high rankings of the French>Croatian and English>Croatian language pairs; either very few or no TFNs are found in any other translations from English or French into Croatian. Although six of the 23 (26.09%) total high>high translations rank among the top 14 texts sorted according to TFN RF$_{\text{text}}$, a higher percentage (9/23; 31.13%) contain no TFNs at all.

There is a sizeable difference in TFN RF$_{\text{text}}$ between the highest-ranking text (English>French; *Le crime de Lord Arthur Savile*; TFN RF$_{\text{text}}$ = 66.307) and the second-highest text (Swedish>English; *Strife and Peace*; TFN RF$_{\text{text}}$ = 38.081). *Le crime de Lord Arthur Savile* (1893) – Albert Savine's French translation of a short story collection by Oscar Wilde – presents a noteworthy outlier. Though the text does not contain the highest AF of TFNs (23) – ranking only behind *Gospođa Bovary*'s 34 – its short length of just under 35,000 tokens renders the RF substantially higher than the other texts. The vast majority of the text's TFNs (19/23; 82.60%) indicate where a word or phrase was already presented in French in the source text – "*En français dans le texte*". This phenomenon leads to the exceptionally high frequency of TFNs, given Savine's decision to flag in the target text all instances of French in the original text in this manner. Two other French translations in the corpus employ single instances of this same strategy for marking TL language usage occurring in source texts: the English>French *Le Portrait de Dorian Gray* (also authored by Wilde) and the Italian>French *Feu Mathias Pascal*. Wilde's usage of French in his works is reflective of his personal background and perhaps the era's literary zeitgeist. As with many of his European contemporaries, the Irish author learned French as a child and even wrote in this acquired language occasionally. In this regard, the paratextual foreignization in *Le crime de Lord Arthur Savile* may owe more to the coincidental alignment between the TL and the source text's multilingual aesthetics than to the translator's reactivity to the language status differential between the SL and TL. This case illustrates that paratextual elements of translated text may be primarily contingent on other factors or circumstances besides language power dynamics.

In fact, the results indicate that the strategy of using TFNs not only seems to be independent from language status, but also proves equally embraced and rejected as a viable translation strategy. As mentioned previously, just over half of all translations (66/122; 54.10%) contain no TFNs whatsoever, and there appears to be no clear pattern regarding which language pairs or status relationships facilitate the mere presence of TFNs. The nearly perfectly even split between translations containing TFNs and those lacking them is a strong reminder that this paratextual phenomenon is contingent on the same set of basic productive constraints that precludes the (non-)appearance of translational loanwords (TLWs); translators' use of paratexts such as TFNs preemptively hinges on a positive alignment between their combined inclination and editorial permissions to employ such translation strategies in the first place. As the wealth of literature on paratexts in translation makes clear, there is a wide range of factors and agents that influence whether and/or which paratextual elements appear in translated and non-translated literary texts. The highly balanced distribution between translations that employ TFNs and those that do not thus points toward the same confounding variables which undoubtedly influence the presence of translational loanwords (TLWs) in translated texts. It is likelier that paratextual foreignization in translation is first and foremost influenced by some combination of translator choices, editorial preferences, and/or publisher constraints (see Nergaard 2013) than by any overarching SL/TL power differentials that are disconnected from translated texts' local production processes.

7.6. Limitations

The present study's operationalization of paratextual foreignization reflects the same limitation as that of lexical interference: if TFNs reflect the degree of paratextual foreignization in translation, then it is not immediately apparent how paratexts – whether footnotes or other types – correspond to domestication. It may be argued that other paratexts constitute evidence of paratextual domestication, if their contents somehow lower the visibility of the text's status as a translation. In the present study, this possibility is precluded by the decision to focus exclusively on translator-attributed footnotes as evidence of increased translator visibility and thus paratextual foreignization, regardless of other paratexts' content or functions.

The missing context regarding the circumstances under which each translation in the corpus was commissioned may also be a limitation. For example, it is possible that some Croatian translations included in the e-lektire project were initially commissioned by the project itself, instead of being previous translations reformatted and further annotated for educational purposes. This lack of a clear distinction between the roles of the translator and the educator further annotating the text would complicate the task of distinguishing TFNs from footnotes that are arguably more likely attributable to other senders. Furthermore, the annotated Croatian translations perhaps best exemplify the present study's shortcoming in operationalizing paratextual foreignization, given the variety of other possible senders and addressees of paratexts in the translations.

In a similar vein, it must be emphasized that the present study's treatment of the relation between paratexts and their antecedents in source texts is simplistic. Regardless of whether paratexts or their referents are written in the SL or rendered into the TL, these identified TFNs are treated uniformly as equal evidence of paratextual foreignization. Relatedly, the study discounts as TFNs those footnotes which contain TL renderings of passages in other (non-SL) languages reproduced in the main body of the target text, as in the Croatian>Swedish translation *Återkomsten*. Footnotes accompanying the inclusion of other languages beyond translations' respective SLs and TLs may bear some other complex relation to paratextual foreignization. The effects of these types of footnotes may also be best determined with the aid of more comprehensive metadata or parallel corpus methodology.

Paratextual foreignization may even depend on the ways in which different paratexts (e.g., imprints, prefaces, footnotes) combine to reveal the roles of various actors involved in the texts' production. Of course, this suggestion harks back to the observation that some paratexts are naturally more visible than others. While corpus-based methodologies may be adequate for the purpose of identifying paratexts in translation and characterizing them in general terms, they are also inherently limited in their ability to determine the effects of these elements on the reception of the text. Particularly in the realm of translator visibility via paratexts, psycholinguistic research on readers' perception may prove necessary in addressing these concerns.

7.7. Chapter conclusion

This study finds neither statistically significant nor consistently positive associations between SL status and paratextual foreignization in translation. It is perhaps noteworthy that the language pairs of the top 14 texts with the highest TFN RFs_{text} include at least one high-status language. Still, given the failure of the fixed SL and TL analyses to reveal any consistent or statistically significant associations, no generalizations may be drawn from the text-by-text analysis.

The lack of any clear correlation between comparative SL status and the RF of TFNs in the corpus suggests that some other factor or combination of factors may be a better predictor of paratextual foreignization in translation. While further (multivariate) research is needed in this regard, the combination of the explanatory variable's lack of predictive power and the even balance between the number of translations with TFNs and those without may give some indication that TFNs are perceived as a polarizing translation strategy among actors involved in the production of translated texts, regardless of overarching sociolinguistic power dynamics. Future studies may place greater emphasis on the nature of the translation production process' influence on this paratextual element.

The historically controversial and fairly easily identifiable nature of TFNs makes them an ideal means of operationalizing foreignization in translation studies research. Further research should incorporate other types of paratexts as evidence of foreignization in translation. TFNs and other types of paratextual foreignization may simply arise over the course of texts' production processes, varying according to the translator, publisher, and other involved actors. It would be useful yet undoubtedly highly challenging to measure the variability of paratextual features of translations according to these various interrelated factors; this task would naturally require a sufficiently large and representative corpus designed to balance these metadata.

Future studies may also adapt Batchelor's full typology of paratexts into a gradient of paratextual foreignization *and* domestication, depending on their visibility or perceived disruption of the text's reception. It is also worth employing a methodology that takes into account the broadest range of possible senders and functions (see Paloposki 2010), as well as fully considering their implications in terms of indicating SL- or TL-oriented translation strategies embodied paratextually – i.e., foreignization or domestication. Still, as with translation studies writ large, future research on paratexts in translation may both require and inform more sophisticated and multidimensional frameworks to replace the doggedly binaristic SL-/TL-oriented characterizations that persist in the discipline. Freeth (2022) for instance offers a thorough argument for the necessity of expanding Venuti's dualistic paradigm to include other agents in the production process in order to characterize translator visibility as evidenced in the paratexts of digitized literary translations. Given the observed links between (non-SL) code switches and translations' footnotes, the multilingual nature of source texts and target texts alike merits further consideration for its implications for the paradigm of translator (in)visibility. As Batchelor (2018, 159) herself notes, the tendency for SL and TL readerships to be conceptualized as monolingual, culturally isolated entities "ignores the fact that many cultural products in circulation in the source or target culture will have readers and viewers from multiple cultures, particularly when the language used in the cultural product is a global one." The demonstrable deficiencies of presupposing the existence of monolithic languages and cultures undermine the validity of the SL-TL paradigm which serves this thesis, as will be further explored in the concluding chapter.

8. Thesis conclusion

This thesis has conducted a systematic, corpus-based investigation into the relationship between SL influence (i.e., interference and foreignization) and SL status in translation. The project's central research question was formulated as follows:

> RQ1: Is SL status positively associated with the degree of SL influence exhibited in literary translations?

This research question was accompanied by a central hypothesis predicting a positive association between SL status and SL influence on the lexical, syntactic, and paratextual features of translated texts. The collective results of the project's three constituent studies did not produce evidence in support of this hypothesis. The series of theoretical and practical steps necessary to operationalize key concepts and answer this research question empirically are summarized in the following section.

8.1. Summary of theoretical foundation and methodology

Proceeding from Bourdieu's concept of linguistic capital as the theoretical basis of language power, the thesis first examined the role of language power as a determinant of cross-linguistic influence (CLI) in language contact scenarios. It then related dual aspects of the Bourdieusian perspective of language power to sociolinguistic accounts of language status and language prestige, identifying the former as an ideal and operationalizable variable in the context of this thesis. Key factors of language status were subsequently identified in a brief review of sociolinguistic literature. Language status was conceptualized as the relative positionings of languages within competitive hierarchy, where competing languages encounter one another in language contact scenarios. Translation was highlighted as a particular form of language contact.

The thesis then highlighted the recurrent theme of language power in translation studies. A review of the literature illustrated the discipline's long tradition of conceptualizing translation strategies along a continuum spanning the polarities of

SL- and TL-oriented approaches. “SL influence” was designated as the overarching term for empirical identifiers of SL-oriented translation strategies, which were later clarified to refer to interference on the lexical and syntactic levels and foreignization on the paratextual level. Following the proliferation of descriptive translation studies and the discipline’s concurrent cultural turn, scholars began to assert the tendency for translators to devise strategies prioritizing (i.e., oriented toward) the features of the more powerful language in their respective language pairs. This foundational hypothesis was formalized in Toury’s law of interference and reformulated in Baker’s description of normalization (i.e., the exaggeration of TL linguistic conventions in translation) as a supposedly universal feature of translation. Venuti’s famous assertion of the tendency for anglophone translators to adopt domesticating strategies in light of the global dominance of English was also demonstrated to reflect the underlying logic of Toury and Baker. It was subsequently demonstrated that the introduction of corpus-based methodology to translation studies was explicitly intended to provide the empirical basis to test hypotheses such as Toury’s and Baker’s. Limited corpus-based translation research measured the effects of language power relations in translation with respect to isolated linguistic features of specific language pairs, yet the aspect of Toury’s law of interference involving language power relations had remained unexamined systematically. The thesis set out to fill this research gap.

The project then surveyed a series of key developments in translation studies and machine translation that maintained the theme of language power asymmetries while seemingly ignoring descriptive translation studies’ foundational hypothesis positing the correlation between language power relations and SL influence in translation. It demonstrated the ways in which this oversight represents a substantial opportunity to enhance researchers’ capacity to foster better understanding and explainability of the characteristics of neural machine translation (NMT) and large language model (LLM) output on the basis of empirical data – an area of increasing importance, given the recent rise of generative AI. Against this background, the thesis returned to Toury’s law of interference, framing this hypothesis as a simple bivariate correlation.

Language status – or SL status, in particular – constituted the project’s explanatory variable. A language status assessment model was developed on the basis of the EGIDS framework (Lewis and Simons 2010) and other insights from sociolinguistics gleaned from the literature review. The thesis then selected a range of languages – English, French, German, Italian, Swedish, Croatian, and Irish – and assessed their status in hierarchical and ordinal terms. Subsequently, a multilingual comparable corpus of translated and original literary prose fiction published in the mid-19th to early-20th centuries was constructed.

SL influence constituted the project's response variable, and was operationalized in various forms on the lexical, syntactic, and paratextual levels. On the lexical level, SL influence was conceptualized as a unilateral form of interference and operationalized as the relative frequency (RF) of translator-attributed loanwords, referred to here as translational loanwords (TLWs). On the syntactic level, SL influence was conceptualized as the diametric opposite of the exaggeration of TL syntactic features, positing translations' syntactic compositions as being identifiable along a continuous spectrum spanning the polarities of interference and normalization. It was operationalized using a novel metric called the syntactic interference/normalization coefficient (SINC), which entails microstructural and aggregate comparisons between the POS n-grams RF distributions of translations and comparable SL and TL texts. On the paratextual level, SL influence was conceptualized as a unilateral form of foreignization and operationalized as the RF of translator-attributed footnotes, or translational footnotes (TFNs).

With language status conceptualized as an ordinal variable, the project applied Kendall's tau to test for a hypothesized positive association between the two variables, pointing to the impossibility of testing for Pearson's moment-product correlation. Because the nature of ordinal variables entails that the distance between rankings is unknown, it was not possible to standardize status differentials between SLs and TLs and apply the statistical test to all data points simultaneously. As such, the thesis devised a series of primary and secondary analyses for assessing the validity of its hypothesis. The primary data analysis consisted of a pair of complementary tests (the fixed TL analysis and fixed SL analysis) formulated around two subhypotheses. First, subcorpora were formed for all translations into each given TL in the corpus; a positive association between SL status and SL influence was then hypothesized and tested for in each subcorpus. Second, subcorpora were formed for all translations from each given SL in the corpus; a negative association between TL status and SL influence was then hypothesized and tested for in each subcorpus. Secondary analyses grouped translations by status pair (SP), language pair, then texts individually, ranking each according to their (collective or individual) degree of SL influence. Without explicitly testing for associations between the variables, it was anticipated that the results of the secondary analyses would generally reflect that translations from comparatively higher-status SLs status tend to exhibit higher degrees of SL influence.

8.2. Summary of results and contributions

In the study presented in Chapter 5, lexical interference was measured in terms of the RF of translational loanwords (TLWs). Confirming expectations, it was determined that TL translators generally respond to SL status in the anticipated manner, as the levels of lexical interference in the translated texts tended to increase as SL status increased: statistically significant, positive associations were detected in all but one of the fixed TL subcorpora. These results lent substantial evidence in support of subhypothesis I. Conversely, in half of the fixed SL subcorpora, translations tended to exhibit higher levels of lexical interference as TL increased, directly contradicting subhypothesis II. The study's ranking of the SPs, language pairs, and texts according to their levels of lexical interference revealed that translations in language pairs involving a high-status language (i.e., English or French) as either the SL or TL tended to exhibit the highest degrees of lexical interference. In conclusion, the findings may be tentatively interpreted as providing modest evidence of the positive association between SL status and lexical interference in translation (when TLs are held constant), but the more granular analyses point to a more nuanced relationship between the variables, generally rejecting the study's central hypothesis.

In the study presented in Chapter 6, syntactic interference/normalization was measured in terms of a systematic, multiscalar comparison between POS n-gram frequency distributions of translations and those of comparable SL and TL texts, using a novel metric called the syntactic interference/normalization coefficient (SINC). The only affirmative statistically significant finding aligning with the study's hypothesis was a positive association between SL status and syntactic interference in the fixed German TL subcorpus (supporting subhypothesis I). A statistically significant negative association was detected in the fixed English TL subcorpus (contradicting subhypothesis I), and statistically significant positive associations between TL status and syntactic interference were detected in the fixed English SL subcorpus and the fixed Swedish SL subcorpus (both contradicting subhypothesis II). The study's secondary analyses did not indicate any consistent relationship between the variables. The results of this study therefore did not provide evidence in support of the hypothesis that SL status is positively associated with syntactic interference.

In the study presented in Chapter 7, paratextual foreignization was measured in terms of the RF of translator footnotes (TFNs). No statistically significant associations were detected among the fixed TL and fixed SL subcorpora. Moreover, the study's secondary analyses did not suggest any clear or consistent relationship between the variables either. The results of this study therefore did not provide any evidence in support of the hypothesis that SL status is positively associated with paratextual foreignization. The study's conclusion speculated that paratextual foreignization – namely, the inclusion and frequency of footnotes attributable to the translator – is more likely determined by particular conditions and actors intervening in the text's production process than overarching language power relations.

Beyond the results of the project's constituent studies, the research framework exemplified in this thesis offers several substantial contributions to descriptive translation studies and its offshoots. The novel language status assessment model developed in Chapter 3 may be applied to diverse multilingual geopolitical contexts, enabling the stable operationalization of language status as an explanatory variable in corpus-based or other empirically-oriented translation research. Furthermore, the project's various operationalizations of SL influence have been devised to be language-agnostic, meaning that they may also be replicated for different language pairs in future corpus-based studies of translational behavior. In addition to this summary of the thesis' results and main contributions, it is necessary to conduct a conclusive discussion synthesizing and jointly interpreting the findings of the project's individual studies.

8.3. Overarching discussion

The projects' findings indicate that translations are most sensitive to language power dynamics on the lexical level. This outcome aligns with Hoffer's (2002) observation that loanwords constitute the preeminent and most common byproduct of language contact more broadly. For the large majority of the TLs (i.e., receiving languages) investigated in the project, their translations from higher-status SLs tended to contain higher RFs of translator-attributed loanwords. However, the tendency for some SLs to induce more translator-attributed loanwords as TL status *increased* undermined the prospects of a straightforward relationship between language status and lexical interference. A more nuanced view of the relationship between language status and lexical interference emerged following more granular analyses, which revealed that language pairs involving high-status languages, whether as the SL or TL, exhibited the highest levels of lexical interference. That translations between high-status languages – as well as translations from lower-status into higher-status languages – contained high frequencies of TLWs was an unexpected outcome. In consideration of language power relations, it may be the case that the most powerful (i.e., high-status) languages are secure enough in their positions that translators do not feel pressured to protect their linguistic capital by avoiding foreign lexical items. From this perspective, target-culture norms in English and French may conceptualize loanwords in translation as desirable exoticisms instead of threats to their linguistic capital.

This tendency for translations between English and French in particular to contain frequent loanwords may be a product of the time period in question. Rollason (2005) notes that the long-established, bidirectional transmission of lexical items between Britain and France has more recently given way to a unilateral importation of English-language words into the French lexicon, due not to Anglo-French relations but rather the globalizing cultural and economic dominance of the United States beginning in the late twentieth century. These "reluctantly accepted Anglicisms" are appropriated despite the "massive [French] government-sponsored promotion for the French language" and the "puristic, normative policy of French institutions and academies" amidst the language's receding influence on the world stage (Snell-Hornby 2006, 140). At the same time, Rollason (2005, 52) projects that the same unilateral influence of the

English lexicon has been even more pronounced on other European languages like German and Italian – the project's medium-status languages. The results of this thesis thus perhaps partially corroborate the notion that translation strategies regarding lexical interference broadly align with broader trends in language change: both are driven by the comparative dominance of source/donating languages. Furthermore, these observed trends support Bourdieu's assertion that prevailing linguistic practices in language contact scenarios exhibit bias toward languages with higher linguistic capital. In line with Hoffer's (2002) overview of loanwords as the preeminent byproduct of language power asymmetries in language contact scenarios, the project's conclusion regarding the sensitivity of loanword frequencies to language status indicates that this translation phenomenon offers a promising area for future research, as examined further in Section 8.5.

Although the results of Chapter 5's study offer only ambiguous support for its central hypothesis, the contrast between the fixed TL analysis' affirmative findings and the fixed SL analysis' inconclusive results points toward a broader theoretical takeaway. This outcome lends credence to Toury's assertion that translation strategies are principally formulated according to TL norms, as translations grouped according to TL consistently matched their levels of lexical interference to SL status, while translations grouped by SL did not display any reliable patterns. That is, the project's findings support the view of translations as "facts of the culture that would host them" (Toury 2012, 18), where translators' decisions are primarily governed by target-culture norms determining which translation practices are considered appropriate or acceptable. This framing offers a highly plausible explanation for the study's results, given that the same trend was observed in nearly all of the fixed TL subcorpora. It suggests that comparisons of linguistic features between sets of translations from various SLs into the same TL may offer the most fruitful grounds for descriptive translation research – at least in terms of lexical interference.

The combination of the (partially) affirmative results of Chapter 5's study with the other studies' inconclusive results confirms Toury's (2012, 315) suspicion that levels of interference are unlikely to be uniform across different linguistic levels. Considering normalization as the inverse of interference on the syntactic level, the findings undermine Baker's (1996, 183) claim that the overuse of "typical grammatical structures" constitutes one of the most apparent forms of normalization, which is allegedly more prominent in translations into higher-status TLs. It is perhaps also necessary to examine the project's outcome in light of the contrast between operationalizations of different forms of SL influence, such as the deliberate use and perceived intrusion of loanwords in translation versus the aggregate, frequency-based comparisons of syntactic structures' distributions. The first of these phenomena undoubtedly reflects a very conscious decision by the translator, given the SL term's self-evident disruption of TL lexical conventions. In this respect, the alignment or refutation of the target system's translation norms regarding the acceptability of loanwords would be much more blatant. On the contrary, it is highly unlikely that translators would consciously adjust the entire distribution of target texts' syntactic structures in direct relation to those of comparable SL or TL texts. It may be anticipated that translators of certain target cultures are more likely to replicate certain syntactic structures that are evident in their SLs, and that language power relations would play some determinable role in this equation. Even so, the inconclusive results of the study in Chapter 6 perhaps indicate that syntactic interference/normalization in translation is more attributable to the subtler cognitive phenomenon of syntactic priming as determined by syntactic similarities and differences between SLs and TLs (see Gries and Kootstra 2017).

Although no relationship between language power relations and paratextual foreignization was detected, it is worth observing that some translator-attributed footnotes were directly linked to loanwords (i.e., lexical interference), a strategy which is discussed by Hermans (2007, 44). For example, the TLW *puderkammer* in the DE>EN translation *Joseph II. and his Court* is accompanied by a footnote providing a literal translation for the German term ("powder-room"). The IT>DE translation *Una donna* flags allegedly untranslatable wordplay in the source text by simply reproducing the term in the target text with an explanatory footnote (*Ins Deutsche nicht übertragbares Wortspiel.* [Wordplay not transferable to German.]). As such, the translator's visibility – achieved through foreignizing strategies such as the use of footnotes – sometimes coincides with interference on the linguistic levels, particularly with respect to foreign lexical items. This observation suggests that translators sometimes feel compelled to explain or even *justify* their use of loanwords via the addition of paratexts, perhaps indicating an awareness of the potentially controversial reception of their deviation from TL conventions. From this perspective, translators' paratexts may be used to infer their perceptions of target-culture translation norms.

Finally, the results shed light on the project's initial distinction between language status and language prestige. The lone Irish>English translation (*The Dirty Dust*) exhibited a high degree of lexical interference (TLW RF_{text} = 28.626), which was a highly surprising result in the context of the study's hypothesis, given that the text was translated from the lowest-status language into the highest-status language. Still, knowledge of the historical circumstances of this particular language pair may contextualize this result; the translation of Irish-language works into English has traditionally been viewed as a means of promoting Irish literature out of reverence for the language (Tymoczko 1999/2014; Fhrighil et al. 2020). Despite its low status, Irish may be thus understood as holding a high *prestige* among certain English-language translators – typically, those in Ireland seeking to boost the language's revival. The lone Irish>German translation of this same source text (*Cré na Cille*) also exhibits a moderate degree of lexical interference (TLW RF_{text} = 9.602), with the text ranked at 31 out of 122 total translations. It may be the case that SL prestige is a better predictor of lexical interference in translation than SL status in certain cases. Although this speculative pronouncement has been made on the basis of merely one translation, the low levels of lexical interference in the English>Irish and German>Irish translations lend further support. It is also possible that lexical interference exhibited in the English and German translations of the Irish novel *Cré na Cille* may be attributed to translators' reverence for the source text or author in particular, as both are highly respected in the Irish literary tradition (Byrne 2018). Regardless, the corpus' small numbers of translations into and from Irish preclude more definitive takeaways, meaning that further research is needed in this area. More generally, the project's results must be contextualized with respect to its overarching methodological limitations before any future research avenues are recommended.

8.4. Overarching limitations and critical reflections

Whereas the limitations of each individual study in the thesis are explored in their respective chapters, this section reflects on the broader limitations of the project's methodological approach in relation to the complex, interdisciplinary nature of its research question.

The operationalization of language status as a measurable explanatory variable in empirical translation research requires additional scrutiny. For instance, the language status assessment model's treatment of languages as indivisible entities, regardless of domain or dialect, is open to further reflection. The novel language status assessment model introduced in this thesis is predicated on a coarse hierarchy among geopolitical scales and social domains; however, this assumed hierarchy may be simplistic, given the complexity of relations between disparate social domains in conferring languages' their power. Returning to the example of modern Irish, which enjoys official status in both the Republic of Ireland and the European Union, it is worth reiterating that the language's hold on the various social domains in the EGIDS model is highly debatable. As emphasized in Chapter 3, each level on the (E)GIDS scale entails the capture and stability of all lower levels. There is certainly a case to be made that Irish is currently firmly implanted at the National level (Level 1), seeing as policy dictates its national use in education, work, mass media, and government. In practice, however, the use of Irish in these various settings is highly limited. A similar conundrum arises for the potential argument that Irish constitutes a Regional language (Level 2), given its wider use in the Gaeltacht. While the project ultimately designated Irish as an Educational language (Level 4), reasoning that the language was primarily promoted as a basic literacy via education during the relevant time period, the gradual implementation of its official standing in institutional policies complicates this picture. Undoubtedly, other ambiguous cases abound with respect to testing the language status assessment model's validity.

While the project justified its focus on European languages and literary translation on pragmatic grounds, this decision may nevertheless be perceived to run afoul of calls for translation studies to shed its historical inclinations toward both Eurocentrism and literature as adequately representative of the supposedly universal nature of translation features (see Zanettin 2012, 22; Van Doorslaer and Flynn 2013; Tymoczko 2018, 248). As noted by Chesterman (2004, 43), many supposedly universal features of translation have been observed in literary translations and extrapolated to other domains. It must therefore be reiterated that the findings of this thesis should not be generalized beyond its limited context. The postcolonial turn in translation studies criticized the descriptive branch's fixation on European languages and cultures as well as its unquestioning assumption of nation-states as the principal divisions between literary systems, which are arguably both reflected in this thesis (Hermans 2019, 146).

The decision to conceptualize language power as abstracted from any specific domain (e.g., literature) perhaps warrants further scrutiny. Although this theoretical maneuver was justified with reference to prominent sociolinguistic views, it is arguable that Casanova's narrower concept of literary capital better suits the project's focus on literary translation, particularly given its direct lineage to Bourdieu's work, which provides the basis for conceptualizing language power in this thesis. The precise extent of the literary systems' independence from language power or linguistic capital more broadly is not entirely clear in the work of Bourdieu and descriptive translation scholars. For Even-Zohar (1990, 66-68), "[p]olitical and/or economic power may play a role" in determining the centrality of literary systems, such as with the British and French empires imposing their literatures on colonies, but these factors are not strictly necessary. The thesis perhaps offers a relatively superficial treatment of this complex theoretical query.

It is worth further discussing the limitations of the project's corpus design as outlined in Chapter 4. The reliance on comparable corpus methodology in order to identify interference is problematic, as interference is typically determined by directly comparing translations with their source texts (Chesterman 2004, 39). There is a limited number of parallel text pairs in the project's multilingual corpus, and no direct comparisons are made between translations and their source texts. Moreover, the entire concept of comparable corpora may be problematized, as there are many criteria beyond simply genre or text type that may be used to establish comparability, and the conventions of a given genre or text type may be dissimilar across languages and cultures (see Zanettin 2012, 48; Lefer and Vogeleer 2013, 15; López-Arroyo 2020). The necessary inclusion of a diverse range of languages in this thesis arguably compounds these comparability challenges.

It is also potentially problematic that the comparable corpora representing the project's selected languages are arguably neither properly genre-controlled nor register-controlled (see Lefer and Vogeleer 2013). While the corpus constructed for this thesis included historical novels (particularly translated from German into English), Van Poucke (2011, 108) highlights the different literary subgenres typical of the selected languages during the time period in question, and decides to explicitly exclude historical novels in anticipation of their deviant linguistic features. Other texts included in the corpus (e.g., H. G. Wells' *The War of the Worlds*) are arguably more appropriately categorized in distinct subgenres such as science fiction, thus constituting another distorting variable unaccounted for in the project. Another frequently avoided misstep in corpus-based translation research is the false comparison between original and translated texts reflecting different registers (Bernardini and Ferraresi 2011, 228). Biber (1995) identifies multiple registers within fictional texts (narrative, situation-dependent, non-abstract, and edited), though others treat fiction as a single register. In her wide-reaching study on register variation, Neumann (2014, 84-85) insists on fiction being a self-contained register, while Hansen-Schirra et al. (2011, 143) likewise treat fictional texts as a single register in the CroCo Corpus. "Fiction" is likewise defined as an overarching register in the Dutch Parallel Corpus, but as Delaere and De Sutter (2017, 9) point out, this register only contains four texts translated from French, and the lack of information on the manner in which different registers were defined poses a significant methodological problem. Ideally, the multilingual corpus might have been constructed on the basis of a systematic balance of specific subgenres and observable registers. However, the text selection process for constructing the corpus was primarily based on availability, which is perhaps an inevitable consequence of a wide-reaching project of this nature; the inability to conduct the text selection process systematically potentially undermines the representativeness of the corpus.

The decision to strive for representativeness in the comparable subcorpora on the basis of SL authors' popularity in international markets may be subject to further scrutiny. This theoretical basis was adopted from Casanova (2002), who also notes that authors such as James Joyce and Franz Kafka were hardly recognizable in their domestic literary markets until foreign translations catalyzed their international acclaim – in Kafka's case, posthumously. Such cases complicate the seemingly straightforward goal of constructing a representative sample of a literary system for a specific timeframe. It is perhaps also relevant that some authors included in the comparable subcorpora were L2 speakers, as some researchers may consider this fact to muddle the prospects of achieving representativeness in the corpus. For instance, Joseph Conrad is a prominent author famous for writing in his third language, English, and his writing style is said to reflect influence from his first two languages, Polish and French (Gardner-Chloros and Weston 2015, 185). More generally, the construction of any sample intended to represent a single national literature is intrinsically problematic, as the "idea of a single national culture mystifies the fact that these cultures are internally diverse" (Baer 2023, 227). Such complications point to the necessity of confronting "what exactly is normalized and with respect to what norms" in translation research applying comparable corpus methodology (Lefer and Vogeleer 2013, 17).

Despite best efforts during the text selection process, it is possible that covertly indirect translations were included in the corpus, in which case the inclusion of a mediating or pivot language in an indirect translation chain between would be expected to dilute or distort the manifestation SL influence in the ultimate target text (see Hadley 2017). As translations between low-status languages are generally more likely to require mediation through a high-status pivot language (Whyatt and Pavlović 2021), it may be expected that these translations in the corpus present the greatest risk for covertly indirect translations. As such, conclusions drawn from multilingual corpus – which generally contains fewer translations into, from, and between its selected low-status languages – may be further cast into doubt. Moreover, the wide range of texts' publication dates as well as the temporal discrepancies between the comparable texts and translations (particularly for low-status languages) lend an inadvertently diachronic dimension to the corpus, thereby potentially distorting the studies' findings.

The separation of linguistic levels for the operationalization of interference and foreignization constitutes another potentially significant limitation of the project. It has long been observed in corpus-based translation research that lexical and grammatical features are highly interdependent, hence researchers typically investigate so-called lexico-grammatical features (see Kenny 2001). The interdependence of paratextual features with linguistic features of translation is also evident, as footnotes appear to be a common companion to the use of loanwords in translation. This relationship is likely also related to external factors, and process-oriented research once again proves imperative: Van Poucke (2011, 107) notes that Dutch publishers generally prohibit the use of footnotes to "explain foreign words" in literary translation, as this paratextual strategy is associated with scientific texts. The rigid separation of translations' lexical, syntactic, and paratextual features inhibits the analysis of the overarching relationship between language power relations and SL influence.

There are many potential confounding variables that have been excluded from the scope of this thesis, the most prominent of which are briefly discussed here. Although the SINC methodology introduced and applied in Chapter 6 attempted to neutralize the effects of SL-TL distance, this factor could play a role in facilitating SL influence in any form, and would ideally be handled more systematically. Moreover, while this thesis has addressed the social constraints acting on translators (i.e., SL-TL power relations), it has necessarily neglected the cognitive constraints, which also impact the manifestation of SL influence in translation (Kotze 2021, 115). Although the project has made reasonable attempts to isolate the effects of translation specifically from general language contact in general, as in Chapter 5's identification of translator-attributed loanwords, it is very challenging to accurately distinguish these forces, as evident in Kotze's (ibid., 122) summary of the relevant literature.

The bivariate design is another significant limitation of the project. As previously mentioned in Chapter 4, recent corpus-based translation research implements multifactorial analyses in an attempt to systematize the multitude of potential confounding factors that complicate straightforward causal relationships between two variables (De Sutter and Lefer 2020). In fact, advanced statistical approaches may overturn the findings of earlier research. Conducting a multifactorial analysis for Baker and Olohan's (2000) bivariate study on explicitation, De Sutter and Lefer (2020, 13) refute the original authors' conclusion, finding instead that the use of *that* – whether in original or translated texts – is more attributable to "the complexity of the syntactic environment and on the basis of register" than simply texts' translation status. The authors characterize the tendency for corpus-based translation research to take the form of bivariate (i.e., monofactorial) observational studies as a limitation which adjacent fields have long abandoned, and argue instead that research designs involving corpus methods should be multifactorial, interdisciplinary, and multi-methodological (De Sutter and Lefer 2020, 5-6). The bivariate design of this thesis constitutes a significant shortcoming.

Perhaps another reason why Toury's law of interference has remained untested is that its underlying dichotomous epistemology has been widely problematized, with scholars having since proposed a diverse range of alternative epistemologies (see Marín García 2023). Consequently, the core dichotomies underpinning this project – SL/TL power relations and SL-/TL-oriented translation strategies – may be viewed as obsolete by more current perspectives. Blumczynski and Hassani (2019) illustrate translation studies' traditional embrace of dualistic thinking and highlight the many ways in which this epistemological base proves insufficient or misleading in translation research. It may be argued on these grounds that the thesis' research design commits the same fundamental error as Schleiermacher, whose writer-reader binary overlooks the translator as the "hidden middle term" (Pym 1995, 5). Investigations into translators' discernable styles also from corpus methods have also been conducted using comparative frequencies (see Baker 2000), and style has even recently been examined in relation to post-editing machine-translated literature (Kenny and Winters 2020). The translator's presence was tepidly acknowledged in Chapter 7's study on paratextual foreignization, yet the possibility of distinct translator styles influencing the translations' compositions was otherwise disregarded. This exclusion owed to the adoption of Toury's theory, which hypothesizes the uniformity of translation strategies for translators in a given target culture. The composition of the corpus did not adequately reflect this presupposition, as many language pairs have very few texts with certain translators being overrepresented.

Even taking the descriptive translation studies agenda on its own terms, problems remain. The resolute empirical tradition in translation studies has continually drawn criticism of the legitimacy and feasibility of its scientific aspirations. As one of the most ardent critics in this vein, Pym (2007, 42) accuses descriptive and corpus-based translation research as embodying "a rather quaint empiricism" that "rarely transcend[s] positivist notions of science." It is possible that scholars at the heart of descriptive translation studies "have perhaps over-reacted against traditional prescriptivism in their desire to place Translation Studies on a more scientific basis" (Chesterman 2004, 36). Such concerns have remained firmly in place since the early aughts of corpus-based translation studies, which Tymoczko (1998) contextualizes in relation to the turn away from positivist scientific approaches over the second half of the 20th century, drawing particular attention to the dangers of its pursuit of "laws" of translation. It is worth reiterating Toury's (2012, 300-301) clarification of translational laws as inherently *probabilistic* in nature and tenuously supporting a "gradually unfolding theory" of translation encompassing an ever-expanding set of interacting variables. Furthermore, calls from Toury and Even-Zohar to pursue case studies reflecting the fullest variety of possible translation contexts and specific historical and cultural circumstances may be said to reflect the "postpositivist nature" at the core of descriptive translation studies (Tymoczko 2014, 41-42). The renewed embrace of Toury's (2012, 80) maximally inclusive concept of translation as being defined by target cultures may also help empirical translation researchers guard against accusations of positivism, given its adaptability to varying cultural interpretations of what translation is and should be.

The limits of the project's bivariate design have already been alluded to in Chapter 4. Regardless of the project's outcome, any presumed causal relationship between language status as an explanatory variable and SL influence as a response variable derived from corpus-based research alone should be subject to scrutiny as well. Toury (2012, 5-6) lays out the prerequisites for deriving explanations of translational behavior from empirical data, stressing the necessity of jointly investigating the "functions, processes and products" of translation, as their interdependence is necessary to disentangle in order to generate explanations. Translation norms serve as "explanatory hypotheses for actual behavior and its perceptible manifestations" in each of these dimensions (ibid., 65). In this regard, the project's focus on translation products – translated texts and their linguistic compositions, with minimal metadata provided – preemptively limits the explanatory power of its results.

The "explanatory power" toward which descriptive translation studies has always been oriented may be problematized as well. Chesterman (2008) portrays the multiplicity of perspectives on the nature of explanation, causality, and the relationship between the two. He argues that, while explanation may take different forms, explanatory power necessarily involves the "establishing of *relations* of different kinds, between the *explanandum* and various other phenomena or variables" (Chesterman 2008, 376). The widespread yet implicit belief in empirical methods' capacity to fully encapsulate these relations is perhaps why an increasingly large constituency supports the idea that "translation studies has remained firmly embedded within the reductionist model, not so much in its search for universal laws but rather in its search for decomposing systems into elementary, simple units" (Marais and Meylaerts 2018, 2). Translation studies has undoubtedly undergone a much-needed expansion since its descriptive branch was coined and codified. This expansion has included "dethroning the (literary) written text as the primary product of translation" and reconceptualizing translation as a "complex, unpredictable *process* (rather than as a product)", while also "overcoming the *binaries* (source-target, original-translation, domestication-foreignization, for example) that have traditionally delimited its field of study" (Meylaerts and Marais 2023, 1). Since the project arguably embodies these conceptual anachronisms, any interpretation of its findings must be contextualized according to translation studies' undeniable theoretical progress since these concepts were first introduced.

8.5. Avenues for further research

The robustness of the language status assessment model may assist in translation studies' ongoing process of "[o]pening itself more broadly to the whole world with its six or seven thousand languages" (Tymoczko 2018, 249). Accordingly, the most natural direction for future research is the investigation of the effects of language power relations on translations in other language pairs using this model. Given the language status assessment model's supposed ability to categorize all languages for any constellation, future studies could range from highly localized contexts to the global level, i.e., the entire "world language system" (De Swaan 2001). Assessing language power dynamics for the global language constellation may necessitate some fine-tuning to the model, however. With the United Nations' six official languages evenly ranked at

Level 0 (International), it is perhaps valid to reformulate the EGIDS typology within the language status assessment model so as to categorize English in the present age as the singularly global or “hypercentral” language (see Heilbron 1999; De Swaan 2001; Crystal 2003; Phillipson 2008; Seidlhofer 2013; House 2018). Although the model’s second-order sorting criterion would elevate English according to its number of worldwide speakers anyway, its widely-recognized global dominance compels its distinction from other merely international languages.

As alluded to in the previous section on the project’s overarching limitations, future corpus-based investigations of the impacts of language power relations in translation should incorporate other variables. The multiplicity of possible confounding variables may be tempered by strategically incorporating new variables anticipated to have the greatest effect on SL influence. Although language status has served as the expression of language power relevant to this thesis, language prestige deserves equal consideration. Conceptualized in terms of speakers’ attitudes and/or ideologies, language prestige proves more elusive in its capacity to be operationalized consistently in empirical research frameworks. It must necessarily be operationalized via direct engagement with speakers or translators in order to gauge their perceptions, such as by surveys on language attitudes. Beyond this other dimension of language power, corpus-based research on SL influence in translation should implement other common explanatory variables as suggested by De Sutter and Lefer (2020, 6), such as textual function, register, genre, and domain. Future works may also devise methods for systematically measuring the effects of language distance on SL influence, as well as lexical and structural priming, which would necessitate the use of parallel corpora.

The universality (i.e., cross-lingual comparability) of operationalizations is essential to future studies on the levels of SL influence in translations of diverse language pairs. Following the observation of the limitations in separating different linguistic levels, additional research should investigate SL influence on various linguistic levels jointly. For instance, future studies may further distinguish between the two loanword translation strategies described by Baker (2018, 34-35) – loanwords with an accompanying explanation (in the form of, e.g., a footnote or parenthetical gloss) and those without. The interaction between lexis and syntax in terms of SL influence on translated texts provides another opportunity. As Hoey (2011, 154) observes, translation researchers tend to focus on lexical features, only occasionally examining them in connection with grammar, generally maintaining a division between the two. Identifying common lexicogrammatical features among a diverse range of language pairs for the sake of assessing the impacts of language power dynamics may prove prohibitively difficult, however. It may be the case that the wider the range of languages to be examined, the more generic operationalizations of SL influence must be, such that systematic research on language status and linguistic patterns will tend to be restricted to individual linguistic levels as illustrated in this thesis' three constituent studies. Nonetheless, there are many potential research opportunities concentrated within these three linguistic dimensions.

Naturally, the SINC framework deserves the most scrutiny, as it is a completely new method and extremely ambitious in its scope. One possible way of testing the effectiveness of SINC in characterizing levels of syntactic interference/normalization in translation would be to assess the manner in which SINC scores relate to human judgments of fluency, as fluency ratings may approximate the perception of translations' syntactic similarity to comparable TL texts (i.e., normalization). However, fluency ratings also encompass not only syntactic structures but also their lexical contents, meaning that the actual collocations behind POS n-grams would need to be examined. It would also be beneficial to compare measurements of syntactic interference obtained via SINC with those obtained via Toral's (2019) perplexity-based method. Relatedly, the SINC methodology may be easily adjusted to compare the syntactic compositions of human- and machine-generated translations of the same source text, as a source text's POS n-gram frequency distribution could replace that of the comparable SL subcorpus in the SINC framework, and the human- and machine-generated translations could still be compared with a comparable TL subcorpus. Future work in this area may modify SINC calculations to include morphological information, as machine-translated text has been shown to conform more closely to the morphosyntactic structures of source texts than human-translated texts (Luo et al. 2024).

The operationalization of paratextual foreignization as translator-attributed footnotes is straightforwardly replicable across traditional literary texts and the like, yet it is unclear how the concept may be operationalized with respect to digitally-mediated translations, such as NMT/LLM output or web-scraped parallel corpora.

Perhaps the operationalizations of paratextual foreignization in these contexts are best formulated in accordance with Freeth's (2023, 420) notion of collateral paratextuality, which offers "a mechanism through which to analyse complex constellations of paratextual materials found in digital spaces." Rather than focusing on the criteria that define paratexts as such, his framing portrays paratextuality as a set of relations that may be "formed inadvertently, in parallel or in addition to another without the conscious intervention of the creator" (Freeth 2023, 426). To operationalize paratextual foreignization within this complexity of relations would undoubtedly be a highly complicated endeavor.

As the effects of language status were strongest in terms of translations' lexical features, this dimension offers the most promising area for future research. The uncovered tendency for language power relations to manifest as lexical interference in translation encourages further research on the translation of specialized terminology in multilingual settings and how it relates to status differences between languages. The inclusion of language status as an additional variable in corpus-based research on EU legal translation may prove particularly fruitful, as recent research has detected strong lexical interference in this context (see Pontrandolfo 2021; Seracini 2021). This domain has the advantage of offering pre-existing parallel corpora including a variety of languages. Contexts reflecting the intersection between legal translation and cultural translation may also present unique opportunities for researching the complexity of power relations in translation: Roshdy (2023) examines English translations of texts pertaining to Islamic finance law through a postcolonial lens, finding a strong tendency toward the use of loanwords. Such corpus-based studies provide a solid foundation for more interdisciplinary approaches toward the investigation of the interactions between language power relations and translation. Furthermore, the difficulties of harmonizing terminological resources with NMT output are increasingly of interest to researchers (see Čulo and Nitzke 2016; Haque et al. 2020; Doğru 2022; Bane et al. 2023), and language power relations may have an even more complex effect in this area (see Schneider 2022).

Empirical research on the effects of language power dynamics on the linguistic features of NMT and multilingual LLM output is undeniably vital to understanding translation and language contact in the digital age. Thompson et al. (2024) estimate that the web contains a staggering amount of machine-translated text, particularly for low-resource languages. This trend suggests that multilingual NMT systems trained on multitudes of indiscriminately web-scraped data could be inadvertently recycling and compounding the translation norms embedded in MT output. Moreover, most LLMs are overwhelmingly English-centric, with an estimated 93% of GPT-3's training data consisting of monolingual English text (Brown et al. 2020, 6). The emergent translation and multilingual capabilities of LLMs may therefore amplify the linguistic features of English, given the application of cross-lingual transfer learning to highly skewed training data sets. The rapidly increasing use of these technologies makes this research avenue particularly urgent.

There are also numerous opportunities for conducting applied research in contexts in which language power dynamics are expected to exert significant influence, especially those more overtly politicized than literary translation. In this regard, the international development sector is one particularly promising area for further inquiry: Bourdieu's linguistic capital provides an ideal framing for understanding how multilingual practices in international development characterize and perpetuate "unequal aid encounters" (Roth 2019). In his influential *Translation Theory and Development Studies*, Marais (2014, 7) lays out the book's goal of "situat[ing] translation as a factor in the political economy of the day, the day-to-day efforts of people to adapt to the power configurations within which they were born or had been forced," citing Gentzler's (2008) critique of translation studies' tendency to confine itself unnecessarily to literary texts. The suitability of development contexts for future inquiries into language power dynamics may be illustrated by way of reference to recent qualitative literature on the role of translation in development.

Mainstream development initiatives are overwhelmingly dominated by non-governmental organizations (NGOs) based in the Global North and favor major international languages – primarily English and, to a lesser extent, French and Spanish (Tesseur 2022). Qualitative research confirms that development practices and discourse are shaped by hegemonic languages, often placing the onus of translation on local beneficiaries in the Global South (Footitt 2019, 391). Translation in this context is thus carried out under a strong pressure for beneficiaries of NGO-led projects to adapt to the esoteric vocabulary of the development sector, creating a barrier to meaningful participation or input when such concepts prove incompatible with local language systems (Tesseur and Crack 2020). Naturally, the adaptation of donor terminology tends to take the form of loanwords. Future studies may investigate the frequency, types, and contexts of loanwords in light of inequalities between language communities in development settings, even combining empirical (i.e., corpus-based) methods with more qualitative, ethnographic approaches.

In one particularly relevant case study, Tesseur and Crack (2020) conduct a series of interviews to demonstrate how English buzzwords conceived by NGOs and international project funders shape development strategies and discourse, permeating efforts to implement projects in Kyrgyzstan and Malawi. In Kyrgyzstan, Kyrgyz is much more widely spoken than Russian, though the latter is more high-status as it is the preferred language in domains such as government and commerce (Tesseur and Crack 2020, 29). Though English competency is exceedingly rare among Kyrgyzs, the strong preference toward English among international NGOs and development funding bodies has resulted in a high number of English loanwords infiltrating discourse among local NGOs in the country, where English buzzwords enter Russian as loanwords or calques and are subsequently “transposed” from Russian into Kyrgyz (Tesseur and Crack 2020, 30-31). Interviewees cite the translation of key English terms (e.g., “advocacy” and “stakeholders”) into Kyrgyz as a major challenge in their work, with some asserting that these “translation issues” contribute substantially to the “limited development of Kyrgyz as a language” (ibid., 31). The social and political complexities of these situations evince the merits of House’s (2011, 206) call for translation scholars to treat quantitative, corpus-based research as merely a launchpad for more interdisciplinary translation research.

The development context portrayed here demonstrates not only the potentially adverse social consequences of lexical interference induced by language power asymmetries, but also the interrelatedness of translation and other forms of language contact in these settings; multilingual development encounters involve cross-linguistic communication in many forms, including written translation, oral interpretation, language education, bi- and even trilingualism. As such, it is again worth contextualizing translation in relation to other forms of language contact, where SL influence – whether on the lexical or other linguistic levels – reflects a type of cross-linguistic influence (CLI).

8.6. Concluding remarks

This doctoral thesis began by reflecting on the historical competition between languages, situating translation among the many forms of language contact in which these power struggles emerge. The multitude of perspectives on the intersection between language and power in sociology, sociolinguistics, translation studies, natural language processing, and other related fields indicates any attempted empirical description of this relationship, such as the one presented in this project, reflects a necessarily limited view of an irreducibly complex and interdisciplinary subject. For similar reasons, there persists a well-worn skepticism about translation studies' ability to develop meaningfully predictive abilities with respect to the linguistic features of translated texts, especially given the increasingly popular view that posits translation as an irreducibly complex process (see Marais and Meylaerts 2018). These concerns may be alleviated with a renewed emphasis on the original aim of descriptive translation research, which aspired to the continual refinement of tentatively formulated, probabilistic, and context-dependent laws of translation.

To acknowledge and embrace the irreducible complexity of translation does not require dispensing with the simpler models that characterized earlier translation theories. Instead, these dichotomous approaches may be "extend[ed] into a higher dimensionality" in which translation phenomena (e.g., interference and normalization) may not only be contrasted against their presumed opposites but also placed within entirely different perspectives (Blumczynski and Hassani 2019, 15-16). This multidimensional approach echoes the notion that corpus research "should not be seen as an end in itself, but as a starting point for continuing richly (re)contextualized qualitative work" (House 2011, 206).

Corpus-based methods entered translation studies as a direct response to the limitations of the reliance on close readings and isolated examples which underpinned previous translation research. The primary strength of corpus methodology is that it provides one of the most highly effective means of identifying and characterizing CLI in various forms of language contact – namely, the comparison of linguistic patterns' frequencies within and across languages over time (Kotze 2021). Moreover, cross-lingual frequency comparisons and their diachronic trajectories reflect the intended probabilistic nature of translation laws given their intrinsically relative nature. Historical corpora have allowed linguists to track CLI over time, particularly with respect to significant developments in the mass media landscape – e.g., radio and television – and the manner in which they create new forms of language and cultural contact (Hoffer 2002).

Today's increasing automation of multilingual communication perhaps warrants distinguishing multilingual AI – including not only NMT but also LLMs, given their multilingual capabilities – as a unique form of language contact, set apart from human translation by its technical complexity, opacity, and thus far highly unpredictable behaviors. These technologies are being used more and more to translate and generate multilingual digital content, with Google Translate alone processing some 150 billion words daily, handily surpassing the collective outputs of professional translators (Asscher 2022, 1). Amidst the already massive infrastructures and expanding reach of language-based AI, the digital sphere represents another link in a series of intertangled terrains upon which territorial struggles between languages continuously unfold. Renewed emphasis on descriptive translation studies' theoretical foundations as well as the descriptive power and interdisciplinary compatibility of corpus methodology could uphold efforts to excavate fossilized evidence of these conflicts, regardless of their sites.

9. References

Adegbija, Efurosibina E. 1994. *Language Attitudes in Sub-Saharan Africa: A Sociolinguistic Overview*. Clevedon: Multilingual Matters.

Aikhenvald, Alexandra Y. 2007. “Grammars in Contact: A Cross-Linguistic Typology.” In Grammars in Contact: A Cross-Linguistic Typology, edited by Alexandra Y. Aikhenvald and R. M. W. Dixon, 4:1–66. Explorations in Linguistic Typology Series. Oxford: Oxford University Press.

Alexander, Ronelle. 2006. *Bosnian, Croatian, Serbian, a Grammar: With Sociolinguistic Commentary*. Madison: University of Wisconsin Press.

Alvstad, Cecilia. 2012. “The Strategic Moves of Paratexts: World Literature through Swedish Eyes.” *Translation Studies* 5 (1): 78–94. https://doi.org/10.1080/14781700.2012.628817.

Ammon, Ulrich, and Marlis Hellinger. 1992. *Status Change of Languages*. Berlin: De Gruyter.

Ammon, Ulrich. 1989. “Towards a Descriptive Framework for the Status/Function (Social Position) of a Language Within a Country.” In *Status and Function of Languages and Language Varieties*, edited by Ulrich Ammon, 21–106. Berlin: De Gruyter.

Ammon, Ulrich. 1992. “On the Status and Changes in the Status of German as a Language of Diplomacy.” In *Status Change of Languages*, edited by Ulrich Ammon and Marlis Hellinger, 421–38. Berlin: De Gruyter.

Asad, Talal. 1986. “The Concept of Cultural Translation in British Social Anthropology.” In *Writing Culture : The Poetics and Politics of Ethnography : A School of American Research Advanced Seminar*, edited by James Clifford and George E. Marcus, 141–64. Berkeley: University of California Press.

Asscher, Omri. 2022. “The explanatory power of descriptive translation studies in the machine translation era.” *Perspectives*: 1–17. https://doi.org/10.1080/0907676X.2022.2136005.

Asscher, Omri. 2023. “The position of machine translation in translation studies: A definitional perspective.” *Translation Spaces* 12 (1): 1–20. https://doi.org/10.1075/ts.22035.ass.

Assis Rosa, Alexandra. 2010. “Descriptive Translation Studies (DTS).” In *Handbook of Translation Studies*, 94–104. Amsterdam: John Benjamins.

Assis Rosa, Alexandra. 2023. “Descriptive Approaches.” In *The Routledge Handbook of Translation Theory and Concepts*, 185–207. Routledge Handbooks in Translation and Interpreting Studies. Abingdon: Routledge.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. “Corpus Design Criteria.” *Literary and Linguistic Computing* 7 (1): 1–16. https://doi.org/10.1093/llc/7.1.1.

Backus, Ed. 2015. “A Usage-Based Approach to Code-Switching: The Need for Reconciling Structure and Function.” In *Code-Switching Between Structural and Sociolinguistic Perspectives:*, edited by Gerald Stell and Kofi Yakpo, 43:19–38. linguae & litterae. Berlin: De Gruyter. https://doi.org/10.1515/9783110346879.

Baer, Brian James. 2023. “Cultural Approaches.” In *The Routledge Handbook of Translation Theory and Concepts*, 224–40. Routledge Handbooks in Translation and Interpreting Studies. Abingdon: Routledge.

Baker, Mona. 1993. “Corpus Linguistics and Translation Studies: Implications and Applications.” In *Text and Technology: In Honour of John Sinclair*, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–50. Amsterdam: John Benjamins.

Baker, Mona. 1996. “Corpus-based translation studies: The challenges that lie ahead.” In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, edited by Harold Somers, 175–86. Amsterdam: John Benjamins.

Baker, Mona. 2000. “Towards a Methodology for Investigating the Style of a Literary Translator.” *Target. International Journal of Translation Studies* 12 (2): 241–66. https://doi.org/10.1075/target.12.2.04bak.

Baker, Mona. 2004. “A corpus-based view of similarity and difference in translation.” *International Journal of Corpus Linguistics* 9 (2): 167–93. https://doi.org/10.1075/ijcl.9.2.02bak.

Baker, Mona. 2018. *In Other Words: A Coursebook on Translation*. 3rd ed. Abingdon: Routledge.

Baker, Mona, and Maeve Olohan. 2000. “Reporting That in Translated English:

Evidence for Subconscious Processes of Explicitation?" *Across Languages and Cultures* 1 (2): 141–58. https://doi.org/10.1556/Acr.1.2000.2.1.

Bane, Fred, Anna Zaretskaya, Tània Blanch Miró, Celia Soler Uguet, and João Torres. 2023. "Coming to Terms with Glossary Enforcement: A Study of Three Approaches to Enforcing Terminology in NMT." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 345–53. Tampere: European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.34.

Batchelor, Kathryn. 2018. *Translation and Paratexts*. Translation Theories Explored. Boca Raton: Routledge.

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, et al. 2022. "Building Machine Translation Systems for the Next Thousand Languages." arXiv. https://doi.org/10.48550/arXiv.2205.03983.

Becher, Viktor, Juliane House, and Svenja Kranich. 2009. "Convergence and Divergence of Communicative Norms through Language Contact in Translation." In *Convergence and Divergence in Language Contact Situations*, edited by Kurt Braunmüller and Juliane House, 125–52. Amsterdam: John Benjamins.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. "Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French." *Computer Speech & Language* 49 (May): 52–70. https://doi.org/10.1016/j.csl.2017.11.004.

Bernardini, Silvia, and Adriano Ferraresi. 2011. "Practice, Description and Theory Come Together – Normalization or Interference in Italian Technical Translation?" *Meta : Journal Des Traducteurs / Meta: Translators' Journal* 56 (2): 226–46. https://doi.org/10.7202/1006174ar.

Berruto, Gaetano. 2018. "The Languages and Dialects of Italy." In *Sociolinguistics; Linguistic Varieties; Romance Languages*, edited by Wendy Ayres-Bennett and Janice Carruthers, 494–525. Berlin: De Gruyter.

Biber, Douglas. 1993. "Representativeness in Corpus Design." *Literary and Linguistic Computing* 8 (4): 243–57. https://doi.org/10.1093/llc/8.4.243.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Bilodeau, Isabelle. 2019. "Bending Conventions: Agency and Self-Portrayals in Japanese Translator Commentary." *Japan Forum* 31 (1): 64–85. https://doi.org/10.1080/09555803.2018.1530280.

Bisazza, Arianna, and Marcello Federico. 2016. "A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena." *Computational Linguistics* 42 (2): 163–205. https://doi.org/10.1162/COLI_a_00245.

Blasi, Damián, Antonios Anastasopoulos, and Graham Neubig. 2021. "Systematic Inequalities in Language Technology Performance across the World's Languages." arXiv. https://doi.org/10.48550/arXiv.2110.06733.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 5454–76. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485.

Blommaert, Jan. 2015. "Pierre Bourdieu and Language in Society." In *Handbook of Pragmatics*, edited by Jan-Ola Östman and Jef Verschueren, 1–16. Working Papers in Urban Language and Literacies. Amsterdam: John Benjamins.

Blumczynski, Piotr, and Ghodrat Hassani. 2019. "Towards a meta-theoretical model for translationA multidimensional approach." *Target. International Journal of Translation Studies* 31 (3): 328–51. https://doi.org/10.1075/target.17031.blu.

Bourdieu, Pierre. 1986. "The Forms of Capital." In *Handbook of Theory and Research for the Sociology of Education*, edited by John G. Richardson, translated by Richard Nice, 241–58. Greenwood Press.

Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Edited by John B. Thompson. Translated by Gino Raymond and Matthew Adamson. Cambridge: Harvard University Press.

Brenzinger, M., A. Yamamoto, N. Aikawa, D. Koundiouba, A. Minasyan, A. Dwyer, C. Grinevald, et al. 2003. "Language Vitality and Endangerment." In . Paris:

UNESCO Ad Hoc Expert Group on Endangered Languages. Accessed March 23, 2024. https://ich.unesco.org/doc/src/00120-EN.pdf.

Breuilly, John. 2017. "Modern Empires and Nation-States." *Thesis Eleven* 139 (1): 11–29. https://doi.org/10.1177/0725513617700036.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Brooke, Julian, Adam Hammond, and Graeme Hirst. 2015. "GutenTag: An NLP-Driven Tool for Digital Humanities Research in the Project Gutenberg Corpus." In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 42–47. Denver: Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-0705.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, 33: 1877–1901.

Buts, Jan, and Henry Jones. 2021. "From Text to Data: Mediality in Corpus-Based Translation Studies." *MonTI. Monografías de Traducción e Interpretación*, 13: 301–29. https://doi.org/10.6035/MonTI.2021.13.10.

Byrne, Eoin. 2018. "'Éistear Le Mo Ghlór!': Máirtín Ó Cadhain's Cré Na Cille and Postcolonial Modernisms." *Irish Studies Review* 26 (3): 335–46. https://doi.org/10.1080/09670882.2018.1474727.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. "Re-Evaluating the Role of Bleu in Machine Translation Research." In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*, 249–56. Association for Computational Linguistics. Accessed March 23, 2024. https://www.research.ed.ac.uk/en/publications/re-evaluating-the-role-of-bleu-in-machine-translation-research.

Casanova, Pascale. 2002. "Consécration et accumulation de capital littéraire. La traduction comme échange inégal." *Actes de la recherche en sciences sociales* 144 (4): 7–20. https://doi.org/10.3917/arss.144.0007.

Casanova, Pascale. 2002. “Consecration and Accumulation of Literary Capital: Translation as Unequal Exchange.” Translated by Siobhan Brownlie. In Venuti, Lawrence. *The Translation Studies Reader*. 2021. 4th ed. 407–423. London: Routledge.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. “Approaches to Human and Machine Translation Quality Assessment.” In *Translation Quality Assessment*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 1:9–38. Machine Translation: Technologies and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7_2.

Castilho, Sheila, and Natália Resende. 2022. “Post-Editese in Literary Translations.” *Information* 13 (2): 1–22. https://doi.org/10.3390/info13020066.

Catford, John Cunnison. 1965. *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. London: Oxford University Press.

Chen, Baoguo, Yuefang Jia, Zhu Wang, Susan Dunlap, and Jeong-Ah Shin. 2013. “Is Word-Order Similarity Necessary for Cross-Linguistic Structural Priming?” *Second Language Research* 29 (4): 375–89. https://doi.org/10.1177/0267658313491962.

Chesterman, Andrew. 2004. “Beyond the Particular.” In *Translation Universals: Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki, 33–49. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.48.

Chesterman, Andrew. 2016. *Memes of Translation: The Spread of Ideas in Translation Theory. Revised Edition*. 2nd ed. Amsterdam: John Benjamins.

Chlumská, Lucie. 2018. “Prominent POS-Grams and n-Grams in Translated Czech in the Mirror of the English Source Texts.” In *Taming the Corpus: From Inflection and Lexis to Interpretation*, edited by Masako Fidler and Václav Cvrček, 99–117. Quantitative Methods in the Humanities and Social Sciences. Cham: Springer. https://doi.org/10.1007/978-3-319-98017-1_6.

Choudhury, Monojit, and Amit Deshpande. 2021. “How Linguistically Fair Are Multilingual Pre-Trained Language Models?” *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (14): 12710–18. https://doi.org/10.1609/aaai.v35i14.17505.

Coldiron, A.E.B. 2012. “Visibility Now: Historicizing Foreign Presences in Translation.” *Translation Studies* 5 (2): 189–200. https://doi.org/10.1080/14781700.2012.663602.

Conrad, Sebastian. 2011. *German Colonialism: A Short History*. Translated by Sorcha O’Hagan. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139030618.

Córdoba Serrano, María Sierra. 2010. “Translation as a Measure of Literary Domination: The Case of Quebec Literature Translated in Spain (1975-2004).” *MonTi: Monografías de Traducción e Interpretación*, no. 2: 249–81. https://doi.org/10.6035/MonTI.2010.2.12.

Corpas Pastor, Gloria, and Míriam Seghiri Domínguez. 2010. “Size Matters: A Quantitative Approach to Corpus Representativeness.” In *Lengua, Traducción, Recepción En Honor de Julio César Santoyo*, edited by Rosa Rabadán, 111–45. León: Publicaciones Universidad de León.

Costa-Jussà, Marta R., and Mireia Farrús. 2014. “Statistical Machine Translation Enhancements through Linguistic Levels: A Survey.” *ACM Computing Surveys* 46 (3): 42:1-42:28. https://doi.org/10.1145/2518130.

Croatian Bureau of Statistics - Državni Zavod za Statistiku Republike Hrvatski. 2018. “2018 Statistical Yearbook of the Republic of Croatia.” Zagreb: Croatian Bureau of Statistics. Accessed March 23, 2024. https://www.dzs.hr/Eng/Publication/stat_year.htm.

Cronin, Michael. 1998. “The Cracked Looking Glass of Servants.” *The Translator* 4 (2): 145–62. https://doi.org/10.1080/13556509.1998.10799017.

Crystal, David. 2003. *English as a Global Language*. 2nd ed. Cambridge: Cambridge University Press.

Čulo, Oliver, and Jean Nitzke. 2016. “Patterns of Terminological Variation in Post-Editing and of Cognate Use in Machine Translation in Contrast to Human Translation.” In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, 106–14. https://aclanthology.org/W16-3401.

D’Alessandro, Roberta. 2021. “Syntactic Change in Contact: Romance.” *Annual Review of Linguistics* 7 (1): 309–28. https://doi.org/10.1146/annurev-linguistics-

011619-030311.

Daems, Joke, Sonia Vandepitte, Robert J. Hartsuiker, and Lieve Macken. 2017. "Identifying the Machine Translation Error Types with the Greatest Impact on Post-Editing Effort." *Frontiers in Psychology* 8. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.01282.

Danchev, Andrei. 1984. "Translation and Syntactic Change." In *Historical Syntax*, edited by Jacek Fisiak, 47–60. Berlin: De Gruyter.

Dancey, Christine, and John Reidy. 2020. *Statistics Without Maths for Psychology*. Harlow: Pearson Education.

Darquennes, Jeroen. 2006. "German as a Lingua Franca." *Annual Review of Applied Linguistics* 26: 61–77. https://doi.org/10.1017/S0267190506000043.

Delaere, Isabelle, and Gert de Sutter. 2017. "Variability of English Loanword Use in Belgian Dutch Translations : Measuring the Effect of Source Language and Register." In *Empirical Translation Studies : New Methodological and Theoretical Traditions*, 300:81–112. Berlin: De Gruyter. https://doi.org/10.1515/9783110459586-004.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* 47 (2): 255–308. https://doi.org/10.1162/coli_a_00402.

de Sutter, Gert, Patrick Goethals, Torsten Leuschner, and Sonia Vandepitte. 2012. "Towards Methodologically More Rigorous Corpus-Based Translation Studies." *Across Languages and Cultures* 13 (2): 137–43. https://doi.org/10.1556/Acr.13.2012.2.1.

de Sutter, Gert, and Marie-Aude Lefer. 2020. "On the Need for a New Research Agenda for Corpus-Based Translation Studies: A Multi-Methodological, Multifactorial and Interdisciplinary Approach." *Perspectives* 28 (1): 1–23. https://doi.org/10.1080/0907676X.2019.1611891.

de Sutter, Gert, and Marc Van de Velde. 2008. "Do the Mechanisms That Govern Syntactic Choices Differ between Original and Translated Language? A Corpus-Based Translation Study of PP Extraposition in Dutch and German." In *Proceedings of the International Symposium on Using Corpora in Contrastive* *and Translation Studies (UCCTS)*, 1–38. Hangzhou. http://hdl.handle.net/1854/LU-537751.

de Sutter, Gert, and Eline Vermeire. 2019. "Grammatical Optionality in Translations: A Multifactorial Corpus Analysis of That/Zero Alternation in English Using the MuPDAR Approach." In *New Empirical Perspectives on Translation and Interpreting*, 13–37. New York: Routledge.

de Swaan, Abram. 2001. *Words of the World: The Global Language System*. Maldwell: Blackwell Publishers.

Dimitriu, Rodica. 2009. "Translators' Prefaces as Documentary Sources for Translation Studies." *Perspectives* 17 (3): 193–206. https://doi.org/10.1080/09076760903255304.

Direction générale Statistique - Statistics Belgium. 2017. "Recensement- Census." Accessed March 23, 2024. https://statbel.fgov.be/fr/propos-de-statbel/que-faisons-nous/recensement-census.

Doğru, Gökhan. 2022. "Translation Quality Regarding Low-Resource, Custom Machine Translations: A Fine-Grained Comparative Study on Turkish-to-English Statistical and Neural Machine Translation Systems." *Istanbul University Journal of Translation Studies* 0 (17): 95–115. https://doi.org/10.26650/iujts.2022.1182687.

e-Lektire. 2024. "Portal e-lektire i pilot projekt e-Škole." 2024. Accessed March 23, 2024. https://lektire.skole.hr/e-lektire-i-pilot-projekt-e-skole/.

Edwards, John. 1994. *Multilingualism*. London: Taylor & Francis Group.

Edwards, John. 1996. "Language, prestige and stigma." In *Kontaktlinguistik*, 12:703–8. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin: De Gruyter.

Even-Zohar, Itamar. 1979. "Polysystem Theory." *Poetics Today* 1 (1): 287–310. https://doi.org/10.2307/1772051.

Even-Zohar, Itamar. 1990. "Polysystem Studies." *Poetics Today* 11 (1).

Evert, Stefan, and Stella Neumann. 2017. "The Impact of Translation Direction on Characteristics of Translated Texts. A Multivariate Analysis for English and German." In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47–80. Trends in Linguistics Studies and

Monographs. Berlin: De Gruyter. https://doi-org.dcu.idm.oclc.org/10.1515/9783110459586-003.

Faisal, Fahim, Yinkai Wang, and Antonios Anastasopoulos. 2022. "Dataset Geography: Mapping Language Data to Language Users." arXiv. https://doi.org/10.48550/arXiv.2112.03497.

Fhrighil, Ríóna Ní, Anne O'Connor, and Michelle Milan. 2020. "Translation in Ireland: Historical and Contemporary Perspectives." *Translation Studies* 13 (2): 129–37. https://doi.org/10.1080/14781700.2020.1751261.

Fishman, Joshua A. 1991. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Multilingual Matters 76. Clevedon: Multilingual Matters.

Footitt, Hilary. 2019. "Translation and the Contact Zones of International Development." *The Translator* 25 (4): 385–400. https://doi.org/10.1080/13556509.2020.1758505.

Frankenberg-Garcia, Ana. 2005. "A Corpus-Based Study of Loan Words in Original and Translated Texts." In *Proceedings from the Corpus Linguistics Conference Series; Corpus Linguistics 2005*, 19. Birmingham.

Freeth, Peter Jonathan. 2022. "Beyond Invisibility: The Position and Role of the Literary Translator in the Digital Paratextual Space." PhD, University of Leeds. https://etheses.whiterose.ac.uk/30818/.

Freeth, Peter Jonathan. 2023. "Between Consciously Crafted and the Vastness of Context: Collateral Paratextuality and Its Implications for Translation Studies." *Translation Studies* 16 (3): 419–35. https://doi.org/10.1080/14781700.2023.2194882.

Gardner-Chloros, Penelope. 2009. *Code-Switching*. Cambridge: Cambridge University Press.

Gardner-Chloros, Penelope, and Daniel Weston. 2015. "Code-Switching and Multilingualism in Literature." *Language and Literature* 24 (3): 182–93. https://doi.org/10.1177/0963947015585065.

Gaspari, Federico, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. "Introducing the Digital Language Equality Metric: Technological Factors." In *Workshop Towards Digital Language*

Equality within the 13th Language Resources and Evaluation Conference, 1–12. Marseille, France: European Language Resources Association (ELRA). https://aclanthology.org/2022.tdle-1.1.

Genette, Gerard. 1997. *Paratexts: Thresholds of Interpretation*. Translated by Jane E. Lewin. Cambridge: Cambridge University Press.

Gentzler, Edwin. 2008. *Translation and Identity in the Americas: New Directions in Translation Theory*. London: Routledge. https://doi.org/10.4324/9780203609941.

Gentzler, Edwin, and Maria Tymoczko. 2002. "Introduction." In *Translation and Power*, edited by Maria Tymoczko and Edwin Gentzler, xi–xxviii. Amherst: University of Massachusetts Press.

Gibadullin, Ilshat, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. "A Survey of Methods to Leverage Monolingual Data in Low-Resource Neural Machine Translation." In *International Conference on Advanced Technologies for Humanity*. https://doi.org/10.48550/arXiv.1910.00373.

Gómez Capuz, Juan. 1997. "Towards a Typological Classification of Linguistic Borrowing (Illustrated with Anglicisms in Romance Languages)." *Revista Alicantina de Estudios Ingleses*, no. 10: 81. https://doi.org/10.14198/raei.1997.10.08.

Görlach, Manfred. 2003. *English Words Abroad*. Vol. 7. Terminology and Lexicography Research and Practice. Amsterdam: John Benjamins.

Gottlieb, Henrik. 2004. "Danish Echoes of English." *Nordic Journal of English Studies* 3 (2): 39–66.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation." *Transactions of the Association for Computational Linguistics* 10: 522–38. https://doi.org/10.1162/tacl_a_00474.

Gries, Stefan Th., and Gerrit Jan Kootstra. 2017. "Structural Priming within and across Languages: A Corpus-Based Perspective." *Bilingualism: Language and Cognition* 20 (2): 235–50. https://doi.org/10.1017/S1366728916001085.

Hadley, James. 2017. "Indirect Translation and Discursive Identity: Proposing the

Concatenation Effect Hypothesis." *Translation Studies* 10 (2): 183–97. https://doi.org/10.1080/14781700.2016.1273794.

Hadley, James Luke. 2023. *Systematically Analysing Indirect Translations: Putting the Concatenation Effect Hypothesis to the Test*. New York: Routledge. https://doi.org/10.4324/9780429282768.

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. "Survey of Low-Resource Machine Translation." *Computational Linguistics* 48 (3): 673–732. https://doi.org/10.1162/coli_a_00446.

Hansen-Schirra, Silvia. 2011. "Between Normalization and Shining-through: Specific Properties of English-German Translations and Their Influence on the Target Language." In *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*, edited by Svenja Kranich, 135–62. Amsterdam: John Benjamins.

Haque, Rejwanul, Mohammed Hasanuzzaman, and Andy Way. 2020. "Analysing Terminology Translation Errors in Statistical and Neural Machine Translation." *Machine Translation* 34 (2): 149–95. https://doi.org/10.1007/s10590-020-09251-z.

Haspelmath, Martin. 2009. "Lexical Borrowing: Concepts and Issues." In *Loanwords in the World's Languages: A Comparative Handbook*, edited by Martin Haspelmath and Uri Tadmor. Berlin: De Gruyter.

Haugen, Einar. 1950. "The Analysis of Linguistic Borrowing." *Language* 26 (2): 210–31. https://doi.org/10.2307/410058.

Haugen, Einar. 1972. *The Ecology of Language*. Redwood City: Stanford University Press.

He, Shilin, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael R. Lyu, and Shuming Shi. 2019. "Towards Understanding Neural Machine Translation with Word Importance." arXiv. https://doi.org/10.48550/arXiv.1909.00326.

Heilbron, Johan. 1999. "Towards a Sociology of Translation: Book Translations as a Cultural World-System." *European Journal of Social Theory* 2 (4): 429–44. https://doi.org/10.1177/136843199002004002.

Heilbron, Johan, and Gisèle Sapiro. 2007. "Outline for a Sociology of Translation:

Current Issues and Future Prospects." In *Constructing a Sociology of Translation*, 93–107. John Benjamins. https://www.jbe-platform.com/content/books/9789027292063-btl.74.07hei.

Hermans, Theo. 1999. *Translation in Systems: Descriptive and Systemic Approaches Explained*. Manchester: St. Jerome.

Hermans, Theo. 2007. *The Conference of the Tongues*. Manchester: St. Jerome.

Hermans, Theo. 2019. "Descriptive Translation Studies." In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker and Gabriela Saldanha, 3rd ed., 143–47. London: Routledge. https://doi.org/10.4324/9781315678627.

Hickey, Raymond. 2010. "Contact and Language Shift." In *The Handbook of Language Contact*, edited by Raymond Hickey, 149–69. Blackwell Handbooks in Linguistics. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781444318159.ch7.

Hickey, Raymond, and Carolina P. Amador-Moreno. 2020. "Linguistic Identities in Ireland – Contexts and Issues." In *Irish Identities: Sociolinguistic Perspectives*, edited by Raymond Hickey and Carolina P. Amador-Moreno, 3–20. Berlin: De Gruyter. https://doi.org/10.1515/9781501507687.

Hindley, Reg. 1990. *The Death of the Irish Language: A Qualified Obituary*. London: Routledge.

Hjorth-Andersen, Christian. 2006. "The Relative Importance of the European Languages." In *14th International Conference of the ACEI*. Vienna. Accessed March 23, 2024. https://www.economics.ku.dk/research/publications/wp/2006/0623.pdf/.

Hoey, Michael. 2011. "Lexical Priming and Translation." In *Corpus-Based Translation Studies: Research and Applications*, edited by Alet Kruger, Kim Wallmach, and Jeremy Munday, 153–68. Bloomsbury Advances in Translation. London: Continuum International.

Hoffer, Bates L. 2002. "Language Borrowing and Language Diffusion: An Overview." *Intercultural Communication Studies* 11 (4): 1–37.

House, Juliane. 2011. "Using Translation and Parallel Text Corpora to Investigate the Influence of Global English on Textual Norms in Other Languages." In *Corpus-Based Translation Studies: Research and Applications*, edited by Alet

Kruger, Kim Wallmach, and Jeremy Munday, 189–208. London: Continuum International.

House, Juliane. 2018. "The Impact of English as a Global Lingua Franca on Intercultural Communication." In *Intercultural Communication in Asia: Education, Language and Values*, edited by Andy Curtis and Roland Sussex, 97–114. Multilingual Education. Cham: Springer International. https://doi.org/10.1007/978-3-319-69995-0_6.

Howard, Rosaleen, Raquel de Pedro Ricoy, and Luis Andrade Ciudad. 2018. "Translation Policy and Indigenous Languages in Hispanic Latin America." *International Journal of the Sociology of Language* 2018 (251): 19–36. https://doi.org/10.1515/ijsl-2018-0002.

Humbley, John. 1974. "Vers une typologie de l'emprunt linguistique." *Cahiers de Lexicologie* 25: 46–70.

Inghilleri, Moira. 2023. "Sociological Approaches." In *The Routledge Handbook of Translation Theory and Concepts*, 241–62. Routledge Handbooks in Translation and Interpreting Studies. Abingdon: Routledge.

Institut National d'Études Démographiques. n.d. "Population de la France-séries longues." Ined - Institut National d'études Démographiques. Accessed March 23, 2024. https://www.ined.fr/en/everything_about_population/data/online-databases/france-population-long-series/.

Istituto Nazionale di Statistica. 2012. "Capitolo 2: Popolazione." In *L'Italia in 150 anni. Sommario di statistiche storiche 1861-2010*. Accessed March 23, 2024. https://www.istat.it/it/archivio/228440.

Ives, Peter. 2006. "'Global English': Linguistic Imperialism or Practical Lingua Franca?" *Studies in Language and Capitalism* 1: 121–41.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, et al. 2017. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." arXiv. http://arxiv.org/abs/1611.04558.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." arXiv. https://doi.org/10.48550/arXiv.2004.09095.

Kahane, Henry. 1986. “A Typology of the Prestige Language.” *Language* 62 (3): 495–508. https://doi.org/10.2307/415474.

Kanana Erastus, Fridah. 2013. “Examining African Languages as Tools for National Development: The Case of Kiswahili.” *Journal of Pan African Studies* 6 (6): 41–68.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation: A Corpus-Based Study*. Manchester: St. Jerome.

Kenny, Dorothy. 2022. “Human and Machine Translation.” In *Machine Translation for Everyone*, edited by Dorothy Kenny, 121–40. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6653406.

Kenny, Dorothy, and Marion Winters. 2020. “Machine Translation, Ethics and the Literary Translator’s Voice.” *Translation Spaces* 9 (1): 123–49. https://doi.org/10.1075/ts.00024.ken.

Kew, Tannon, Florian Schottmann, and Rico Sennrich. 2023. “Turning English-Centric LLMs Into Polyglots: How Much Multilinguality Is Needed?” arXiv. https://arxiv.org/abs/2312.12683v1.

Kocmi, Tom, and Christian Federmann. 2023. “Large Language Models Are State-of-the-Art Evaluators of Translation Quality.” arXiv. https://doi.org/10.48550/arXiv.2302.14520.

Koehn, Philipp, and Rebecca Knowles. 2017. “Six Challenges for Neural Machine Translation.” In *Proceedings of the First Workshop on Neural Machine Translation*, 28–39. Vancouver: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3204.

Kotze, Haidee. 2021. “Translation, Contact Linguistics and Cognition.” In *The Routledge Handbook of Translation and Cognition*, edited by Fábio Alves and Arnt Lykke Jakobsen, 113–32. Routledge Handbooks in Translation and Interpreting Studies. London: Routledge.

Kroch, Anthony S. 2001. “Syntactic Change.” In *The Handbook of Contemporary Syntactic Theory*, edited by Mark Baltin and Chris Collins, 698–729. Blackwell Handbooks in Linguistics. Oxford: Blackwell Publishers. https://doi.org/10.1002/9780470756416.ch22.

Kruger, Haidee. 2017. “The Effects of Editorial Intervention. Implications for

Studies of the Features of Translated Language." In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, edited by Gert de Sutter, Marie-Aude Lefer, and Isabelle Delaere, 113–55. Berlin: De Gruyter, Inc.

Kumar, Krishan. 2010. "Nation-States as Empires, Empires as Nation-States: Two Principles, One Practice?" *Theory and Society* 39 (2): 119–43. https://doi.org/10.1007/s11186-009-9102-8.

Laviosa, Sara. 2008. "Universals." In *Routledge Encyclopedia of Translation Studies*, edited by Mona Baker and Gabriela Saldanha, 2nd ed., 306–10. London: Routledge.

Lee, David. 2001. "Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle." *Language Learning and Technology* 5 (3): 37–72.

Lefer, Marie-Aude, and Svetlana Vogeleer. 2013. "Interference and Normalization in Genre-Controlled Multilingual Corpora." *Belgian Journal of Linguistics* 27: 1–21.

Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012. "Language Models for Machine Translation: Original vs. Translated Texts." *Computational Linguistics* 38 (4): 799–825. https://doi.org/10.1162/COLI_a_00111.

Lewis, Melvyn, and Gary Simons. 2010. "Assessing Endangerment: Expanding Fishman's GIDS." *Revue Roumaine de Linguistique* 55 (2): 103–20. https://doi.org/10.1017/CBO9780511783364.003.

Lin, Yu-Hsiang, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, et al. 2019. "Choosing Transfer Languages for Cross-Lingual Learning." arXiv. https://doi.org/10.48550/arXiv.1905.12688.

López-Arroyo, Belén. 2020. "Can Comparable Corpora Be Compared?" *Ibérica*, no. 39 (January): 43–68. https://doi.org/10.17398/2340-2784.39.43.

Luo, Jiaming, Colin Cherry, and George Foster. 2024. "To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation." arXiv. https://doi.org/10.48550/arXiv.2401.01419.

Mackey, William F. 1989. "Determining the Status and Function of Languages in Multinational Societies." In *Status and Function of Languages and Language*

Varieties, 3–20. Grundlagen Der Kommunikation Und Kognition / Foundations of Communication and Cognition. Berlin: De Gruyter. https://doi.org/10.1515/9783110860252.3.

Maia, Belinda. 1998. “Word Order and the First Person Singular in Portuguese and English.” *Meta : Journal Des Traducteurs / Meta: Translators’ Journal* 43 (4): 589–601. https://doi.org/10.7202/003539ar.

Maier, Robert M., Martin J. Pickering, and Robert J. Hartsuiker. 2017. “Does Translation Involve Structural Priming?” *Quarterly Journal of Experimental Psychology* 70 (8): 1575–89. https://doi.org/10.1080/17470218.2016.1194439.

Malmkjær, Kirsten. 2023. “Linguistic Approaches.” In *The Routledge Handbook of Translation Theory and Concepts*, 155–68. Routledge Handbooks in Translation and Interpreting Studies. Abingdon: Routledge.

Marais, Kobus. 2014. *Translation Theory and Development Studies: A Complexity Theory Approach*. London: Routledge. https://doi.org/10.4324/9780203768280.

Marais, Kobus, and Reine Meylaerts. 2018. *Complexity Thinking in Translation Studies: Methodological Considerations*. Routledge Advances in Translation and Interpreting Studies. London: Routledge.

Marín García, Álvaro. 2023. “Epistemological Positions.” In *The Routledge Handbook of Translation Theory and Concepts*, 13–27. Routledge Handbooks in Translation and Interpreting Studies. Abingdon: Routledge.

Meylaerts, Reine, and Kobus Marais, eds. 2023. *The Routledge Handbook of Translation Theory and Concepts*. London: Routledge. https://doi.org/10.4324/9781003161448.

Martin, Alison E. 2006. “Annotation and Authority: Georg Forster’s Footnotes to the Nachrichten von Den Pelew-Inseln (1789).” *Translation and Literature* 15 (2): 177–201. https://doi.org/10.3366/tal.2006.0021.

Mauranen, Anna. 2004. “Corpora, Universals and Interference.” In *Translation Universals: Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki. Philadelphia: John Benjamins.

McEnery, Tony, and Richard Xiao. 2008. “Parallel and Comparable Corpora: What Is Happening?” In *Incorporating Corpora: The Linguist and the Translator*, edited by Gunilla M. Anderman and Margaret Rogers, 18–31. Clevedon:

Multilingual Matters.

McRae, Ellen. 2006. “The Role of Translators’ Prefaces to Contemporary Literary Translations into English.” Thesis, The University of Auckland. https://researchspace.auckland.ac.nz/handle/2292/5972.

Mellinger, Christopher D., and Thomas A. Hanson. 2016. *Quantitative Research Methods in Translation and Interpreting Studies*. London: Routledge.

Merrill, Christi A. 2019. “Postcolonialism.” In *Routledge Encyclopedia of Translation Studies*, 3rd ed., 428–32. London: Routledge.

Mullen, Alex. 2012. “Introduction: Multiple Languages, Multiple Identities.” In *Multilingualism in the Graeco-Roman Worlds*, edited by Alex Mullen and Patrick James. Cambridge: Cambridge University Press.

Navigli, Roberto, Simone Conia, and Björn Ross. 2023. “Biases in Large Language Models: Origins, Inventory and Discussion.” *Journal of Data and Information Quality* 15 (2): 1–21. https://doi.org/10.1145/3597307.

Nergaard, Siri. 2013. “The (In)Visible Publisher in Translations: The Publisher’s Multiple Translational Voices.” In *Authorial and Editorial Translation 2: Editorial and Publishing Practices*, edited by Hanne Jansen and Anna Wegener, 177–208. Quebec: Éditions québécoises de l’œuvre. http://hdl.handle.net/10315/26585.

Neumann, Stella. 2014. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German*. Vol. 251. Trends in Linguistics. Studies and Monographs. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110238594.

Newmark, Peter. 1988. *A Textbook of Translation*. Hertfordshire: Prentice Hall.

Niranjana, Tejaswini. 1992. *Siting Translation: History, Post-Structuralism, and the Colonial Context*. Berkeley: University of California Press.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, et al. 2022. “No Language Left Behind: Scaling Human-Centered Machine Translation.” arXiv. https://doi.org/10.48550/arXiv.2207.04672.

Ó Buachalla, Séamas. 1984. “Educational Policy and the Role of the Irish Language from 1831 to 1981.” *European Journal of Education* 19 (1): 75–92.

https://doi.org/10.2307/1503260.

Ó Laoire, Muiris. 2005. “The Language Planning Situation in Ireland.” *Current Issues in Language Planning* 6 (3): 251–314. https://doi.org/10.1080/14664200508668284.

O’Leary, Philip. 1990. “‘The Dead Generations’: Irish History in the Gaelic Revival.” *Proceedings of the Harvard Celtic Colloquium* 10: 88–145.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.

Ostler, Nicholas. 2005. *Empires of the Word: A Language History of the World*. Harper Collins.

Östman, Jan-Ola, and Leila Mattfolk. 2011. “Ideologies of Standardisation: Finland Swedish and Swedish-Language Finland.” In *Standard Languages and Language Standards in a Changing Europe*, edited by Tore Kristiansen and Nikolas Coupland, 75–82. Standard Language Ideology in Contemporary Europe. Oslo: Novus.

Paloposki, Outi. 2010. “The Translator’s Footprints.” In *Translators’ Agency*, edited by Tuija Kinnunen and Kaisa Koskinen, 86–107. Tampere Studies in Language, Translation and Culture, Series B4. Tampere: Tampere University Press.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311. Philadelphia: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pellatt, Valerie, ed. 2013. *Text, Extratext, Metatext and Paratext in Translation*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Pennycook, Alastair. 2022. “Critical Applied Linguistics in the 2020s.” *Critical Inquiry in Language Studies* 19 (1): 1–21. https://doi.org/10.1080/15427587.2022.2030232.

Phillipson, Robert. 2004. “English in Globalization: Three Approaches.” *Journal of Language, Identity & Education* 3 (1): 73–84. https://doi.org/10.1207/s15327701jlie0301_4.

Phillipson, Robert. 2008. "The Linguistic Imperialism of Neoliberal Empire." *Critical Inquiry in Language Studies* 5 (1): 1–43. https://doi.org/10.1080/15427580701696886.

Pięta, Hanna. 2017. "Theoretical, Methodological and Terminological Issues in Researching Indirect Translation: A Critical Annotated Bibliography." *Translation Studies* 10 (2): 198–216. https://doi.org/10.1080/14781700.2017.1285248.

Podlevskikh Carlström, Malin. 2022. "The (in)Visibility of Translation and Translators in the Swedish Publication of Post-Soviet Russian Literature: An Analysis of Peritexts." *STRIDON: Studies in Translation and Interpreting* 2 (2): 45–74. https://doi.org/10.4312/stridon.2.2.45-74.

Pontrandolfo, Gianluca. 2021. "National and EU Judicial Phraseology under the Magnifying Glass: A Corpus-Assisted Analysis of Complex Prepositions in Spanish." *Perspectives* 29 (2): 260–77. https://doi.org/10.1080/0907676X.2020.1815816.

Poplack, Shana, and Nathalie Dion. 2012. "Myths and Facts about Loanword Development." *Language Variation and Change* 24 (3): 279–315. https://doi.org/10.1017/S095439451200018X.

Popović, Maja, and Hermann Ney. 2006. "POS-Based Word Reorderings for Statistical Machine Translation." In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 06), 1278–83. Genoa.

Popović, Maja, and Hermann Ney. 2011. "Towards Automatic Error Analysis of Machine Translation Output." *Computational Linguistics* 37 (4): 657–88. https://doi.org/10.1162/COLI_a_00072.

Pratt, Chris. 1980. *El anglicismo en el español peninsular contemporáneo*. Madrid: Gredos.

Pym, Anthony. 1995. "Schleiermacher and the Problem of 'Blendlinge.'" *Translation and Literature* 4 (1): 5–30.

Pym, Anthony. 2004. *The Moving Text: Localization, Translation, and Distribution*. Amsterdam: John Benjamins. https://doi.org/10.1075/btl.49.

Pym, Anthony. 2007. "Chapter 2: Philosophy and Translation." In *A Companion to*

Translation Studies, edited by Piotr Kuhiwczak and Karin Littau, 24–44. Bristol: Multilingual Matters. https://doi.org/10.21832/9781853599583-004.

Rafael, Vicente L. 1988. *Contracting Colonialism: Translation and Christian Conversion in Tagalog Society under Early Spanish Rule*. Ithaca: Cornell University Press.

Robinson, Nathaniel R., Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. “ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages.” arXiv. https://doi.org/10.48550/arXiv.2309.07423.

Rockenberger, Annika. 2014. “Video Game Framings.” In *Examining Paratextual Theory and Its Applications in Digital Culture*, edited by Desrochers Nadine and Daniel Apollon, 252–86. Hershey: IGI Global.

Rollason, Christopher. 2005. “Unequal Systems: On the Problem of Anglicisms in Contemporary French Usage.” In *In and Out of English: For Better, for Worse*, edited by Gunilla Anderman and Margaret Rogers, 39–56. Clevedon: Multilingual Matters.

Roshdy, Rana. 2023. “Translating Islamic Law: The Postcolonial Quest for Minority Representation.” PhD, Dublin City University. https://doras.dcu.ie/28896/.

Roth, Silke. 2019. “Linguistic Capital and Inequality in Aid Relations.” *Sociological Research Online* 24 (1): 38–54. https://doi.org/10.1177/1360780418803958.

Said, Edward W. 1994. *Culture and Imperialism*. London: Vintage.

Schleiermacher, Friedrich. 1816/2012. “On the Different Methods of Translating.” In *The Translation Studies Reader*, edited by Lawrence Venuti, translated by Susan Bernofsky, 3rd ed., 43–63. Abingdon: Routledge.

Schneider, Britta. 2022. “Multilingualism and AI: The Regimentation of Language in the Age of Digital Capitalism.” *Signs and Society* 10 (3): 362–87. https://doi.org/10.1086/721757.

Seidlhofer, Barbara. 2013. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Sensch, Jürgen. 2007. “History of the German Population since 1815. Data Compilation on the Basis of Published Studies Using Official Statistics and Sources.” Köln: GESIS Data Archive. ZA8171 Data file Version 1.0.0. Accessed March 23, 2024. https://doi.org/10.4232/1.8171.

Seracini, Francesca L. 2021. “Phraseology in Multilingual EU Legislation: A Corpus-Based Study of Translated Multi-Word Terms.” *Perspectives* 29 (2): 245–59. https://doi.org/10.1080/0907676X.2020.1800058.

Serigos, Jacqueline Rae Larsen. 2017. “Applying Corpus and Computational Methods to Loanword Research : New Approaches to Anglicisms in Spanish.” PhD dissertation, The University of Texas at Austin. https://doi.org/10.15781/T26970F0H.

Sidorov, Grigori. 2019. *Syntactic N-Grams in Computational Linguistics*. SpringerBriefs in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14771-6.

SIL International. 2024a. “History of the Ethnologue.” Ethnologue. Accessed March 23, 2024. https://www.ethnologue.com/history.

SIL International. 2024b. “Methodology.” Ethnologue. Accessed April 7, 2022. https://www.ethnologue.com/methodology.

Simpson, Andrew. 2008. *Language and National Identity in Africa*. Oxford: Oxford University Press.

Snell-Hornby, Mary. 2006. *The Turns of Translation Studies: New Paradigms Or Shifting Viewpoints?* Amsterdam: John Benjamins.

Spivak, Gayatri Chakroavorty. 1993. “The Politics of Translation.” In *The Translation Studies Reader*, in Venuti, Lawrence. 2012. *The Translation Studies Reader*. 3rd edition. New York: Routledge.

Srivastava, Neelam. 2018. *Italian Colonialism and Resistances to Empire, 1930-1970*. Cambridge Imperial and Post-Colonial Studies Series. London: Springer.

Stahlberg, Felix, Danielle Saunders, and Bill Byrne. 2018. “An Operation Sequence Model for Explainable Neural Machine Translation.” arXiv. https://doi.org/10.48550/arXiv.1808.09688.

Statistics Finland. n.d. “Vital Statistics by Year and Information, 1749-2022.” Statistics Finland. Accessed March 23, 2024. https://pxdata.stat.fi/PxWeb/pxweb/en/StatFin/StatFin__synt/statfin_synt_pxt_12dx.px/.

Statistics Sweden. n.d. “Population and Population Changes 1749–2023.” Statistikmyndigheten SCB. Accessed March 23, 2024.

https://www.scb.se/en/finding-statistics/statistics-by-subject-area/population/population-composition/population-statistics/pong/tables-and-graphs/population-statistics---summary/population-and-population-changes/.

Statistik Austria. 2022. "Demographisches Jahrbuch 2022." Statistik Austria. Accessed March 23, 2024. https://www.statistik.at/services/tools/services/publikationen/detail/1703.

Stewart, William. 1968. "A Sociolinguistic Typology for Describing National Multilingualism." In *Readings in the Sociology of Language*, edited by Joshua A. Fishman. Paris: Mouton Publishers.

The Swiss Federal Statistical Office. 2021a. "Bilanz der ständigen Wohnbevölkerung, 1861-2020." Bundesamt für Statistik. September 1, 2021. Accessed March 23, 2024. https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/tabellen.assetdetail.18344355.html.

The Swiss Federal Statistical Office. 2021b. "Hauptsprachen Seit 1910." Federal Statistical Office. January 25, 2021. Accessed March 23, 2024. https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/tables.assetdetail.15384561.html.

Tahir-Gürçağlar, Şehnaz. 2002. "What Texts Don't Tell: The Uses of Paratexts in Translation Research." In *Crosscultural Transgressions*, edited by Theo Hermans, 2nd ed., 44–60. Research Models in Translation Studies 2: Historical and Ideological Issues. Manchester: St. Jerome.

Teich, Elke. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. Text, Translation, Computational Processing 5. Berlin and New York: Mouton de Gruyter.

Tesseur, Wine. 2022. *Translation as Social Justice: Translation Policies and Practices in Non-Governmental Organisations*. New Perspectives in Translation and Interpreting Studies 5. New York: Routledge. https://doi.org/10.4324/9781003125822.

Tesseur, Wine, and Angela Crack. 2020. "'These Are All Outside Words': Translating Development Discourse in NGOs' Projects in Kyrgyzstan and Malawi." *Journal for Translation Studies in Africa* 1: 25–42.

https://doi.org/10.38140/jtsa.1.4332.

Thompson, John B. 1991. "Editor's Introduction." In *Language and Symbolic Power*, by Pierre Bourdieu, 1–31. Cambridge: Harvard University Press.

Thompson, Brian, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. "A Shocking Amount of the Web Is Machine Translated: Insights from Multi-Way Parallelism." arXiv. https://doi.org/10.48550/arXiv.2401.05749.

Tirkkonen-Condit, Sonja. 2004. "Unique Items–over-or under-Represented in Translated Language?" In *Translation Universals: Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki. Amsterdam: John Benjamins Publishing.

Toledano Buendía, Carmen. 2013. "Listening to the Voice of the Translator: A Description of Translator's Notes as Paratextual Elements." *Translation & Interpreting* 5 (2): 149–62.

Toral, Antonio. 2019. "Post-Editese: An Exacerbated Translationese." In *Proceedings of Machine Translation Summit XVII: Research Track*, edited by Mikel Forcada, Andy Way, Barry Haddow, and Rico Sennrich, 273–81. Dublin: European Association for Machine Translation. https://aclanthology.org/W19-6627.

Toury, Gideon. 2012. *Descriptive Translation Studies - and Beyond: Revised Edition*. 2nd ed. Amsterdam: John Benjamins.

Toury, Gideon. 2004. "Probabilistic Explanations in Translation Studies: Welcome as They Are, Would They Qualify as Universals?" In *Translation Universals: Do They Exist?*, edited by Anna Mauranen and Pekka Kujamäki, 15–32. Amsterdam: John Benjamins.

Tymoczko, Maria. 1998. "Computerized Corpora and the Future of Translation Studies." *Meta : journal des traducteurs / Meta: Translators' Journal* 43 (4): 652–60. https://doi.org/10.7202/004515ar.

Tymoczko, Maria. 1999/2014. *Translation in a Postcolonial Context: Early Irish Literature in English Translation*. 2nd ed. London: Routledge. https://doi.org/10.4324/9781315538624.

Tymoczko, Maria. 2014. *Enlarging Translation, Empowering Translators*. London:

Routledge. https://doi.org/10.4324/9781315759494.

Tymoczko, Maria. 2018. "The History of Internationalization in Translation Studies and Its Impact on Translation Theory." In *A History of Modern Translation Knowledge: Sources, Concepts, Effects*, edited by Lieven D'hulst and Yves Gambier, 153–69. Amsterdam: John Benjamins.

Macrory, Ian. 2010. "Annual Abstract of Statistics." 146. London: Office for National Statistics. UK. https://webarchive.nationalarchives.gov.uk/ukgwa/20151014002904mp_/http://www.ons.gov.uk/ons/rel/ctu/annual-abstract-of-statistics/no-146--2010-edition/index.html.

van Doorslaer, Luc, and Peter Flynn, eds. 2013. *Eurocentrism in Translation Studies*. Vol. 54. Benjamins Current Topics. Amsterdam: John Benjamins.

van Hout, Roeland, and Pieter Muysken. 1994. "Modeling Lexical Borrowability." *Language Variation and Change* 6 (1): 39–62. https://doi.org/10.1017/S0954394500001575.

van Oost, Astrid, Annelore Willems, and Gert de Sutter. 2016. "Asymmetric Syntactic Patterns in German-Dutch Translation: A Corpus-Based Study of the Interaction between Normalisation and Shining Through." *International Journal of Translation* 28 (1–2): 7–25. http://hdl.handle.net/1854/LU-7222949.

van Poucke, Piet. 2011. "Translation and Linguistic Innovation : The Rise and Fall of Russian Loanwords in Literary Translation into Dutch." In *TransUD-Arbeiten Zur Theorie Und Praxis Des Übersetzens Und Dolmetschens*, edited by Pekka Kujamäki, Leena Kolehmainen, Esa Penttilä, and Hannu Kemppanen, 39:101–20. Berlin: Frank & Timme. http://hdl.handle.net/1854/LU-1168479.

van Poucke, Piet. 2012. "Measuring Foreignization in Literary Translation: An Attempt to Operationalize the Concept of Foreignization." In *Domestication and Foreignization in Translation Studies*, edited by Hannu Kemppanen, Marja Jänis, and Alexandra Belikova, 46:139–57. Berlin: Frank & Timme. http://hdl.handle.net/1854/LU-2017381.

Venuti, Lawrence. 1995. *The Translator's Invisibility: A History of Translation*. New York: Routledge.

Venuti, Lawrence. 2012. *The Translation Studies Reader*. 3rd edition. New York:

Routledge.

Vinay, Jean-Paul, and Jean Darbelnet. 1958/2000. “A Methodology for Translation.” In *The Translation Studies Reader*, edited by Lawrence Venuti, translated by Juan C. Sager and M.-J. Hamel, 84–93. London: Routledge.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. “On the Features of Translationese.” *Digital Scholarship in the Humanities* 30 (1): 98–118. https://doi.org/10.1093/llc/fqt031.

Volkart, Lise, and Pierrette Bouillon. 2023. “Are Post-Editese Features Really Universal?” In *Proceedings of the International Conference HiT-IT 2023*, 294–304. Naples. https://doi.org/10.26615/issn.2683-0078.2023_025.

Way, Andy. 2018. “Quality Expectations of Machine Translation.” In *Translation Quality Assessment: From Principles to Practice*, edited by Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, 159–78. Machine Translation: Technologies and Applications. Cham: Springer. https://doi.org/10.48550/arXiv.1803.08409.

Whyatt, Bogusława, and Nataša Pavlović. 2021. “Translating Languages of Low Diffusion: Current and Future Avenues.” *The Interpreter and Translator Trainer* 15 (2): 141–53. https://doi.org/10.1080/1750399X.2021.1917172.

Winters, Marion. 2004. “F. Scott Fitzgerald’s Die Schönen Und Verdammten: A Corpus-Based Study of Loan Words and Code Switches as Features of Translators’ Style.” *Language Matters* 35 (1): 248–58. https://doi.org/10.1080/10228190408566215.

Wolf, Michaela. 2007. “Bourdieu’s ‘Rules of Game’: An Introspection into Methodological Questions of Translation Sociology.” *Matraga* 20: 130–45.

Wright, Sue. 2006. “French as a Lingua Franca.” *Annual Review of Applied Linguistics* 26: 35–60. https://doi.org/10.1017/S0267190506000031.

Yuste Frías, José. 2012. “Paratextual Elements in Translation: Paratranslating Titles in Children’s Literature.” In *Translation Peripheries: Paratextual Elements in Translation*, edited by Anna Gil-Bardají, Pilar Orero, and Sara Rovira-Esteva, 117–34. Bern: Peter Lang.

Zanettin, Federico. 2012. *Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Routledge.

Zanettin, Federico. 2013. “Corpus Methods for Descriptive Translation Studies.” *Procedia - Social and Behavioral Sciences*, Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC 2013), 95: 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618.

Zenner, Eline. 2013. “Cognitive Contact Linguistics. The Macro, Meso and Micro Influence of English on Dutch.” PhD, KU Leuven. https://lirias.kuleuven.be/1821288.

Appendix A

The project's associated Github directory is accessible via the following link: https://github.com/mattriemland/Riemland-DCU-doctoral-thesis-materials.

The contents of this repository are outlined in the README.txt file.