Promotoren	Prof. dr. Gert De Sutter Vakgroep Vertalen, tolken en communicatie Prof. dr. Johan De Caluwe Vakgroep Taalkunde
Decaan	Prof. dr. Marc Boone
Rector	Prof. dr. Anne De Paepe

UNIVERSITEIT GENT

Faculteit Letteren & Wijsbegeerte

Isabelle Delaere

Do translations walk the line?

Visually exploring translated and non-translated texts in search of norm conformity

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de vertaalwetenschap

2015

Acknowledgements

It is a cliché the size of an average PhD, but I feel incredibly lucky to have been surrounded these past few years by a great number of people who have supported me in one way or another. I would like to seize the opportunity to express my profound gratitude to everyone who, somewhere along the way, pushed me a little further towards the finish line.

Foremost, I would like to thank my supervisors, Gert De Sutter and Johan De Caluwe, as this dissertation owes greatly to them. Thank you, Gert and Johan, for the time you have invested in reading my drafts and for your ideas for improvement. Your complementary perspectives and insightful remarks were of incalculable value to this thesis and I would like to thank you both for your enthusiasm as well as your constructive criticism.

To Bernard De Clerck and Kris Heylen, I would like to express my earnest gratitude for being in my Doctoral Guidance Committee. Your helpful comments and suggestions as well as the time you invested in this project were greatly appreciated.

I would also like to take the opportunity to thank Sandra Halverson, Stella Neumann and Bart Defrancq who were kind enough to accept the invitation to be in my jury.

I am grateful for the useful feedback from anonymous reviewers, conference audiences, etc. which I received over the past four years. Closer to home, I feel extremely fortunate to have been surrounded by many wonderful colleagues in Ghent, too many to mention.

Heartfelt thanks go out to (former) LT3 colleagues, many of whom became friends: Dries and Peter, for all the little things that mattered in the end. Véronique, for believing in me even though I did not have that much to offer on paper. No one can lighten up a room with laughter like you can. Els, Lieve and Kathelijne for your wisdom regarding life, and for the courageous and exemplary superwomen you are. Marjan, Sarah, Cynthia and Joke for being young and keeping us fresh. Klaartje for fruitful breaks at work and listening to my personal ramblings.

Special thanks go out to two truly amazing people: Bart, allow me to refer to Desmet 2014, p. vi when I say: the pleasure was all mine. I know you would advise against too

much praise, so I will stick to this: don't ever change. *Schmorfie*, you possess an endless optimism and enthusiasm which keeps all of us going and it is by no means an understatement when I say that you are the most generous and selfless person I have ever met. As Dries would have it: if the two of you didn't exist already, they would have to invent you. Seriously.

I would also like to thank my colleagues Stefaan, Rita, Annik, Myriam, Filip, Sylvianne, Daan, Nicole, Liselotte, Evy, Pauline and Ruud, who, each in their own way, contributed to the warm and amicable atmosphere in the Dutch department. I specifically would like to thank the people with whom I have shared not only an office, but also cake, lunches and dinner parties: Lore, whom I admire for being so straightforward and for her no-nonsense attitude. I'm pretty sure your awesomeness will lead to achieving great things in life. Only few of us know how to work a tight schedule around a set of *eendenbillen*. Annelore, surely for your analytical skills, but definitely for always being sincere and for your ability to lighten up a room with one of your little remarks. Thank you for looking after me and cheering me up when my mood turned gloomy. Koen, thank you for your never-ending support, especially with regard to the wonderful world of statistics, and for your patience and explanatory skills, all of which have been instrumental to the technical aspects of this thesis. Lynn, for carrying on where I left off by exploring new territories for our research and for your positive energy in the office.

I feel fortunate to be surrounded by a circle of friends, whom I want to thank for always being kind enough to pretend to know what this PhD was actually about, but more importantly, for making life worthwhile: Frederik, my kindred spirit, thank you for the countless pub nights filled with chit-chat (both light and heavy), laughter and the occasional dance moves. Grazas, An-Sofie, for always checking in on me and for our many joint adventures, I am sure there are many yet to come. Pieter, for your unwavering support (if not for this PhD project, than for my virtually non-existing foosball skills). Thank you Flurt and Filimberley for the countless dinner parties, and celebrations of all kinds.

These acknowledgements would be nowhere near complete without thanking my family, by blood, in remembrance or by destiny, who offered me a warm nest, full of opportunities. I would like to thank my in-laws for accepting me as one of their own and for many a Sunday afternoon well spent. Sofie & Pjan, we have never been closer even though you moved to the other side of the planet.

A special thanks to my sister, Ties, for clearing the path for me in so many directions and always looking after me. Etienne, for taking such good care of us. My most heartfelt gratitude goes out to my mom, who went well out of her way to make sure I had lemonade, even though life gave her plenty of lemons.

Of course, I would like to take the opportunity to say thank you to Alex, for singing songs to the cat and reminding me that all is good. Definitely for making me a better person. But most of all, for making every day a happy day. You are, by far, the very best thing that ever happened to me. However, I have to admit that giving me a gorgeous daughter, Ada, was not in your best interest where being my favorite human being on this planet is concerned.

Finally, and perhaps not entirely conform to the norm for this particular genre, I would like to direct the biggest thanks to my supervisor once again. Truth be told, I do not even know where to begin to express my gratitude. Thank you, Gert, for your unwavering support. You were my rock.

List of Abbreviations

GSD	General Standard Dutch
CBTS	Corpus-based Translation Studies
DTS	Descriptive Translation Studies
BSD	Belgian Standard Dutch
TAO	Taaladviesoverleg
DPC	Dutch Parallel Corpus

List of Tables

List of Figures

Table of Contents

Introduction

Does the language used in translations differ fundamentally from non-translated language? This has been one of the main research questions in Translation Studies for the past few decades (see e.g. Baker, 2004; Laviosa, 1998; Malmkjaer, 1997; Olohan & Baker, 2000). More specifically, translation studies scholars have been investigating whether there are differences between translations and non-translations on a lexical (e.g. Bernardini & Ferraresi, 2011), grammatical (e.g S. Hansen & Hansen-Schirra, 2012) and discursive (Hatim & Munday, 2004) level. Many studies focusing on these differences were inspired by Baker's seminal paper of 1993 in which she described the so-called universal features of translation as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (1993, p. 243). On the one hand, Baker's paper had a profound impact on the research field and led to a number of significant developments as well as a boost in research output. On the other hand, however, the highly specific research program which ensued also suffered from a few shortcomings. In the following paragraphs, we will start by discussing the most important developments and continue by exposing some of the shortcomings and how this thesis seeks to remedy these issues.

First of all, Baker suggested that translation studies, as a discipline, was ready for the methodology and the techniques applied in corpus linguistics, which led to the introduction of the corpus-based approach to the field.

Secondly, Baker refers to how the typical, static, notion of equivalence between source and target text should be replaced by a more dynamic concept, viz. a functional equivalence between a translation and its target language and culture (1993, p. 236). More specifically, Descriptive Translation Studies were traditionally mainly source-text-oriented as they investigated the equivalence between a translation and its source text. In other words, translations were compared with their source texts when verifying whether there are linguistic differences between the former and the latter (see e.g. Blum-Kulka, 1986; Catford, 1965; Vinay & Darbelnet, 1995). This type of comparative research was often based on a rather negative kind of reasoning and investigated

translation shifts as "all that a translation *could have had* in common with its source but *does not*"(Toury, 1995, p. 84). By changing the notion of equivalence, Baker suggested moving the field's research focus away from comparing the target text with its source text to comparing translated texts in a given language with non-translated texts in that same language.

Additionally, Baker's encouraging the field of translation studies to focus on identifying the aforementioned translation universals as a necessary means to elucidate "the nature of translated text as a mediated communicative event" (1993, p. 243) resulted in a booming research field. This enthusiastic call resulted in numerous publications reporting on descriptive findings with regard to the existence of translation universals such as explicitation (e.g Mauranen, 2000; Olohan & Baker, 2000; Zufferey & Cartoni, 2014), simplification (e.g. Laviosa, 2002), and normalization (e.g Bernardini & Ferraresi, 2011; Kenny, 2001). Normalization as a concept has been interpreted in various ways and can be summarized succinctly as follows: translations show an overall tendency towards normalized language (see Section 1.4 for a more detailed summary). Moreover, the norm-conformity hypothesis, i.e. the central hypothesis in this thesis, builds on this hypothesized tendency to normalize and predicts that translations will generally opt for a linguistic alternative which is norm-conform rather than an alternative which is not.

It should be noted, however, that the existence of translation universals is not generally accepted and various studies have been published which question the validity of the features' universal status or the way research was conducted (e.g. Becher, 2010; House, 2008; Kruger & van Rooy, 2012). There are a number of issues, all of which are related to the multidimensional aspect of translations, and the following paragraphs set out the details with regard to these issues, how they can be addressed and how Corpus-Based Translation Studies can benefit from tackling them.

First, the use of modern age techniques such as computerized analyses, digital corpora and elaborate empirical methods allow us to investigate translated versus non-translated texts in a completely different manner, revealing a more complex, multidimensional reality with regard to translations. Unlike what many of the early studies appeared to assume, it has been shown that the factor *translation* (versus *non-translation*) is not the only factor at play when it comes to linguistic differences between a translated and a non-translated text and that other factors are likely to cause linguistic differences as well, such as the specific language pair under investigation (e.g. Lefer, 2012), the translator as an individual (Rybicki, 2012) or the cultural system within which the translators perform their task (Lindqvist, 2010). Furthermore, as Baker originally defined the features of translation as not being "the result of interference from specific linguistic systems" (1993, p. 243), this more or less ruled out the potential influence of other factors, which therefore received little attention. Genre and source

language, for example, are two aspects which were originally excluded as possible determining factors and as a result, the effect of these factors has not received much attention, and the combined effect even less so.

The investigation of the first aspect, i.e. genre, is motivated by the fact that a given genre can be distinguished from another genre on the basis of a number of non-linguistic concepts, or situational characteristics (Biber & Conrad, 2009) which often result in specific linguistic characteristics, hence the probability of their importance for the observed differences between texts in a corpus as indicated by, for example, Teich (2003), Puurtinen (2003), Bernardini & Ferraresi (2011), Kruger (2012), and Neumann (2014). Even though research in corpus-based translation studies has established the importance of this extralinguistic factor (see e.g. Becher, 2010; Jenset & McGillivray, 2012; Lefer & Vogeleer, 2013; Puurtinen, 2003; Trosborg, 1997), the number of in-depth studies that are supported by thorough statistical methods is, to the best of our knowledge, rather limited (Diwersy, Evert, & Neumann, 2014; Kruger & van Rooy, 2012; Lefer, 2012).

In addition to genre, the specific influence of the source language has also received little attention, which is surprising as it is not unlikely that linguistic differences between language A and language B might have an influence on their respective translations into language C (see e.g. Cappelle, 2012; Lefer, 2012; Volansky, Ordan, & Wintner, 2013; Zufferey & Cartoni, 2014). Furthermore, a large number of the published studies on translation universals were carried out using English as the language under investigation. However, if one wants to find proof for a feature which is supposedly universal, one should investigate translations in as many languages as possible as "a universal feature is one that is found in translations regardless of language pairs [...]" (Chesterman, 2004b, p. 3).

A second open question concerns the fact that the majority of the research was done on a (near) univariate level, e.g. by comparing the use of only one or a few linguistic variable(s) in translated versus non-translated language (see e.g. Olohan & Baker, 2000 who focus on the absence vs. presence of the reporting that as an operationalization of the explicitation hypothesis). However, if one wants to obtain results and conclusions which can be generalized and which are not dependent on the behavior of a single variable, one should apply multivariate research, i.e. the investigation of multiple variables simultaneously, as this type of approach can provide insights that are more broadly applicable.

A third, more technique-related research gap is linked to the first and the second hiatus mentioned above and concerns the fact that, as translations are the result of a multidimensional activity, they should ideally be investigated by means of advanced statistical techniques which can capture and analyze this multidimensionality. Put differently, highly specific techniques should be applied which allow for analyzing the

multidimensional aspects of the relation between translations and non-translations, both with regard to multiple factors and multiple variables (cf. Gries, 2010).

Research goals

The main goal of this dissertation is to map norm-conforming behavior on a lexical level in translated and non-translated Belgian Dutch text material. More specifically, we want to examine norm-conforming behavior with regard to two types of norms, viz. prescriptive and descriptive norms. The difference between these two types of norms could be summarized as follows: prescriptive norms prescribe how things should be done whereas descriptive norms describe how things are done. This thesis focuses on *prescriptive* norm-conforming behavior by verifying to what extent translations and non-translations conform to the linguistic norms issued by the Dutch Language Union which prescribe whether certain lexemes, expressions, etc. belong to the standard language or not. We investigate *descriptive* norm-conforming behavior on the other hand, by examining to what degree translations conform to genre-determined norms which describe what is prevailing language usage in a given genre and what is not.

Additionally, we want to address the long-standing need for multifactorial research in Corpus-Based Translation Studies by investigating two extra factors besides *translation status,* i.e. whether a text is translated or not. The first additional factor under investigation is the influence of *genre* on the lexical choices in translated and non-translated Belgian Dutch; a language which has received little to no attention in Translation Studies and could therefore reveal interesting results. The second factor under investigation is *source language* and how this factor has an influence on the lexical choices in a number of genres.

In particular, we set out to confirm or reject a number of highly specific hypotheses, each of which has been operationalized in a case study (more details with regard to these hypotheses are provided below). Furthermore, we want to show how multivariate analysis and more particularly profile-based correspondence analysis can be used in corpus-based translation studies, how this highly specific approach is particularly suited for investigating linguistic preferences, and how generating results based on the aggregated behavior of multiple variables can lead to conclusions which are more broadly applicable in comparison to observations based on the behavior of a single variable. Moreover, we want to show how profile-based correspondence analysis and, more specifically, the two-dimensional representation of its results, can provide visual insight into the similarities and differences between the various language varieties in

our dataset. This approach will provide answers to a number of specific research questions such as: “Are Instructive Texts which were translated from French lexically similar to journalistic texts which were translated from French?”, “Or do translated Instructive Texts behave uniformly, irrespective of the source language?”, “Or do Instructive Texts always behave like Instructive Texts, irrespective of source language or translation status?”

Research design

In order to reach the research goals described above, four corpus-based case studies were carried out to investigate norm-conforming behavior. As texts, and particularly translations, are considered “primary products of norm-regulated behavior, and can therefore be taken as immediate representations thereof” (Toury, 1995, p. 65), a corpus-based approach seems highly suitable for investigating whether there are differences in norm-conforming behavior between various genres, e.g. Journalistic Texts, and language varieties, e.g. Belgian Dutch translated from French.

In the first case study we investigate the dispersion of formal versus neutral language based on the underlying assumption that, when given a choice between a formal and a neutral lexical variant for a given concept, the translation will adopt the descriptive, genre-specific norm and behave accordingly. For example, if non-translated journalistic texts show a preference for neutral language, we hypothesize that translated journalistic texts will show a similar preference.

In the second case study we look at the dispersion of more endogenous lexical alternatives versus English loanwords under the hypothesis that translations will conform to the descriptive, genre-specific norms. We hypothesize, for example, that when non-translated Tourist Information shows a preference for English loanwords, translations will follow this trend and translated Tourist Information will show a similar preference for English loanwords.

The remaining two case studies are also based on the norm-conformity hypothesis, which predicts that translations tend to show norm-conforming behavior. However, these two case studies focus on prescriptive norms, rather than descriptive norms, as was the case in the first two case studies. More specifically, we investigate to what extent translations conform to the norms with regard to General Standard Dutch (as opposed to the regionally more restricted variety Belgian Standard Dutch and the non-accepted variety non-standard Dutch), an investigation which was divided over two sub-hypotheses and matching case studies. The main hypothesis for these last two case studies is: translated texts makes more use of standard language than non-translated

texts. More specifically, for the third case study we operationalized this main hypothesis as follows: "when given the choice between a General Standard Dutch alternative and a non-standard Dutch alternative, a translation will make use of the General Standard Dutch alternative more often than a non-translation".

The fourth case study is similar to the third case study as it is based on the sub-hypothesis that translations, when compared to non-translations, will opt more often for the General Standard Dutch alternative instead of the regionally more restricted Belgian Dutch alternative.

The methodology we used was identical for the four case studies. It existed of five phases: (i) hypothesis and variables selection, (ii) data extraction, (iii) manual validation of the data, (iv) statistical processing of the data and (v) interpreting the results.

First, we gather a large, varied collection of linguistic variables in the shape of profiles, i.e. sets of alternatives which have the same function or meaning, such as the near-synonyms *indien* vs *als* (if), *echter* vs *maar* (but), and *reeds* vs *al* (already).

Second, we use a corpus to examine the distribution of these profiles in a number of genres, both in non-translated Belgian Dutch and in Belgian Dutch translated from French and from English. The data were extracted from the DPC (Dutch Parallel Corpus) which is a 10-million-word, sentence-aligned parallel corpus for the language pairs Dutch-English and Dutch-French, with Dutch as the central language and the contents of which are divided into genres (Macken, De Clercq, & Paulussen, 2011).

As a third step, we manually validated the extracted data in order to ensure that only those attestations which contain a variant which can be replaced by the other variants in a given profile were included in the dataset. In other words, if a given variant could not be replaced by the other variant(s) in the profile, the attestation was discarded. For example, one of the profiles in the second case study is *unit* vs. *afdeling.* In Dutch, both terms can be used to refer to a department or unit and can be replaced by one another without changing the meaning of a given sentence. In the following expression, however, *unit* cannot be replaced by *afdeling* as *unit* is used in a specific context and carries a different meaning: "... the patient received two units of blood...". During the manual validation, this attestation of *unit* would be discarded from the dataset. This manual validation step was carried out so as to ensure we are measuring differences between a profile's formal alternatives and not their meanings.

In the penultimate step we want to present an exploratory statistical technique which is ideally suited for this type of research, i.e. profile-based correspondence analysis (Plevoets, 2008). Profile-based correspondence analysis is an extended version of correspondence analysis (Greenacre, 2007) and investigates words or constructions not as autonomous pieces of information, but always in relation to synonymous words or constructions (by means of relative proportions). This exploratory data analysis, based on the profiles' frequencies, results in a two-dimensional plot which offers a

visualization of the linguistic distances between the various factors in our dataset, i.e. the genres and the (source) language varieties. The genres are Legal Texts, Instructive Texts, Journalistic Texts, Political Speeches, Broad Commercial Communication, Specialized Communication and Tourist Information. The (source) language varieties are Belgian Dutch translated from French, Belgian Dutch translated from English and non-translated Belgian Dutch.

Finally, these plots allow us to verify whether our hypotheses with regard to norm-conforming behavior can be confirmed or not. For example, if a plot shows no significant distance between translated Belgian Dutch and non-translated Belgian Dutch, then we cannot say that the norm-conforming behavior is significantly different between these two language varieties. In addition to computing the main plots which show the effects between the genres and the (source) language varieties, we also computed the interactions between the genres and the (source) language varieties. These additional analyses provide information regarding the differences or similarities between, for example, translated Journalistic Texts and non-translated Journalistic Texts and allow us to verify whether there are language varieties which behave differently when compared to the general trends that can be observed in the main plot.

Structure of the thesis

This dissertation is structured as follows:

Chapter 1 provides the reader with a brief survey of the related research, starting with the early research in the field of Translation Studies and working its way to the current hot topics in corpus-based translation studies. Chapter 1 also provides the theoretical background for the foundations of the central norm-conformity hypothesis which is, amongst others, based on Toury's law of growing standardization (1995) and Pym's risk aversion hypothesis (2008).

Chapter 2, Research Design provides the reader with information concerning the corpus and its contents, the statistical techniques that were used and the sources that were consulted to select the variables.

Chapters 3, 4, 5 and 6 present the four case studies that were carried out in this study: (i) a case study which investigates the use of formal versus neutral language; (ii) a case study on more endogenous lexical items versus their equivalent English loanwords; (iii) a case study investigating the dispersion of General Standard Dutch versus non-standard Dutch and, finally; (iv) a case study which investigates the dispersion of General Standard Dutch versus Belgian Standard Dutch.

Finally, in Chapter 7 the overall results are summarized and the general trends and conclusions are discussed. Furthermore, this chapter gives an interpretation of potential explanatory factors for these results. Additionally, this chapter covers the shortcomings of this study, how they could be dealt with in the future and what the perspectives are for future work.

Chapter 1
Background

In this chapter, both the theoretical and empirical background of this study will be presented. More particularly, four aspects will be discussed so as to be able to situate the research presented in this study. First of all, without going into detail, Section 1.1 aims to sketch the highlights with regard to Corpus-Based Translation Studies, showing how the field relates to both Corpus Linguistics and Descriptive Translation Studies. Section 1.2 provides relevant details on one of the core issues in Corpus-Based Translation Studies, viz. the so-called universal features of translations, or translation universals, as well as some theoretical and methodological issues related to this concept. The research presented in this dissertation wants to tackle the issues related to the features' assumed universal character by adding weight to the claim that translating is in fact a multidimensional activity and that both source language and genre have an influence on the linguistic behavior in translations, all of which is described in Section 1.3. The final section in this chapter, Section 1.4, provides background information on the linguistic reality in Flanders on the one hand and the central norm-conformity hypothesis and its related concepts on the other hand.

1.1 Studying translation through corpora

CBTS (Corpus-based Translation Studies) is a rather young research discipline and mainly resulted from two other, related fields of research: Corpus Linguistics on the one hand and Descriptive Translation Studies on the other hand (see e.g. Laviosa, 2002). In the following paragraphs, both fields will be briefly discussed so as to be able to arrive at a better understanding of the research traditions that are typical of CBTS.

1.1.1 Corpus Linguistics

CBTS has been inspired by and benefitted from the methodologies of modern day corpus linguistics, which is linked with the introduction of computers and networks in the 1980s and 1990s and which can be described as "a branch of general linguistics that involves the analysis of large machine-readable corpora of running text, using a variety of software tools designed specifically for textual analysis" (Laviosa, 2002, p. 6).

In Baker (1995) a definition regarding the coverage of the term "corpus" is provided: first, a corpus consists of a collection of texts which can be analyzed automatically or semi-automatically due to their machine-readable form. Second, a corpus is not necessarily limited to written texts, but may contain spoken material as well. Third, a corpus can consist of large numbers of texts which come from various sources and deal with multiple topics (p. 225). Ideally, when investigating linguistic phenomena in such corpora, one would want to look at every single text that was ever produced, i.e. the entire population of texts. However, due to obvious practical reasons this is not feasible and corpus linguists are forced to consult other, more realistic, resources. Other definitions by Francis (1992) and the EAGLES research group (Sinclair, 1996) are similar to Baker's definition and refer to a corpus' desired representativeness and how it should be as representative a sample of the population as possible.

However, it should be noted that, no matter the corpus size or how carefully studied or scientifically rigorous its contents, a corpus can never be a hundred percent representative of "language" as a whole and is always limited to some degree. Moreover, the results, any results, of corpus-based research are always dependent on the corpus at hand and the design criteria that were used to build that specific corpus. In addition to being dependent on the corpus, research results are also dependent on the software tools that are used to observe, analyze and process the corpus (Laviosa, 2002, p. 6). It should therefore be noted that the results which are presented in this thesis are also subject to these restrictions. In Section 2.2, for example, we refer to how the quality of the preprocessing steps of a corpus can affect the query results, irrespective of the quality of the query tools applied.

In order to investigate linguistic phenomena in translation studies, various types of corpora are used and Laviosa proposes a "corpus typology for Translation Studies"(2002, p. 34). It is indicated that, for translation studies, the most elementary corpus exists of only two texts or text excerpts, i.e. a source language text (excerpt) and its corresponding target language text (excerpt). These texts can be enriched with additional information depending on the research goal and whether product-based or process-based research is carried out. Laviosa's typology is based on four hierarchical levels, each of which consists of a number of parameters which describe a range of corpus features. The first hierarchical level contains six sets of parameters, all of which refer to the more general features of a text corpus, viz. text length, timeframe, expert

level, number of languages, corpus language and medium. The three ensuing levels consist of increasingly more specific sets of parameters as can be seen in Figure 1, which shows an overview of Laviosa's suggested typology.

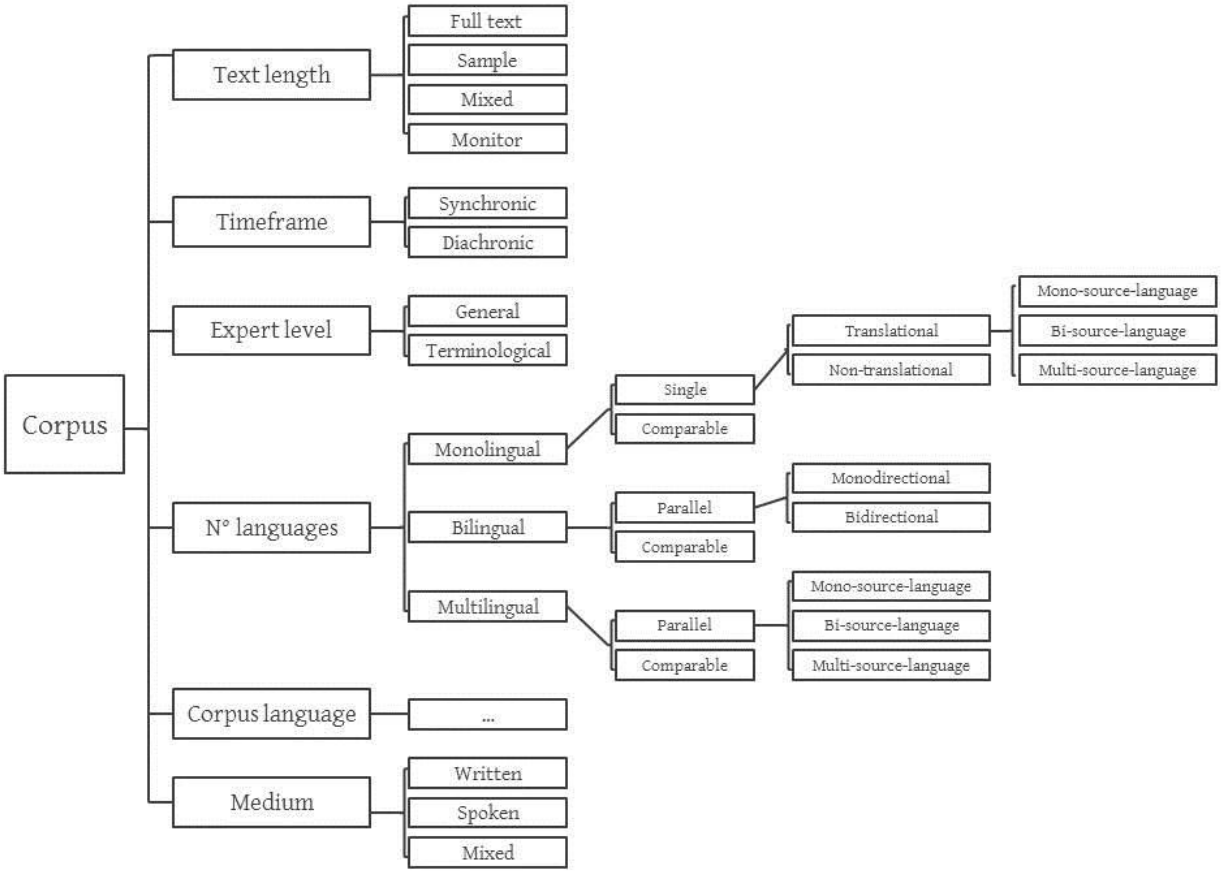

Figure 1 Visualization of the corpus typology as described in Laviosa 2002, p. 34-38

The following paragraphs provide a brief description per corpus type based on Lvaiosa's typology.

TEXT LENGTH

In contrast with full text corpora, which contain complete texts, sample corpora consist of text excerpts which may vary in length depending on the design criteria. A monitor corpus is a dynamic corpus, which means that new material is constantly added causing its size to increase continuously.

TIME FRAME

Synchronic corpora contain material which belongs to a limited period of time and allow researchers to investigate linguistic differences which are not due to language

changing over time, an aspect which can be investigated by applying diachronic corpora, as these contain material which belongs to longer periods of time.

EXPERT LEVEL

General corpora contain text material whose topic allows for everyday language whereas terminological corpora contain texts from specialized domains which leads to the use of technical terminology.

N° of LANGUAGES

Monolingual corpora consist of text material in one single language. When a corpus contains material in two or more languages, it is called a bi- or multilingual corpus.

MONOLINGUAL CORPORA

The difference between a monolingual single and comparable corpus is that the former consists of non-translated text material in one language and that the latter, in addition to the non-translated component, contains a component of translated text material in the same language. In a comparable corpus both components, i.e. the translated and the non-translated one, are built "according to similar design criteria, e.g. according to text genre, topic, time span, distribution of male and female authors, readership, average number of words in each text" (Laviosa, 2002, p. 36), etc.

MONOLINGUAL SINGLE CORPORA

Monolingual single corpora can be either translational or non-translational. A translational corpus contains text material in one language that is certain to have been translated into that language, whereas a non-translational corpus consists of non-translated material only. Furthermore, translational corpora can either contain text material from a single source language (mono-source-language), two different source languages (bi-source-language) or multiple source languages (multi-source-language).

MONOLINGUAL COMPARABLE CORPORA

Monolingual comparable corpora are corpora which consist of two monolingual single corpora: a translational corpus and a non-translational corpus, both in the same language. The two corpora are comparable due to the design criteria which ensure that they are similar in topic, genre, text length, timeframe, etc.

BILINGUAL CORPORA

A bilingual corpus consists of text material in two different languages and can either be parallel or comparable.

BILINGUAL PARALLEL CORPORA

Bilingual parallel corpora can be either mono-directional or bi-directional. A corpus is mono-directional when it consists of text material in language A and its

translation into language B. When, in addition to text material in language A and its translation into language B, the corpus also contains text material in language B and its translation into language A, it is considered a bi-directional corpus.

BILINGUAL COMPARABLE CORPORA

A bilingual comparable corpus contains non-translated text material in two languages which is similar with regard to aspects such as "text genre, topic, time span and communicative function" (Laviosa, 2002, p. 36).

MULTILINGUAL CORPORA

Similar to bilingual parallel corpora, a multilingual parallel corpus contains text material in various languages, e.g. A, B and C, and its translations in multiple languages, e.g. D, E and F whereas a multilingual comparable corpus contains non-translated text material in various languages.

MULTILINGUAL PARALLEL CORPORA

Multilingual parallel corpora are similar to bilingual parallel corpora and can either contain source texts from a single source language (mono-source-language), two different source languages (bi-source-language) or multiple source languages (multi-source-language).

The corpus which was applied in this study is a multilingual parallel corpus (see Section 2.2 for more details).

MULTILINGUAL COMPARABLE CORPORA

Equally similar to bilingual comparable corpora, the text material in a multilingual comparable corpus is collected using specific design criteria which ensure that the texts are indeed comparable where various aspects such as time frame, genre, topic etc. are concerned.

CORPUS LANGUAGE

In theory, a corpus can be built for any language.

MEDIUM

A corpus can consist of material that is (i) written text material; (ii) spoken text material; (iii) a combination of both.

1.1.2 Descriptive Translation Studies

The research presented in this thesis belongs to the field of DTS (Descriptive Translation Studies), which are considered "the branch that concerns itself with the systematic description of three distinct empirical phenomena seen as constituting the object of the

discipline as a whole: the product, the process, and the function of translation" (Laviosa, 2002, p. 10). The elements in this description, viz. translation as a product, translation as a process and the function of translation, can be linked to various sub-disciplines in descriptive translation studies, which will be described briefly in the following paragraphs.

When the research focus lies on translation as a process, and thus ultimately on "trying to find out what happens in the mind of a translator" (Munday, 2012, p. 17), the main goal is to identify every step that is taken between the source text and the translation, or, put differently, the decisions that are being made during the course of a translation. In the early days, investigating the translation process was considered problematic due to what Holmes described as "the little black box" which is the translator's mind (2000, p. 177) and how it is hard to know exactly what is going on in there. More recent research attempts to tackle this problem by using, for example, think-aloud protocols, key-stroke logging, or eye-tracking equipment. Using think-aloud –protocols was often done in the 1990s and basically consists of asking the translator to "verbalize his/her thought processes while translating or immediately afterwards" (Munday, 2012, p. 100). These observations are then recorded, transcribed and used for analysis by the researcher. Keystroke logging software such as Translog or Inputlog records all keystrokes including deletions, corrections, etc. and is therefore a valuable tool for monitoring various types of written language production, including translation processes (see e.g. G. Hansen, 2006; Jakobson, 2005). Applying eye-tracking methods allows a researcher to investigate the translator's focus while (s)he is translating by recording all eye movements, a technique which is believed to reveal information with regard to whether the translation task is cognitively demanding or not (see e.g. O'Brien, 2010).

Investigating the function of translation, has to do with how the translation product and translating as an activity are positioned in the target culture (Laviosa, 2002). In other words, this aspect focuses more on contexts rather than on texts (Munday, 2012). It is therefore not surprising that Holmes originally used the term 'socio-translation studies' to refer to this research area.

Finally, product-oriented descriptive translation studies such as the corpus-based research presented in this thesis focus on translations as finalized products, which can be investigated either in themselves as individual translations or in comparison to other translations. The latter, a comparative analysis, is usually carried out based on multiple translations of the same source text while these translations can exist in one target language or in various target languages (Holmes, 2000). This type of comparison often focuses on translation shifts (Catford, 1965), which are changes between the source text and the target text(s), and which occur somewhere along the translation process. Baker & Saldanha (2008) used a definition by Popovič to identify various elements which are

important where product-oriented translation shifts are concerned: "all that appears as new with respect to the original, or fails to appear where it might have been expected, may be interpreted as a shift" (1970, p. 79).

In *Descriptive Translation Studies and Beyond*, Toury (1995) calls for a common and well-studied methodology as well as systematic research techniques to investigate these translation shifts in order to "ensure that the findings of individual studies will be intersubjectively testable and comparable, and the studies themselves replicable" (p. 3). More specifically, Toury suggests a research program for descriptive translation studies including a three-stage analysis which can be applied to corpus-based (descriptive) translation studies.

Within the first stage, it should be assessed whether a given translation is accepted as a general target language text on the one hand and as a translation in the target culture on the other hand, an assessment which can be carried out by comparing multiple translations in one specific language from a given source text (1995, p. 72-73).

The second stage of Toury's suggested approach consists of an analysis of the relationship between the target text and its source text, where the target text is mapped onto its source text's counterparts. An analysis of this type can by carried out in order to detect so-called translation shifts between the source text and the target text.

Finally, the aim of the third phase of the suggested approach, is to make generalizations about the detected patterns in both the source and the target text. These generalizations could help to determine the specific translation process for a given source text – target text pair by revealing exactly how the translator balanced between invariance and transformation to in search of equivalence.

Munday (2012) refers to Toury's suggested approach and adds an additional, fourth, step to it, i.e. the possibility of repeating phases one to three for other pairs of texts. More specifically, Munday states that replicability and the resulting comparable research results can lead to a "descriptive profile of translations", which can be constructed based on genre, period, author etc. (p. 170). Additionally, this kind of approach will allow for the norms which pertain to each type of translation to be identified and "as more descriptive studies are performed, the ultimate aim is to state laws of behavior for translation in general" (p. 170).

1.1.3 Corpus-based Translation Studies

From the 1990s onwards it was ever more suggested to start implementing the corpus-based approach in descriptive translation studies. In her seminal paper of 1993, for example, Mona Baker states that translation studies as a discipline is about to take a turning point and that

> "this turning point will come as a direct consequence of access to large corpora of both original and translated texts, and of the development of specific methods and tools for interrogating such corpora in ways which are appropriate to the needs of translation scholars. Large corpora will provide theorists of translation with a unique opportunity to observe the object of their study and to explore what it is that makes it different from other objects of study, such as language in general or indeed any other kind of cultural interaction." (Baker, 1993, p. 235).

Baker strongly encouraged applying the methods as well as the resources from corpus linguistics as this would enable translation scholars to both identify those features which are specific of translated language and to bring us to a better understanding of translation and how the processes involved (1993). Toury elaborates on possible guidelines on what exactly such a corpus in translation studies should look like and states that translation studies scholars should investigate translation in reality and not translation in general (1995, p. 32). As a consequence, we should study assumed translations, i.e. "utterances which are presented or regarded as such within the target culture, on no matter what grounds" (1995, p. 32).

Baker was right in predicting that translation studies would take a turn and the number of scholars who applied or at least considered the corpus-based approach grew rapidly (see e.g. Baker, 1995; Kenny, 1998; Laviosa, 1998; Olohan & Baker, 2000; Øverås, 1998; Zanettin, 2000b). It became clear that extended corpus-based studies (i) are an added value for unraveling linguistic phenomena that might not be detected when investigating language by means of single texts and (ii) allows for conclusions that are more broadly applicable in comparison to conclusions which were drawn from the comparison of one translation with its source text.

However, it should be noted that for most corpus-based studies "there is no way of knowing how many different persons were actually involved in the establishment of a translation, playing how many different roles" and although it is common to regard the conjoined entity of all of them as 'the translator', a more process-based approach is needed to investigate the actual role of the translator (Toury, 1995, p. 183).

Depending on the research goals, translation scholars can use different types of corpora, an overview of which is described in Section 1.1.1. Lefer & Vogeleer (2013) refer to Bernardini (2010) who used her experience in both translation studies and contrastive linguistics to plead for a combined use of both monolingual comparable corpora and parallel corpora in translation studies. According to Bernardini, translation scholars can use monolingual comparable corpora to find evidence of features which are typical of translated texts (in comparison to non-translated texts). Parallel corpora, on the other hand, are very well suited for verifying whether the observed features of translated texts are indeed due to a translation process, and were not merely caused by the source text at hand (Bernardini, 2010). According to the typology presented in 1.1.1, the corpus

which was needed for the research presented in this thesis, is a multi-source-language multilingual parallel corpus. There is a small number of officially available corpora for Dutch, and an even smaller number for corpora which consist of Dutch which was produced in Belgium. Furthermore, a distinction should be made between corpora which contain written material and corpora which contain spoken material and/or transcriptions for this spoken material. Some of the publicly available speech corpora for Dutch are: the *Corpus Gesproken Nederlands* (Oostdijk, 2000), the Europarl corpus (Koehn, 2005) and a section of the Eindhoven corpus (Uit den Boogaart, 1975). With regard to contemporary written material, available corpora are the written section of the Eindhoven corpus (Uit den Boogaart, 1975), the Parole Corpus (2004), certain sections of the 38 Million Word Corpus (Kruyt & Dutilh, 1997), and the Twente Nieuws Corpus (2003). However, due to copyright issues, the Twente Nieuws Corpus can no longer be distributed. The number of parallel corpora for Dutch which are publicly available is rather limited: OPUS, a parallel corpus which is publicly available[1], the Triptic Corpus (1999), the MLCC Corpus[2], and the Dutch Parallel Corpus (Macken et al., 2011).

1.2 The concept of Translation Universals

Corpus-based Translation Studies as a research domain was profoundly influenced by Mona Baker's seminal paper of 1993 in which she refers to Even-Zohar's polysystem theory and Toury's notion of norms as key factors in the changing field of translation studies. Even-Zohar's polysystem theory (Even-Zohar, 1978) considers literature to be a dynamic collection of systems rather than a static collection of texts. As a consequence, just as children's literature is related to literature for adults, "translated literature is not disconnected from 'original' literature" (1979, p. 292). Even-Zohar expands on this idea and states that translations as such have a specific way in using the literary repertoire which results in a particular way of adopting specific norms, behaviors and policies. As a consequence, "translated literature may possess modelling principles of its own, which to a certain extent could even be exclusive to it" (Even-Zohar, 1978, p. 118).

In addition to the polysystem theory and its implications for translated literature in general, Baker refers to Toury's norm concept as one of the more important notions

[1] http://opus.lingfil.uu.se/

[2] http://catalog.elra.info/product_info.php?products_id=46

which consider the target system and culture as a starting point instead of focusing on the source-oriented notion of equivalence (1993, p. 239). Norms are those options which are frequently selected by translators at a given moment and in a certain socio-cultural context and which are the result of a translation tradition. Furthermore, Baker states that this translation tradition should be observed and analyzed by means of "a representative body of translated texts in a given language or culture" (1993, p. 240), in other words, by means of a corpus. Baker further refers to Toury who states that communicating in translated utterances "imposes patterns of its own" (1991, p. 50) and Even-Zohar who claims that, in translation, we can observe "patterns which are inexplicable in terms of any of the repertoires involved" (1979, p. 77).

Baker interprets these patterns as "patterns which are not the result of interference from the source or target language" (1993, p. 242) and calls them *universal features of translation*, or "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (1993, p. 243)

Baker initially introduced the following translation universal candidates:

- **Explicitation**, or "a marked rise in the level of explicitness compared to specific source texts and to original texts in general" (p. 243).
- **Simplification**, or "a tendency towards disambiguation and simplification" (p. 244).
- **Conventionality**, or "a strong preference for conventional grammaticality" (p. 244).
- A tendency to "**avoid repetitions**, either by omitting them or rewording them" (p. 244).
- A general tendency to "**exaggerate features of the target language**" (p. 244).
- **A specific type of distribution of certain features** or 'the third code', "which is a result of the confrontation of the source and target codes and which distinguishes a translation from both source texts and original target texts at the same time" (p. 245).

As mentioned in the Introduction, the research program as proposed by Baker resulted in numerous studies and publications, not only on the translation universal candidates mentioned above, but also on new possible candidates and caused a major boost in the field.

In the following paragraphs, both the positive and the negative aspects related to translation universals will be discussed.

1.2.1 Baker's research program

In addition to encouraging the introduction of a corpus-based methodology in translation studies, Baker's paper of 1993 also suggested a specific task to translation scholars. Baker mentioned the fact that translated texts as such are different in comparison to other communicative events in any language and that the features of this

difference should be "explored and recorded" (1993, p. 234). More specifically, Baker described the task to be undertaken as follows:

> "The most important task that awaits the application of corpus techniques in translation studies, it seems to me, is the elucidation of the nature of translated language as mediated communicative event. In order to do this, it will be necessary to develop tools that will enable us to identify universal features of translation, that is features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems."(Baker, 1993, p. 243)

Baker's suggested research program had a profound influence on the field. First of all, the introduction of Corpus-Based Translation Studies as a research discipline marked the beginning of a new era and resulted in a substantial number of studies, leading to promising results. An additional potential advantage of applying a common, corpus-based methodology is the possibility to replicate a given study and to compare the results of various, similar studies.

A second aspect which more or less coincides with this new turn in translation studies is the focus shift from comparing translations to their source text to comparing translated language as such with original, non-translated language. As a result, translations were no longer considered second-hand versions of their originals, but a language variety with its position in the target language system.

A third aspect related to Baker's call for applying new methodologies for detecting translation universals was the sudden rise in research output focused on finding proof of the existence of these universals, which led to thorough analyses of both translated and non-translated texts, which in turn led to numerous interesting descriptive findings. Moreover, these descriptive findings led to formulating (possibly) explanatory hypotheses with regard to the observed results, which could help to arrive at a better understanding of the translation process by means of additional, more experimental research.

In addition to a substantially growing research output focusing on characteristics that are typical of translated texts, (Chesterman, 2004a; Klaudy, 2001; Olohan & Baker, 2000; Øverås, 1998; Tirkkonen-Condit, 2004), this new turn in translation studies resulted in the creation of a rapidly growing number of resources, i.e. corpora, such as for example GEPCOLT, i.e. the German-English Corpus of Literary Texts (Kenny, 2001), TEC, i.e. the Translational English Corpus[3], CEXI, i.e. the English-Italian Translational Corpus (Zanettin, 2000a) , ESPC, i.e. the English-Swedish Parallel Corpus (Altenberg & Aijmer, 2000), CroCo, i.e. Cross-Linguistic Corpora (Hansen-Schirra, Neumann, & Steiner,

[3] http://www.llc.manchester.ac.uk/ctis/research/english-corpus/

2012) or DPC, i.e. Dutch Parallel Corpus (Macken et al., 2011). In turn, building these corpora opened up new opportunities for corpus-based research on translation universals.

1.2.2 Areas for improvement

Aside from the advantages and far-reaching, positive consequences of this research program, the concept of translation universals and the studies which were carried out to find evidence for their existence were subject to criticism as well, and more extensive research revealed a number of research gaps and shortcomings with regard to the corpus-based studies into translation universals. The following paragraphs will provide some background information with regard to the various types of criticism which centered on the concept of translation universals and the concept-related studies which were carried out.

First of all, we will discuss the implications of how the concept 'universal feature of translation' was defined, and some additional issues with regard to this definition. Secondly, we will provide some alternative approaches and thirdly, we will discuss some methodological shortcomings which should be addressed in future research.

A great deal of criticism which was expressed against the concept of translation universals and the research program that surrounded this concept was a direct result of the assumed 'universal' character of these features of translation (see e.g. Becher, 2010; Bernardini & Zanettin, 2004; House, 2008; Kruger & van Rooy, 2012; Mauranen, 2008; Mauranen & Kujamäki, 2004; Tymoczko, 2005). Becher, for example, expresses the need for a clear definition of what the term 'universal tendency' refers to exactly and questions the criteria a given tendency should meet to qualify as 'universal' (2011). The implications of the applied terminology were usefully summarized by Chesterman as follows:

> "In simple terms, we can define a translation universal as a feature that is found (or at least claimed) to characterize all translations: i.e. a feature that distinguishes them from texts that are not translations. More strictly: to qualify as a universal, a feature must remain constant when other parameters vary. In other words, a universal feature is one that is found in translations regardless of language pairs, different text-types, different kinds of translators, different historical periods, and so on."(Chesterman, 2004b, p. 3)

In other words, the fact that features of translation are supposedly universal implies that they should be detected in any study and by means of any corpus. However, plenty of studies have been carried out the results of which were inconclusive or even negative with regard to the occurrence of a given translation universal (see e.g. Kenny, 1999;

Kruger & van Rooy, 2012; Laviosa, 1998; Mauranen, 2000; Øverås, 1998; Puurtinen, 2003, 2004; Williams, 2005).

In addition to the problematic implications of the chosen terminology which received criticism from various angles (see also Becher, 2010; House, 2008), the universal features of translation such as, for example, explicitation are often not clearly defined or lack a precise description of their exact contents or how they should be operationalized (e.g. Chesterman, 2010; Øverås, 1998).

Furthermore, we would like to point out a third issue which concerns the apparent contradiction between various suggested universals. Pym, for example, illustrates this by referring to the universal of simplification which should result in shorter sentences in translations. The universal of explicitation, however, should in longer sentences in translations. It seems hard for both features to be indeed universal and occur simultaneously (2008).

Although House and Tymoczko have taken an extreme position by pointing out that the search for (proof of) translation universals is "futile" and that "the field should give up the search for universals" (House, 2008, p. 11; Tymoczko, 2005, p. 1095, respectively), we would like to choose a middle ground and feel that there is room for an alternative approach such as the one suggested in Laviosa (2002):

> "What universals-based studies intend to unveil is not the existence of all-or-none phenomena, but tendencies, trends, regularities which do not occur in an aseptic, dull environment devoid of singular behaviours, but emerge from a rich, intricate, dynamic world of diversity and contrasts." (p. 78)

Mauranen suggests a similar approach by stating that the translation universals do not "necessarily refer only to absolute laws, which are true without exception" but that most of them are "general or law-like tendencies, or high probabilities of occurrence" (2008, p. 35). Becher takes it a step further and calls for an entirely different approach which no longer focuses on trying to "prove or disprove the allegedly universal status" of a given phenomenon, but which aims at identifying the factors which cause this phenomenon to appear or not (2011, p. 75).

In order to use such an alternative approach, there are some methodological issues which must be solved. For example, law-like tendencies should be investigated in all translational data, irrespective of the specific language under investigation and research should therefore be carried out on as many languages as possible in order to verify whether such tendencies can indeed be found. Not only is there a need for including a wide range of languages, there is a related aspect which may have been overlooked, i.e. the specific language pair under investigation. However, this aspect is highly interesting as the relation between two given languages may be completely different from the relation between two other given languages. Mauranen & Kujamäki

(2004), for example, mention how we should "not draw excessively hasty conclusions on the basis of comparing typologically very close languages only, or a very small range of languages" (see also House, 2008; Lefer, 2012). More particularly, it seems plausible that translating from one Germanic language, say English, into another, say Dutch, results in a translation process which is different from that of translating from a Romance language, say French into a Germanic language, say English.

There is an additional research gap which concerns the fact that previous corpus-based studies on translation universals were often carried out by means of a monofactorial approach. In other words, they focused only on comparing translated text material with non-translated text material without taking other factors such as the source language or the genre into account. However, such a monofactorial approach should be avoided as it fails to investigate the influence of extralinguistic factors on the language usage in translations such as the translator's background, the editorial process of a translation, the genre a translation pertains to, etc. As these factors are highly likely to have an influence on the linguistic characteristics of a given translation, they should be accounted for when presenting the results.

Furthermore, the majority of the corpus-based research that did focus on such extralinguistic factors only focused on one factor at a time while it seems highly interesting to investigate various factors simultaneously and, more particularly, how they interact with one another.

Finally, in addition to the monofactorial research tradition mentioned above, corpus-based translation studies were often carried out by means of a univariate approach, meaning that the results were based on the behavior of one variable only. However, it seems plausible that research which is based on a broader set of variables may lead to conclusions which are more broadly applicable than those which resulted from a univariate approach.

1.3 Multidimensionality in Corpus-Based Translation Studies

As described in the previous paragraphs, it seems that corpus-based translation studies can greatly benefit from a broader research focus and the application of multivariate statistics which can take various (extralinguistic) factors into account simultaneously.

One of these factors is the individual translator, which was investigated by Rybicky, for example, who applied machine-learning stylometric distance methods to translations in an attempt to identify a text's translator. His results showed an overall tendency of the various translations in his corpus to cluster by original author and volume rather than by translation, thus indicating that the individual translator could

not easily by identified. Not only the individual translator, but also the specific cultural system a text is being translated into has been investigated (2012).

In Lindqvist (2010) the following hypothesis was put forward: depending on the cultural system (and their overall translation policies) within which translators perform their task, the meta-textual elements will be translated differently. She compared the translations of a Spanish novel into French, English and Swedish and concludes that this type of study might lead to a better understanding of translational behavior but that additional research is needed as the findings neither confirm nor contradict the hypothesis.

Another relevant extra-linguistic factor is the editorial process the average published text goes through. Kruger, for example, investigated to what degree editorial intervention has an effect on a number of phenomena which are typically perceived as features of translation, viz. explicitation, conventionalization and simplification (to appear). Her findings showed that the main effect of editing is a normalization effect, while little evidence was found of simplification and explicitation as a consequence of editorial intervention.

The factors under investigation in this thesis are source language and genre which are also thought to have a substantial influence on linguistic preferences, irrespective of whether a text is translated or not, and which will be described in the following paragraphs.

1.3.1 Source language

Baker's initial definition of universal features of translation basically excluded source language as a potential factor by saying that these features are not "the result of interference from specific linguistic systems" (Baker, 1993, p. 243). Interference was described in high detail by Toury, who suggested the Law of Interference as a potential law of translation (Toury, 1995). More specifically, Toury formulated the law as follows: "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (p. 275). When interference from the source language in translation is measured in terms of proportionalities and frequencies, rather than in the occasional occurrence of simple interference, this is referred to as "shining through", an assumed property of translation which causes the translation to be oriented towards the source language (see also Hansen-Schirra et al., 2012). Furthermore, source language shining through (Teich, 2003) supposedly contributes to the fact that a translation is different from a comparable, non-translated text in the same language. The concept of interference has often been interpreted in a negative sense, where translations are seen as victims of a strong interference from either the source language or the source text, a perspective which may have had a negative effect on the amount of research which was

carried out to investigate non-negative effects of the source language on the target texts (see also Mauranen, 2008).

However, the influence of the factor source language, i.e. interference, can be investigated in a neutral, abstract and statistical sense and the main goal of this approach is to find out whether significant differences can be detected between language C translated from language A on the one hand and language C translated from language B on the other hand (see also Eskola, 2004). Mauranen even mentions that comparable corpora should include various source languages, so as to be able to distinguish between what is typical of translations in general and what is typical of a specific language pair (2008). In fact, it has been shown that the specific source language is highly likely to have an influence on the linguistic characteristics of a given translation (see e.g. Cappelle, 2012; Lefer, 2012; Mauranen, 2004; Volansky et al., 2013). In Cappelle 2012, for example, the author investigates whether there are differences between English translated from French and English translated from German. The research goal was to verify whether there are frequency differences between the language varieties with respect to verbs expressing manner of motion. More specifically, it was hypothesized that English translated from French contains fewer manner-of-motion verbs than non-translated English whereas no such difference was expected between English translated from German and non-translated English. This hypothesis was based on the fact that both English and German are classified as satellite-framed verbs, whereas French is a verb-framed language. The results showed that the frequencies did indeed vary according to the source language under investigation and Cappelle therefore concluded that "differences between the grammatical systems of languages apparently have their role to play in the translation of motion expressions".

1.3.2 Genre

A second factor which has shown to have a substantial influence on the linguistic preferences of a given text, is the genre the text pertains to. Steiner (2001) and Teich (2003), for example, refer to genre as one of the influencing sources of the properties of translated text while Neumann mentions a translator scholars' demand for research on genre influence on the source as well as the target text, a claim confirmed by her own study (2013). Genres are determined by a variety of non-linguistic, situational characteristics (Biber & Conrad, 2009) and these characteristics result in specific linguistic particularities. Baker (1999), for example wondered whether "certain linguistics or strategies are more likely to occur in certain types of translation genres, like translated fiction, news, inflight magazines" (p. 292). In her concluding remarks, Olohan (2004) too points out that a substantial amount of translations involve non-

literary texts and that, therefore, other genres and texts should also be examined. As a result of investigating other genres, she adds, "we will be able to make more cross-genre comparisons and study the extent to which features of translation may be influenced by genre, text type, etc." (p. 191).

Although the importance of genre as an influential factor has been pointed out (see also e.g. Becher, 2010; House, 2008; Jenset & McGillivray, 2012; Lefer & Vogeleer, 2013; Mauranen, 2008; Puurtinen, 2003; Trosborg, 1997), the number of in-depth studies that are supported by thorough statistical methods is, to the best of our knowledge, still rather limited (for some examples, see Diwersy et al., 2014; Kruger & van Rooy, 2012; Lefer, 2012; Neumann, 2014).

1.4 Norm Conformity

In Section 1.2 the concept of universal features of translation was introduced and, among others, the following feature candidates were put forward: "a strong preference for conventional grammaticality" and "a general tendency to exaggerate features of the target language" (Baker, 1993). In 1995, Toury launched his Law of Growing Standardization (Section 1.4.1.1) which inspired scholars to look for proof of a number of tendencies which are similar to Baker's feature candidates and which are also assumed to be typical of translation such as normalization, conventionality and conservatism. In 2008, Pym proposed the risk aversion hypothesis as a potential explanation for this assumed standardizing behavior of translators and, similarly, Mauranen linked the observed conventionality in translations with a general caution which is typical of translations (Section 1.4.1.2). In the following paragraphs, we will discuss how an assumed norm-conforming behavior of translations synthesizes Toury's Law, Pym's and Mauranen's concept of the cautious translator and a number of features of translation. Furthermore, Section 1.4.2 serves as an introduction to the language situation in Belgium, which is particularly suitable for this type of study on norm-conforming behavior. Finally, all of the aspects mentioned above led us to formulate the norm-conformity hypothesis which is described in Section 1.4.3 which, moreover, introduces two types of norms: descriptive and prescriptive norms.

1.4.1 Theoretical concepts

1.4.1.1 Toury's Law of Growing Standardization

The law of growing standardization was formulated by Gideon Toury (1995, p. 268) as follows: "in translation, source-text textemes tend to be converted into target-language (or target-culture) repertoremes". He elaborates on this concept by adding that "in translation, items tend to be selected on a level which is lower than the one where textual relations have been established in the source text" (p. 269). According to Øverås this comes down to the idea that "translators often fail to capture the complex web of these [source text] relationships and instead produce ready-made, cliché structures, i.e. repertoremes" (1998, p. 582). Later investigations have made use of expressions which refer to related tendencies such as normalization (or conservatism) and conventionality, which will be discussed in the following paragraphs.

Normalization or conservatism has been described as "a term generally used to refer to the translator's sometimes conscious, sometimes unconscious rendering of idiosyncratic text features in such a way as to make them conform to the typical textual characteristics of the target language" (Scott, 1998, p. 112). Kenny (1998) tried to find evidence for normalization at the lexical level and hypothesized that when we compare translations to their source texts, the target texts consist of more toned down vocabulary which causes the target texts to be 'sanitized versions of the originals'. In particular, she described normalization as "the hypothesized tendency of translators to produce translations that are linguistically conservative vis-à-vis their source texts or texts originally produced in the target language" (2000, p. 94). Furthermore, she refers to the observations made by Vanderauwera, who says that foreign texts are usually allowed to be exotic, but only in their packaging (e.g. the book jacket) and not in their linguistic features (1985). Whereas Kenny (1998) focused on normalization on the lexical level, Baker (1993) refers to the tendency of translators to prefer conventional 'grammaticality' and subsequently defined this type of normalization (or conservatism) as the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them (Baker, 1996, p. 183). Baker mentioned findings by Shlesinger, Ben-Shahar, Vanderauwera, Malmkjaer and May to support this tendency in translations. Shlesinger (1991) reported that interpreters tend to round off unfinished sentences by the original speakers, grammaticize ungrammatical sentences and remove hesitations and false starts. According to Ben-Shahar (1994), normalization can also occur on other levels than the grammatical one and marked expressions, not just ungrammatical ones, are also normalized in translations. Vanderauwera (1985), Malmkjaer (1997) and May (1997) found that translators often normalize punctuation, even if non-standard punctuation is used purposefully by the original writer (Baker, 1996). More recently, Bernardini & Ferraresi (2011) investigated the use of Anglicisms in

translated versus non-translated Italian. They used a parallel, comparable corpus and conducted a thorough quantitative and qualitative analysis of the dispersion of 'overt lexical borrowings', 'adapted borrowings', 'semantic loans' and 'morphosyntactic calques' across the different sub-corpora. Their results support the normalization concept in showing that the choices made by translators are more conservative than those of authors of original language.

Summarizing these observations, normalization can be described as "a process within which a (translated) text approximates or even exaggerates some norm of the target register it is translated into, always in terms of some selected textual/linguistic feature" (Hansen-Schirra et al., 2012, p. 4).

The idea of normalization bears close resemblance to the concept of **conventionality**, a concept mentioned by Chesterman who reformulated Toury's law as "translators tend to replace text-specific items with institutionalized items: translations tend to be less idiosyncratic, more conventionalized, than their originals" (1997, p. 72). This idea of conventionality was mentioned by Vanderauwera (1985), Øverås (1998), Kenny (2000), Stewart (2000), and Mauranen (2008). Mauranen, for example, refers to the fact that other scholars have noticed that translations use "generally unmarked grammar, clichés, and typical, common lexis instead of the unusual or the unique" and that in translations, dialect is often replaced by standard language, punctuation is often normalized and target-language specific features are often exaggerated. Stewart's (2000) viewpoint on conventionality can be linked to Chesterman's use of the term "institutionalized items"(1997) as he states that conventionality essentially amounts to the idea that much language use is routine and he links the conventionality hypothesis to the fact that translators are generally thought to produce somewhat less creative language than original writers. Stewart elaborates on the conventionality hypothesis in suggesting that translators who translate into a foreign language deliver even more conventional texts than translators who translate into their mother tongue. Moreover, he mentions how corpora may show a "normalizing, standardizing tendency in translated texts" and that, if translators use corpora "to ensure the validity or existence of collocations, turns of phrase etc.", these corpora can actually contribute to a language which becomes "increasingly flat and conventional" by diminishing the creativity aspect in translations (p. 86).

Combining the research mentioned above, conventionality could be defined as a tendency of translations to stick to conventional, unmarked, common language instead of unusual or unique language.

Although various terminological references, viz. normalization (or conservatism), and conventionality have been introduced, their contents appear to be similar to a certain degree and maybe even overlapping. Overall, it appears that these tendencies may be

attributable to an "umbrella" tendency, i.e. an assumed tendency of translations to, wherever possible, conform to the norm.

1.4.1.2 The concept of the cautious translator

Independently from one another, both Pym and Mauranen put forward a potential explanatory hypothesis with regard to the assumed normalizing behavior of translations. More specifically, in Pym (2005) a suggestion is made to model the concept of explicitation within a framework of risk management, where risk is interpreted as "the probability of an undesired outcome" (p. 34). Pym describes translators as mediators who have a certain communicative function and want to reduce the risk of not successfully achieving their communicative goals. Furthermore, he finds a crucial causal relationship in the concept of risk aversion, which can be formulated as follows: "translators will tend to take risk X in the presence of reward structure Y" (2008, p. 326). Initially, Pym (2005) formulated the risk-aversion hypothesis as an explanation for explicitation, stating that translators make information explicit because they get punished for taking risks, and making information explicit allows them to avoid these risks. However, Pym (2008) elaborates on this and hypothesizes that in addition to explicitation, the standardization tendency is also a risk-averse strategy and the status of the tendency as a possible law depends on the fact whether the translator gets rewarded or not when taking a risk. More specifically, Pym puts forward the following formulation: "translators will tend to avoid risk by standardizing language and/or channeling interference, if and when there are no rewards for them to do otherwise" (p. 326). One of the consequences of the risk-aversion hypothesis is that, when translators have doubts about a certain solution to a translation problem, they might choose to reduce their "personal risk burden" by picking that which is normal or safe.

Similarly, Mauranen (2008) refers to the fact that other scholars have noticed that translations use "generally unmarked grammar, clichés, and typical, common lexis instead of the unusual or the unique" and that in translations, dialect is often replaced by standard language, punctuation is often normalized and target-language specific features are often exaggerated. She links this generally observed tendency towards conventionality with a general caution which is often attributed to translations, thereby referring to the idea that translations are supposed to "avoid margins or periphery and remain safely within the mainstream" (p. 40).

1.4.2 Linguistic normative reality in Flanders

This study was not only motivated by a number of theoretical concepts related to the field of corpus-based translation studies, but also by the highly specific language situation in Belgium, where there are currently three national languages: French,

German and Dutch. French is the official language in the southern region of Belgium, i.e. Wallonia. For the German-speaking community of Belgium, which is situated in Wallonia as well, German is the official language while in the northern region of Belgium, i.e. Flanders, the official language is Dutch. Finally, the Brussels-Capital Region is officially bilingual, where French and Dutch are the official languages.

1.4.2.1 Historical background

For the research presented in this dissertation, it is especially interesting to take a moment to zoom in on the language situation of the northern region, Flanders, where Dutch is the official language and to provide some information with regard to the historical background of this particular region. The summary given here is based on an introductory chapter in Geeraerts et al. (1999) and aims to shed light on a matter that is rather complex. First of all it should be noted that it was not until the renaissance that languages in Europe started to move towards linguistic standardization and in Belgium (even though the nation was not founded until 1830, we will use this reference for practical reasons), the movement towards language standardization was constrained by two main factors.

First, the Dutch-speaking part of Belgium failed to develop its language simultaneously with the Dutch language in the Netherlands which was due to the independence of the latter (circa 1585) while Belgium failed to gain independence and remained politically dependent on other foreign powers for two more centuries.

Secondly, French enjoyed considerable prestige and became the go-to language for everything related to culture and science in Europe . In Belgium too, French played an ever more important role in public fields such as science and political administration. As a result, there was no alternative development of a standard variety of Dutch in Belgium. Only under the reign of Willem I, who strongly promoted the Dutch language, did Dutch become the official language of the Flemish region (circa 1820). Although Belgium gained its independence in 1830, it was not until 1898 that Dutch was declared the second official Belgian language and French was no longer the only official language in Belgium.

From the 1960s onwards the Dutch-speaking area gained political autonomy, which enabled the region to focus on the standardization process of its own language rather than having to protect it from French. Initially, the official political approach was to seek a near complete adoption of the Netherlandic Dutch standard (Taeldeman, 1992), a language policy that sought to clear the language of Belgian variants and have them replaced by the official Standard (Netherlandic) Dutch variants so as to reach a common standard language for both regions. This approach resulted in a series of reference works, regular features in newspapers and even television shows whose main message was “do not say x, but y” in order to purify the language and to create this common standard. Presently matters are less rigid and in addition to General Standard Dutch, i.e.

the variety that is accepted in the entire Dutch-speaking area, there is room for both Belgian Dutch and Netherlandic Dutch varieties as well and these regional varieties are no longer consistently labeled as non-standard Dutch. In particular, the Dutch language consists of the following varieties: General Standard Dutch, Belgian Standard Dutch, Netherlandic Standard Dutch and non-standard Dutch. The language variety which is considered standard language in the entire Dutch-speaking area is General Standard Dutch and can be defined as follows:

> General Standard Dutch is the variety of Dutch which can be used in the public domain, viz. in all important sectors such as administration, administrative departments, administration of justice, education and mass media. In other words, General Standard Dutch is the variety of Dutch which can be used to communicate with people from outside one's familiar surroundings (so-called secondary relationships). Words, expressions, specific pronunciations, or constructions which are considered standard language can essentially easily be used in the aforementioned sectors and situations[4] (own translation).

The language varieties which are considered standard language in Belgium and the Netherlands only are Belgian Standard Dutch and Netherlandic Standard Dutch, respectively. Non-standard Dutch contains those linguistic alternatives which are not accepted as standard language anywhere in the Dutch-speaking area.

Although presently there is room for both Belgian Standard Dutch and Netherlandic Standard Dutch alongside General Standard Dutch, to a certain degree there is still a typical linguistic uncertainty to be found among Belgian Dutch speakers. More specifically, language users in Flanders tend to wonder more often than language users in the Netherlands whether a given expression, word, grammatical construction etc. is considered 'generally accepted language' or not:

> "I refer to the typical Flemish sentiment which is often described as "the Flemish linguistic uncertainty". It is a kind of automatic self-censorship on every language utterance in General Standard Dutch. It is a constant feeling of doubt whether the chosen form or word is "good Dutch". The ironic consequence is that Flemings often opt for the exogenous, unnatural options which causes their language to turn out worse instead of better. To my knowledge, a similar linguistic uncertainty is unknown to Dutchmen." (Deygers, 1998, p. 94, own translation).

Currently, the norms and regulations for General Standard Dutch which are supported by the *Nederlandse Taalunie* (the Dutch Language Union) can be consulted through

[4] http://taaladvies.net/taal/advies/tekst/85#standaardtaal, own translation

various reference works and websites, which we used to select our variables (see Chapter 2 for more details).

1.4.2.2 Norm design

As investigating prescriptive norm-conforming behavior is one of the main research goals of this dissertation, it is particularly important to provide a detailed description of the various types of norms which are used by the Dutch Language Union for writing their guidelines on standard language usage. The norms under investigation can be considered prescriptions which provide guidance when a language user is confronted with multiple options (Renkema, 1985, p. 140). Even though these prescriptions resemble linguistic rules, they are guidelines rather than actual rules and, as is often the case with guidelines, multiple voices can be heard. Renkema originally (1985) formulated seven types of norms which (may) apply if a given situation presents itself in which various linguistic options are possible: the historical norm; the authoritarian norm; the logical norm; the statistical norm; the norm of purity; the effect norm; and the esthetic norm (see also Burger & De Jong, 1991; Deygers, 1998; E. van der Spek, 1990). In the following paragraphs, more detailed information is provided with regard to each norm, how it should be interpreted, and how it can be used as a criterion during the decision process for determining whether something can be considered standard language or not.

According to the **historical norm**, tradition is an important factor and lexemes which have been accepted for a long time, are allowed to stay (Deygers, 1998). In other words, one should not divert from 'old rules' (Renkema, 1985).

Based on the **authoritarian norm**, language users should behave according to the example set by "authoritative speakers, authors and reference works" (Deygers, 1998, p. 80).

The **logical norm** states that language should be logical and, consequently, if something is not logical, it is bad language (Burger & De Jong, 1991, p. 79). An example of an expression in Dutch which should be avoided according to the logical norm, is the use of a double negation.

According to the **statistical norm**, language is everyone's possession and as soon as the majority of the language users use a given lexical item, it should be considered standard language. It appears that the principle behind the statistical norm conflicts with the authoritarian norm, which states that only a minority of the language users, viz. the authoritative community, is entitled to determine what is correct and what is not (Burger & De Jong, 1991).

In accordance with the **norm of purity**, all elements which are not endogenously Dutch should be excluded from the standard language. This norm raises a question

which is relevant with regard to the historical norm as well: "how far in time should one go back for determining what is truly Dutch and what is not?" (Burger & De Jong, 1991).

The purpose of the **effect norm** is to ensure that the "message gets across". In other words, if a reader does not understand a given word, expression or construction, this goes against the effect norm, whose aim is to avoid misunderstandings (Deygers, 1998).

Finally, the most subjective norm, the **esthetic norm**, stipulates that for a linguistic item to be acceptable as standard language, it should be pretty. Because of its subjectivity, the esthetic norm is often supported by one of the other norms (Renkema, 1985).

Taaladvies, i.e. a service by the Dutch Language Union which provides advice on language usage, applies a fixed methodology based on a mixture of the norms mentioned above when formulating their guidelines. It is especially the statistical norm which has an influence on whether a given linguistic item is considered General Standard Dutch, Belgian Standard Dutch, Netherlandic Standard Dutch or non-standard Dutch. The statistical norm is interpreted both in terms of frequency and attitude. More specifically, the *Taaladvies* team verifies by means of various corpora how often a given linguistic item occurs. Additionally, a panel of people with a keen sense of language is consulted so as to investigate the overall attitude with regard to the given linguistic item. More details on this panel and the methodology which is applied by the *Taaladvies* team in general are provided in Chapter 2.

1.4.3 The norm-conformity hypothesis

The previous paragraphs provided more information with regard to the underlying theoretical concepts which led to the central hypothesis under investigation in this study. First of all, there is Toury's law of growing standardization and related concepts such as normalization, conventionality and conservatism, which all more or less come down to the idea that, in general, translations are more conventional than non-translations. Second, Pym's risk aversion hypothesis and Mauranen's related concept of the cautious translator both refer to the idea that within translations there seems to be a tendency to avoid risks and, where possible, to take the safer option. Combining these two concepts with the highly specific language situation in Flanders resulted in the following central hypothesis, i.e. the norm conformity hypothesis:

> When given the choice between a norm-conform option and an interchangeable option which is identical in meaning but not norm conform, translations show a general preference for norm-conforming behavior.

In this dissertation, we distinguish between two types of norms: descriptive and prescriptive norms. To put it extremely briefly: we consider descriptive norms to be

based on how things are done rather than how things should be done, the latter being the idea behind prescriptive norms. The following paragraphs provide more details with regard to the differences between descriptive and prescriptive norms and how we aim to verify for both types to what degree they are being conformed to.

1.4.3.1 Descriptive norms

Malmkjaer (2008) refers to Hewitt (2005) who defines norms from a socio-cultural point of view as follows: "a way of behaving or believing that is normal for a group or culture" (p. 50). In sociology, descriptive norms are often defined as norms which describe actual behavior, irrespective of whether the observed behavior is approved of or not.

One of the goals of this study is to investigate genre variation in translations and we want to find out whether the language usage in translations varies according to the genre the translation belongs to. Or, from a different perspective: we want to verify to what extent a given genre in translation conforms to what is common for that particular genre. As norms often result in regularities of behavior and can be detected by means of linguistic features which themselves however, are not norms (Eskola, 2004), investigating language usage *as is* in various genres, allows us to deduct the descriptive norms for a number of aspects per genre. In order to avoid confusion, we formulated the following definition for the genre-determined descriptive norms under investigation: a descriptive norm is a (mostly) implicit guideline which only applies to a given genre and which is (mostly) picked up inductively by language users of this genre.

1.4.3.2 Prescriptive norms

Whereas the first two case studies focus on investigating genre-determined norms and whether translations conform to these descriptive norms, the last two case studies were carried out to map prescriptive norm-conforming behavior in translations. More particularly, the main goal is to verify whether there is a difference between translations and non-translations and the degree to which they conform to the norms with regard to General Standard Dutch which are issued by *Taaladvies*, a service offered by the Dutch Language Union.

More specifically, these two case studies are based on the linguistic normative reality in Flanders and the research focus here lies on investigating whether translations conform to prescriptive norms, viz. clear guidelines with regard to what is General Standard Dutch and what is not. Unlike the genre-determined norms investigated in the first two case studies, the guidelines with regard to General Standard Dutch are prescriptive and can easily be consulted in various reference works (see Section 2.1.2.3 for more details).

Chapter 2
Research Design

This chapter will provide the details of the methodology used for investigating differences in linguistic norm-conforming behavior between translations and non-translations. As can be seen in Figure 2, we first formulated four main hypotheses, one for each case study, the theoretical background for which is described in Section 1.4. In order to investigate a given hypothesis, an appropriate set of variables (constructed as profiles) was selected. Section 2.1 provides information regarding (i) the various hypotheses that were operationalized in this study and (ii) the variable selection per case study. Section 2.2 gives a detailed overview of the corpus which was used and how the data were extracted from the corpus. In Section 2.3, a brief overview is given with regard to the manual validation of the extracted data. The resulting datasets were analyzed by means of a highly specific exploratory statistical technique, i.e. profile-based correspondence analysis (Plevoets, 2008). Section 2.3.1 goes into the details of these analyses and how they result in two-dimensional plots which visualize the linguistic distances between the language varieties in our dataset and which allow us to interpret the data and generate conclusions with regard to the initial hypotheses.

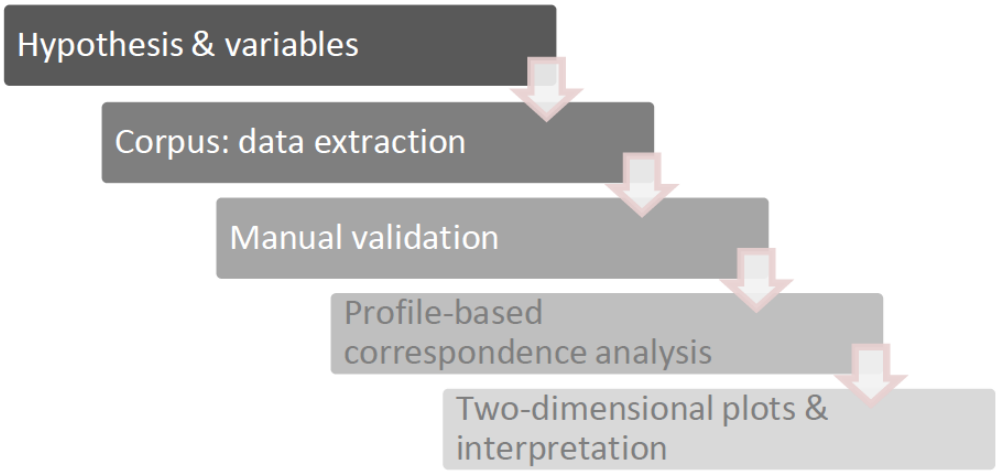

Figure 2 Schematic overview of the research design

2.1 Hypotheses and variable selection

2.1.1 Hypotheses

For each case study, we formulated a hypothesis which is related to the concept of norm conformity and the assumption that, in general, translations behave differently in comparison to non-translations and tend to exhibit linguistic norm-conform behavior. Moreover, our hypotheses predict that translations will show linguistic behavior which is conform to two types of norms, viz. descriptive norms and prescriptive norms. Overall, it should be noted that the proposed hypotheses are based on rather broad assumptions and are therefore of a more general nature, as is typical for the kind of exploratory research which is presented here. However, analyzing the results of each case study resulted in more fine-grained, explanatory hypotheses which are formulated in the concluding remarks of each case study.

In the first case study we investigate the dispersion of formal versus neutral language in our corpus as an operationalization of the following hypothesis: "Genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to formal versus neutral lexemes as their corresponding non-translated genres". For example, if non-translated Journalistic Texts show a preference for neutral language, we hypothesize that translated Journalistic Texts will show a similar preference.

In the second case study we look at the dispersion of more endogenous lexical alternatives versus English loanwords under the hypothesis that "Genres in translation

conform to descriptive, genre-determined norms and show a similar preference with regard to English loanwords and their interchangeable more endogenous lexemes as their corresponding non-translated genres". We hypothesize, for example, that when non-translated Tourist Information shows a preference for English loanwords, translations will follow this trend and translated Tourist Information will show a similar preference for English loanwords.

The language situation in Belgium and more particularly in the Dutch-speaking area is highly specific and fairly unique and provided a solid base for investigating prescriptive norm-conforming behavior in translated versus non-translated language (see also Section 1.4.2). In this case study, a hypothesis was formulated whose main aim was to investigate the dispersion of standard language versus non-standard language: "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable non-standard Dutch lexeme which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts."

The final case study is similar to the third case study as it also focuses on the Belgian language situation to investigate norm-conforming behavior. In this case study, the main goal is to examine the dispersion of General Standard Dutch versus Belgian Standard Dutch (for more details see Section 2.1.2.3). The particular hypothesis was formulated as follows: "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable Belgian Standard Dutch lexeme, which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts".

2.1.2 Variable selection

2.1.2.1 Formal versus neutral language (Case Study I)

Table 1 shows the profile set for the first case study which aims to verify whether translations conform to the descriptive norm with regard to genre-specific formality. As the goal was to investigate norm conformity with regard to formality, we needed profiles which consist of one formal variant and one neutral variant which both frame the same concept. In total, the set contains eight profiles, a detailed survey of which can be found in Section Chapter 3. There were three selection criteria to build this profile set. First, a recent source (e.g. dictionaries, language advice literature) must recognize one variant to be formal and the other to be neutral. Secondly, there is no conflicting information to be found in the remaining sources, and thirdly, the only difference between the variants within one profile is a difference in formality (and not, for example, a regional difference).

Table 1 Case Study I: profile frequencies per genre and (source) language variety.

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL	TOURIST	INSTRUCT		DU_orig	DU<EN	DU<FR		TOTAL
maar	NEUTRAL	BUT	143	971	814	980	5778	487	154		5033	2081	2213		9327
echter	FORMAL		72	310	237	114	676	35	8		716	328	408		1452
proberen	NEUTRAL	TO TRY	6	24	32	69	378	26	17		303	137	112		552
trachten	FORMAL		6	24	17	41	39	6	0		75	26	32		133
nu	NEUTRAL	NOW	36	294	392	493	1528	109	50		1452	924	526		2902
thans	FORMAL		12	92	20	30	15	1	0		36	30	104		170
als	NEUTRAL	IF	49	101	77	35	342	28	88		419	173	128		720
indien	FORMAL		41	75	9	16	9	2	11		73	28	62		163
in	NEUTRAL	IN	38	385	167	535	1730	321	32		1691	1062	455		3208
te	FORMAL		16	75	4	43	91	35	0		137	52	75		264
al	NEUTRAL	ALREADY	80	380	316	380	2323	196	99		1982	860	932		3774
reeds	FORMAL		67	251	87	181	162	12	18		436	186	156		778
moeten + inf	NEUTRAL	TO HAVE TO	115	238	182	200	465	15	23		511	252	475		1238
dienen + inf	FORMAL		209	231	131	41	128	6	80		545	124	157		826
numeral *+ keer*	NEUTRAL	TIMES	23	58	15	34	257	11	13		200	93	118		411
numeral *+ maal*	FORMAL		13	44	6	10	22	1	1		49	12	36		97
TOTAL			926	3553	2506	3202	13943	1291	594		13658	6368	5989		26015

2.1.2.2 English loanwords versus more endogenous lexemes (Case Study II)

The main goal of Case Study II was to verify to what extent translations conform to the genre-determined norm concerning the use of English loanwords versus more endogenous alternatives. Table 2 shows the frequencies of the seven profiles which were gathered for this task. The following steps were taken for building the profile set: First of all, we extracted a list with possible loanwords from the Dutch Parallel Corpus by analyzing the list of lemmas which occur in the corpus. Secondly, we verified the status of these candidate loanwords by means of Van Dale dictionary (den Boon & Geeraerts, 2010-2013) so as to determine whether the loanword candidates were indeed English loanwords. Specific details with regard to how this profile set was built are provided in Section Chapter 4.

Table 2 Case Study II: profile frequencies per genre and (source) language variety

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL		DU_orig	DU<EN	DU<FR		TOTAL
ploeg team	endogenous term loanword	TEAM	2 63	5 115	0 2	3 148	130 270		70 268	10 103	60 227		140 598
dienst-verlening service	endogenous term loanword	SERVICE	18 0	75 14	9 1	117 57	28 18		111 33	43 46	93 11		247 90
baan job	endogenous term loanword	JOB	0 6	4 13	35 45	21 24	74 144		57 142	44 25	33 65		134 232
afdeling unit	endogenous term loanword	DIVISION	23 0	76 2	3 0	62 9	172 11		208 12	65 10	63 0		336 22
o&o r&d	endogenous term loanword	"RESEARCH & DEVELOP-MENT"	1 0	84 15	11 0	23 28	15 15		44 25	90 32	0 1		134 58
partnerschap partnership	endogenous term loanword	PARTNERSHIP	1 0	15 10	7 1	105 59	3 19		75 24	44 50	12 15		131 89
hulpmiddel tool	endogenous term loanword	TOOL	12 0	30 20	1 0	15 11	10 21		33 18	9 12	26 22		68 52
TOTAL			126	478	115	682	930		1120	583	628		2331

2.1.2.3 General Standard Dutch, Belgian Standard Dutch, Netherlandic Standard Dutch and non-standard Dutch (Case Studies III and IV)

The main focus of the last two case studies presented in this dissertation lies on measuring the differences in prescriptive norm conforming behavior in various translated and non-translated genres. More specifically, we want to verify to what degree translations make use of General Standard Dutch as opposed to two other varieties, viz. non-standard Dutch (Case Study III) and Belgian Standard Dutch (Case Study IV). In order to find profiles which consist of a General Standard Dutch alternative on the one hand and a non-standard Dutch or Belgian Standard Dutch on the other, we consulted *Taaladvies*, a language advice service offered by the Dutch Language Union. As shown in Table 3 and Table 4, this resulted in a set of seven profiles for Case Study III and nine profiles for Case Study IV.

Table 3 Case Study III: profile frequencies per genre and (source) language variety

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL	TOURIST		DU_orig	DU<EN	DU<FR		TOTAL
een van de	GSD	ONE OF THE	39	100	45	106	362	52		313	245	146		704
één van de	NSD		37	115	16	165	72	55		202	189	69		460
pv + inf + vd	GSD	VERBAL END GROUP	7	53	11	16	21	1		32	19	58		109
vd + pv + inf	GSD		3	16	10	11	21	1		28	16	18		62
pv + vd + inf	NSD		3	15	1	2	5	1		21	3	3		27
te veel	GSD	TOO MUCH	5	10	12	22	158	4		120	36	55		211
teveel	NSD		3	5	6	14	5	0		24	4	5		33
ten(minste)_goed	GSD	AT LEAST	65	42	5	62	51	1		86	43	97		226
ten(minste)_fout	NSD		5	26	2	8	6	1		21	19	8		48
een beroep doen op	GSD	TO MAKE AN APPEAL TO	18	54	17	51	80	3		87	29	107		223
beroep doen op	NSD		9	7	3	2	7	0		18	3	7		28
zodra	GSD	AS SOON AS	17	31	12	41	103	3		79	37	91		207
van zodra	NSD		5	9	0	1	3	0		12	1	5		18
beginnen + te + inf	GSD	TO START TO	2	1	0	4	19	1		13	6	8		27
beginnen + inf	NSD		0	3	0	1	7	4		9	2	4		15
TOTAL			218	487	140	506	920	127		1065	652	681		2398

Table 4 Case Study IV: profile frequencies per genre and (source) language variety

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL		DU_orig	DU<EN	DU<FR		TOTAL
geraken	BSD	TO GET	0	7	9	12	97		65	35	25		125
raken	GSD		0	15	11	14	199		127	61	51		239
luik	BSD	PART	3	22	5	4	10		18	5	21		44
onderdeel	GSD		12	35	19	45	47		83	42	33		158
tewerkstelling	BSD	EMPLOYMENT	1	3	7	37	13		19	2	41		62
werkgelegenheid	GSD		12	30	55	23	11		60	38	33		131
ten laatste	BSD	AT THE LATEST	21	23	1	8	6		24	8	27		59
uiterlijk	GSD		86	15	1	8	1		73	4	34		111
verwittigen	BSD	TO WARN	4	6	0	0	5		8	0	7		15
waarschuwen	GSD		1	17	6	7	61		30	43	19		92
vragen dat	BSD	TO DEMAND	5	46	7	13	13		41	7	36		84
eisen dat	GSD		5	8	5	3	13		9	13	12		34
proper	BSD	CLEAN	4	0	1	2	12		9	3	7		19
schoon	GSD		1	1	10	9	18		23	14	2		39
telkens	BSD	EACH TIME	2	4	0	2	14		9	3	10		22
telkens als	GSD		2	7	3	1	15		10	11	7		28
eenmaal	BSD	AS SOON AS	0	9	4	1	23		20	10	7		37
zodra	GSD		11	29	12	40	103		76	37	82		195
TOTAL			170	277	156	229	661		704	336	454		1494

Given its importance to this study, the following paragraphs provide elaborate details on the process which determines whether a given linguistic expression is considered

General Standard Dutch, Belgian Standard Dutch or non-standard Dutch as well as the differences between these three language varieties.

Dutch is the official language of three geographical areas: the Netherlands, the northern part of Belgium, i.e. Flanders, and Surinam. In 1980, the Netherlands and the Dutch-speaking area of Belgium founded the *Nederlandse Taalunie* (Dutch Language Union), a cooperation whose purpose is "cooperating on language, language teaching and literature, [as this] can produce considerable benefits" (H. Van der Spek, 2004, p. 3) and in 2004, Surinam became an associate member of the *Taalunie*. The achievements of the *Taalunie* are plentiful, one of them being the development of the website *Taalunieversum*, which consists of various sub-sites.

One of these sections is the *Taaladvie*s website[1], a language advice service which, among other items, contains a large FAQ database with questions and answers related to language issues and correct language usage. According to the *Taalunie*, the database consists over 1200 issues and is still expanding (H. Van der Spek, 2004, p. 12). In addition to information about spelling and grammar issues, *Taaladvies* also provides advice regarding the usability of certain linguistic alternatives according to the specific language variety of Dutch they belong to, viz. General Standard Dutch, Belgian Standard Dutch, Netherlandic Standard Dutch or non-standard Dutch. The *Taalunie* founded the advice database, but holds no responsibility for its contents and does not regard the linguistic advice issued by *Taaladvies* as officially binding, but as reliable guidelines regarding linguistic issues. In order to determine how these FAQs are dealt with, the step-by-step procedure which is used by *Taaladvies* was investigated and is described below in detail.

If a language user asks a question with regard to a certain linguistic phenomenon via the electronic form on the *Taaladvies* website this question is answered by a contributor of *Genootschap Onze Taal* or *Taaltelefoon* based on the information that is readily available in the *Taaladvies* database.

Het Genootschap Onze Taal is a Dutch society of language enthusiasts. *Het Genootschap* publishes a unique journal which deals with various linguistic topics in a scholarly, yet easily readable way. Additionally, *Het Genootschap* offers advice on language usage through conferences, electronic newsletters, books and a website[2]. The Belgian counterpart of *Het Genootschap*, is the *Taaltelefoon*[3], which was founded by the Flemish

[1] http://taaladvies.net/

[2] http://onzetaal.nl/over

[3] http://taaltelefoon.vlaanderen.be/nlapps/docs/default.asp?id=1265&order=

Government. Similar to *Het Genootschap*, the *Taaltelefoon* answers questions related to spelling, punctuation, use of words, grammar, pronunciation, style, etc.

If a question arises which cannot be answered based on previously answered questions and the resulting linguistic advice in the database, additional measures are taken. Moreover, a fixed methodology[4] is applied in order to provide linguistic advice on language usage which is related to the differences between Dutch in Belgium and Dutch in the Netherlands.

a. Belgian as well as Dutch newspapers are consulted by project contributors in order to determine the frequencies of a given linguistic alternative. These newspapers are consulted by means of Mediargus[5] on the one hand and LexisNexis[6] on the other hand. Mediargus is the digital media platform for the Flemish daily press which offers an archive with millions of news articles from newspapers and magazines. Similarly, LexisNexis is a newspaper database which was especially developed for educational purposes and which contains Dutch newspapers only. Both the absolute frequencies and the relative frequencies are then added to a table which provides a survey of the occurrences. These observations are used to determine whether a linguistic alternative can be considered standard language or not in Belgium, in the Netherlands or in general. These searches are not carried out in a fixed set of newspapers, but overall the largest and/or most popular newspapers as well as a number of regional newspapers are consulted. The Belgian newspapers which are consulted are: *Het Belang van Limburg*, *De Gazet van Antwerpen*, *Het Laatste Nieuws*, *Metro* (Belgium), *De Morgen*, *Het Nieuwsblad*, *De Standaard*, and *De Tijd*. The Dutch newspapers are *Het Algemeen Dagblad*, *Het Financiële Dagblad*, *NRC Handelsblad*, *De Telegraaf*, *Trouw*, *Metro* (the Netherlands), *De Volkskrant*, *Haagsche Courant*, *Brabants Dagblad*, and *Dagblad van het Noorden*. If there are data available by means of Mediargus and/or LexisNexis, the following criteria apply:

    - If a given linguistic item is limited to a frequency of less than 5% in comparison to the General Standard Dutch alternative in Mediargus and/or LexisNexis, then this item cannot be considered Belgian or,

[4] The information provided above on the activities of *Taaladvies* was given by one of the contributors and it was stressed that the entire process should not be considered a rigid scientific methodology, and that variations are possible depending on the case at hand.

[5] http://www.mediargus.be/pg/expages/read/About/

[6] http://www.lexisnexis.be/dutch/products/krantenbank.page#

respectively, Netherlandic Standard Dutch. The linguistic item is, in this case, not sent to the panel for approval (see below).

- If a given linguistic item, in comparison to the frequency of the General Standard Dutch alternative, has a frequency of 50% or more in Mediargus and/or LexisNexis, the linguistic item is a possible candidate for the label Belgian or Netherlandic Standard Dutch.

- If a given linguistic item has a frequency of between 5% and 50% in comparison to the frequency of the General Standard Dutch alternative in Mediargus and/or LexisNexis, its status is unclear and the matter is passed on for further investigation to the panel of informants (see below).

Due to methodological limitations, however, it is often impossible to calculate exact numbers for the alternatives. Mediargus, for instance, does not allow for function word searches, such as prepositions (e.g. *op de bus* (on the bus)). Additionally, it is not possible to exclude certain irrelevant meanings of a given word. The corpus-based results are often used as a mere indication of how a given linguistic alternative is used.

b. In addition to the newspaper corpus, various reference works are consulted to determine how a given linguistic item is described. The most recent versions of the following dictionaries are used: *Grote Van Dale* (2010-2013); *Van Dale Hedendaags Nederlands* (2002); *Verschueren* (1996); *Koenen* (2006); *Het Witte Woordenboek Nederlands* (2007); *Prisma Handwoordenboek Nederlands* (2010); *Vlaams-Nederlands woordenboek* (2004). Additionally, several reference works with regard to advice on language usage are consulted: *Woordenboek Correct Taalgebruik (*2004*)*; *Correct Taalgebruik* (2001); *Taalwijzer* (1998); *Stijlboek VRT* (2003).

c. Steps a. and b. do not always result in clear answers to how a given linguistic item should be used. Therefore, the permanent *Taaladvies* contributors select various newspaper articles which contain several linguistic alternatives language users wish to know more about. These articles are then presented to a panel of standard language users who are active in the socio-cultural sector, both in Belgium and the Netherlands. The panel members are asked to correct the articles and, if desired, comment on their corrections. The results

of this query are presented in a table which shows the proportions in terms of percentages, the results of which are interpreted as follows:

- If less than 33% of the panel members from one of the linguistic regions (e.g. Flanders) corrects an alternative, it is considered accepted to a sufficient degree to be labeled standard language in that specific linguistic region (e.g. Standard Belgian Dutch).

- If 50% or more of the panel members from one of the linguistic regions corrects an alternative, it is not considered accepted to a sufficient degree and cannot be labeled standard language in that specific linguistic region.

- If between 33 and 50% of the panel members from one of the linguistic regions corrects an alternative, it is considered a borderline case and it is unclear whether the alternative is accepted or not by the language users. In this stage, it is not decided yet whether the alternative can be considered standard language or not. In certain cases, the obtained information is added to the relevant answer section of the FAQ website, e.g. "status unclear".

In order to avoid a Belgian or a Dutch bias, *Taaladvies* aims at a panel which is evenly represented by Dutchmen and Belgians (of different ages and from different regions in Belgium and the Netherlands). The current panel (2013) consists of 30 Belgian and 37 Dutch permanent members. Occasionally, a member leaves the panel or a new member is added to the panel. Each panel member has either an educational background in language and/or is professionally active with language (e.g. teachers, editors, journalists etc.).

Throughout the process, the corpus-based frequencies and the panel's opinion are the most important factors. The reference works often indicate that the status of a given linguistic alternative is not entirely clear, but the information they provide regarding the alternative's usage is sometimes out of touch with linguistic reality. Especially in reference works containing advice on language usage there is a tendency to disapprove of Belgian alternatives. The corpus-based approach and consulting the panel provide a more balanced image of the general linguistic attitude towards certain alternatives.

The *Taaladvies* contributors use the information provided by the steps in the process to write articles with linguistic advice about a given alternative and how it can be used. These advice articles are then brought up for discussion and presented to the TAO (*Taaladviesoverleg)*. The TAO was installed by the *Taalunie* and is a cooperation between

people with a background in advice on language usage, authors of linguistic reference works and text producers from the Netherlands and Belgium. Finally, the advice articles that are approved by the TAO are published on the *Taaladvies* website and added to the database. It should be noted that the actual texts for the advice articles are the result of a cooperation between contributors of the *Taaltelefoon* and the *Genootschap Onze Taal*.

2.2 Corpus and data extraction

Given this study's research questions, the corpus requirements were highly specific: the corpus should (i) be parallel, (ii) contain various genres, (iii) have translations from at least two source languages, and (iv) should have Belgian Dutch as the central language. There is one existing corpus which meets these highly specific requirements: the Dutch Parallel Corpus (Macken et al., 2011). It is a high-quality, parallel corpus for three languages: Dutch, French and English which consists of five genres. All texts in the corpus have been cleared of copyright by means of written agreements between publishers or authors and the DPC team (De Clercq & Montero Perez, 2010), which makes it readily available for research, as there are no restrictions with regard to publishing the results, sharing the data, etc.

2.2.1.1 Dutch Parallel Corpus

Table 5 Survey of the contents of the Dutch Parallel Corpus per genre and (source) language

Genre	Sub-genre	Basic-level category	Original Dutch	Dutch translated from French	Dutch translated from English
Literature	Fictional texts	Novels	✔	✔	✔
	Non-fictional texts	(auto-) Biographies		✔	
		(self-) Presentations		✔	
		Essayistic texts	✔	✔	
		Expository works	✔		✔
Journalistic Texts	Comment articles	Background articles	✔	✔	✔

		Columns	✔	✔	✔
		Editorials		✔	✔
	News articles		✔		✔
	Informative documents		✔		
Instructive texts	Manuals		✔	✔	✔
	Procedure descriptions		✔	✔	
	Internal legal documents		✔	✔	
Administrative texts	Internal & external correspondence		✔	✔	
	Official speeches		✔		✔
	Proceedings of parliamentary debates		✔	✔	✔
	Yearly reports		✔	✔	✔
	Legislation		✔	✔	
	Minutes of meetings		✔	✔	
External communication	(self-)Presentations		✔	✔	✔
	Informative documents		✔	✔	✔
	Press releases and newsletters		✔	✔	✔
	Promotion and advertising material		✔	✔	✔
	Scientific texts		✔	✔	✔
	Yearly reports		✔		

Table 5 provides an overview of the contents of the Dutch Parallel Corpus based on the core XML-files and shows the genres on the one hand and non-translated Dutch, Dutch

translated from French, and Dutch translated from English on the other hand. The corpus structure and its contents were based both on a user requirements study and a thorough analysis of other parallel corpus projects such as the Europarl corpus and the European Corpus Initiative (Macken et al., 2011, pp. 376-377). Unfortunately, the DPC does not contain text material for every genre and sub-genre for all (source) language varieties, as can be seen in Table 5.

2.2.1.1.1 Genres

Based on the information provided by Macken et al. (2011), the corpus consists of five genres: Literature, Journalistic Texts, Instructive Texts, Administrative Texts and External Communication. Within these genres, a number of sub-genres can be distinguished.

Literature consists of fictional texts and non-fictional texts. Fictional texts contain only one kind of material, i.e. novels. Non-fictional texts on the other hand consist of four kinds of documents: (auto-)biographies; (self-)presentations of organizations, projects or events; essayistic texts; and expository works of a general nature. Whereas the first three types are rather straightforward, the fourth, i.e. expository works of a general nature, is rather vague. Examining the additional metadata for this category showed that there are 35 texts within this category and that these are general (background) articles within a varied range of subjects.

Journalistic Texts contain three basic-level categories, viz. comment articles; news articles; and informative documents of a general nature. Comment articles are divided into background articles; columns; and editorials, while news articles and informative documents of a general nature have no further subdivisions.

Instructive texts are divided into three sub-types, viz. manuals; procedure descriptions; and internal legal documents.

In **administrative texts**, there are six basic-level categories, viz. internal and external correspondence; official speeches; proceedings of parliamentary debates; yearly reports; legislation; and minutes of meetings. The DPC User Manual[7] describes administrative texts as "texts produced within an institutional context, their circulation is usually restricted to internal use or to use within a limited circle of organizations tied to the institution" (DPC Manual, p.4).

External communication is an umbrella category which contains six basic-level categories, viz. (self-)presentations of organizations, projects, events; informative

[7] https://www.kuleuven-kulak.be/dpc/manual/DPC.pdf

documents of a general nature; press releases and newsletters; promotion and advertising material; scientific texts; and yearly reports. The description for external communication as a genre in the DPC manual is as follows: "texts of an informative and/or persuasive nature that are characterized by a wide circulation and meant for external use in general or for peers in a broad sense" (DPC Manual, p. 4).

2.2.1.1.2 (Source) language varieties

In addition to genre-based sub-corpora, the Dutch Parallel Corpus contains text material in three languages: Dutch, French and English. These three sub-corpora consist of regional language varieties: while Dutch is an umbrella category for Belgian Dutch and Netherlandic Dutch, French is an umbrella category for Belgian French and French French, and English contains both British English and American English. However, not all main genres are represented in these regional varieties. More particularly, there are no data available for Fictional Literature, Non-fictional Literature and Journalistic Texts in American English. Similarly, the corpus contains no Fictional Literature in Belgian French.

With regard to these language-based sub-corpora, the DPC can be defined as a parallel corpus, a monolingual comparable corpus and a multilingual comparable corpus as the corpus contains (i) translations and the corresponding source texts; (ii) comparable translated and non-translated texts in a single language; and (iii) comparable translated and non-translated texts across languages. For example, the Dutch sub-corpus consists of three (source) language varieties based on their translation status and source language, viz. non-translated Dutch, Dutch translated from French and Dutch translated from English.

2.2.1.2 Size

DPC contains over ten million words and the corpus is balanced both with regard to genre and translation direction. Additionally, it should be noted that the corpus only contains full texts and makes no use of text samples. A detailed overview of the precise contents in terms of tokens per genre and translation direction can be found in Table 6. According to Macken et al., these word counts "are all based on clean text, i.e. without figures, tables and graphs" (2011).

Table 6 Survey of the DPC contents per genre and translation direction (Macken et al., 2011)

Text Type	SRC→TGT	DU	EN	FR	TOTAL

Administrative Texts	EN→DU	255,155	246,137	0	501,292
	FR→DU	307,886	0	322,438	630,324
	DU→EN	249,410	257,087	0	506,497
	DU→FR	280,584	0	301,270	581,854
	Total	1,093,035	503,224	623,708	2,219,961
External Communication	EN→DU	278,515	272,460	0	550,975
	FR→DU	233,277	0	250,604	483,881
	DU→EN	246,448	255,634	0	502,082
	DU→FR	241,323	0	270,074	511,397
	X→D/E	21,679	20,118	0	41,797
	X→D/E/F	14,192	14,953	15,743	44,888
	Total	1,035,434	563,165	536,421	2,132,020
Instructive Texts	EN→DU	340,097	327,543	0	667,640
	FR→DU	40,487	0	42,017	82,504
	DU→EN	19,011	20,696	0	39,707
	DU→FR	110,278	0	115,034	225,312
	X→D/F	59,791	0	73,758	133,549
	X→D/E	299,996	296,698	0	596,694
	X→D/E/F	138,673	145,103	166,836	450,612
	Total	1,008,333	790,040	397,645	2,196,018
Journalistic Texts	EN→DU	262,768	264,900	0	527,668
	FR→DU	240,785	0	265,530	506,315
	DU→EN	250,580	259,764	0	510,344
	DU→FR	314,989	0	340,319	655,308
	Total	1,069,122	524,664	605,849	2,199,635
Literature	EN→DU	148,488	143,185	0	291,673
	FR→DU	186,799	0	186,620	373,419
	DU→EN	346,802	361,140	0	707,942
	DU→FR	323,158	0	348,343	671,501
	Total	1,005,247	504,325	534,963	2,044,535
Grand Total		5,211,171	2,885,418	2,698,586	10,795,175

2.2.1.3 Linguistic annotations

The DPC has been enriched with the following basic linguistic annotations: lemmas and part-of-speech information. A first step in adding this information to the corpus consists of tokenizing all its texts which comes down to splitting sentences into words while punctuation marks are being stripped off. A second step involves part-of-speech tagging, a process through which all words in the corpus are assigned their morpho-syntactic class. Various existing part-of-speech taggers were used for the languages in the corpus: (i) the D-Coi part-of-speech tagger (Van Den Bosch, Schuurman, & Vandeghinste, 2006) for Dutch; (ii) the combined memory-based part-of-speech tagger from the MBSP toolkit (Daelemans & Van Den Bosch, 2005) for English ; and (iii) the Treetagger (Schmid, 1994) for French. In addition to different taggers, different tag sets were used: (i) the CGN part-of-speech tag set (Van Eynde, Zavrel, & Daelemans, 2000) for Dutch; (ii) the Penn Treebank tag set (Marcus, Santorini, & Marcinkiewicz, 1993) for English; and (iii) the LIMSI parameter file (Allauzen & Bonneau-Maynard, 2008) based on the GRACE part-of-speech tag set (Paroubek, 2000) for French. These tools were used to

provide the lemmas for Dutch and English as well; the lemmas for French, however, have not been added to the corpus.

To evaluate the automatically predicted word classes and lemmas, a sample of the DPC was manually verified. In total 841,000 tokens were evaluated by means of accuracy scores, the results of which can be seen in Table 7 (Macken et al., 2011, p. 384). On average, an accuracy score of 97.6% was reached for the lemmatization and 95.2% for the part-of-speech tagging. For Dutch, the focus language in this study, the accuracy scores are slightly lower than the average scores: 96.5% for the lemmas and 94.8% for the part-of-speech tagging. For this study, it is paramount that these scores are sufficiently high, as some of our corpus queries rely on the linguistic annotations of the corpus and the search results are thus dependent on these annotations. One of the consequences of the 96.5% lemma accuracy score is that, per 100 words, 3.5 words are erroneously lemmatized, i.e. 1 word per 28.6 words. Similarly, a 94.8 accuracy score for the part-of-speech encoding equals 5.2 erroneous part-of-speech tags per 100 words, i.e. 1 word per 19.2 words. The Dutch part of the DPC contains 5,811,095 words and 343,187 sentences, which results in sentences with an average length of 16.9 words. If we match the accuracy scores with these figures, this means that for each 1.7 sentences, 1 word is lemmatized erroneously. Additionally, for each 1.1 sentence, there is 1 word with an erroneous part-of-speech tag. It should be noted that these results may have influenced the output results of our corpus queries when these were based on lemmas and/or part-of-speech codes.

Table 7 Sample size expressed in number of tokens and accuracy scores obtained by the lemmatizers and part-of-speech taggers (Macken et al., 2011, p. 384)

	Sample size (tokens)	Lemmas	Part of Speech
Dutch	211,000	96.5%	94.8%
French	330,000	98.1%	94.6%
English	300,000	98.1%	96.2%

Below, an example can be found of how this linguistic enrichment is shown in the corpus by means of an English sentence and its translations into Dutch and into French.

(1) **As** IN(IN;as) **President** NNP(NNP;President) **of** IN(IN;of) the DT(DT;the) **Parliament** NNP(NNP;Parliament) **,** ,(,;,) I PRP(PRP;I) **am** VBP(VBP;be) **bound** VBN(VBN;bind) **to** TO(TO;to) **tell** VB(VB;tell) **you** PRP(PRP;you) **that** IN(IN;that) I PRP(PRP;I) **believe** VBP(VBP;believe) **the** DT(DT;the) **elections** NNS(NNS;election) **,** ,(,;,) **whatever** WDT(WDT;whatever) **the** DT(DT;the) **specifics** NNS(NNS;specific) **of** IN(IN;of) **individual** JJ(JJ;individual) **results** NNS(NNS;result) **,** ,(,;,) **were** VBD(VBD;be) **disappointing** JJ(JJ;disappointing) **on** IN(IN;on) **two** CD(CD;two) **counts** NNS(NNS;count) **.** .(.;.)

(2) **Als** VZ(VZ(init);als) **Voorzitter** N(N(soort,ev,basis,zijd,stan);voorzitter) **van** VZ(VZ(init);van) **het** LID(LID(bep,stan,evon);het) **Europees** ADJ(ADJ(prenom,basis,zonder);Europees) **Parlement**

N(N(soort,ev,basis,onz,stan);parlement) **kan** WW(WW(pv,tgw,ev);kunnen) **ik** VNW(VNW(pers,pron,nomin,vol,1,ev);ik) **echter** BW(BW();echter) **niet** BW(BW();niet) **verhelen** WW(WW(inf,vrij,zonder);verhelen) **dat** VG(VG(onder);dat) **de** LID(LID(bep,stan,rest);de) **verkiezingen** N(N(soort,mv,basis);verkiezing) **,** LET(LET();,) **ongeacht** VZ(VZ(init);ongeacht) **de** LID(LID(bep,stan,rest);de) **details** N(N(soort,mv,basis);detail) **van** VZ(VZ(init);van) **afzonderlijke** ADJ(ADJ(prenom,basis,met-e,stan);afzonderlijk) **uitslagen** N(N(soort,mv,basis);uitslag) **,** LET(LET();,) **naar** VZ(VZ(init);naar) **mijn** VNW(VNW(bez,det,stan,vol,1,ev,prenom,zonder,agr);mijn) **opvatting** N(N(soort,ev,basis,zijd,stan);opvatting) **op** VZ(VZ(init);op) **twee** TW(TW(hoofd,prenom,stan);twee) **punten** N(N(soort,mv,basis);punt) **teleurstellend** ADJ(ADJ(vrij,basis,zonder);teleurstellend) **waren** WW(WW(pv,verl,mv);zijn) **.** LET(LET();.)

(3) **Als** VZ(VZ(init);als) **Voorzitter** N(N(soort,ev,basis,zijd,stan);voorzitter) **van** VZ(VZ(init);van) **het** LID(LID(bep,stan,evon);het) **Europees** ADJ(ADJ(prenom,basis,zonder);Europees) **Parlement** N(N(soort,ev,basis,onz,stan);parlement) **kan** WW(WW(pv,tgw,ev);kunnen) **ik** VNW(VNW(pers,pron,nomin,vol,1,ev);ik) **echter** BW(BW();echter) **niet** BW(BW();niet) **verhelen** WW(WW(inf,vrij,zonder);verhelen) **dat** VG(VG(onder);dat) **de** LID(LID(bep,stan,rest);de) **verkiezingen** N(N(soort,mv,basis);verkiezing) **,** LET(LET();,) **ongeacht** VZ(VZ(init);ongeacht) **de** LID(LID(bep,stan,rest);de) **details** N(N(soort,mv,basis);detail) **van** VZ(VZ(init);van) **afzonderlijke** ADJ(ADJ(prenom,basis,met-e,stan);afzonderlijk) **uitslagen** N(N(soort,mv,basis);uitslag) **,** LET(LET();,) **naar** VZ(VZ(init);naar) **mijn** VNW(VNW(bez,det,stan,vol,1,ev,prenom,zonder,agr);mijn) **opvatting** N(N(soort,ev,basis,zijd,stan);opvatting) **op** VZ(VZ(init);op) **twee** TW(TW(hoofd,prenom,stan);twee) **punten** N(N(soort,mv,basis);punt) **teleurstellend** ADJ(ADJ(vrij,basis,zonder);teleurstellend) **waren** WW(WW(pv,verl,mv);zijn) **.** LET(LET();.)

2.2.2 Problem areas regarding the Dutch Parallel Corpus

Although the Dutch Parallel Corpus is a high-quality corpus, there are a number of issues we would like to address as we found that they affected the results of a pilot study which we carried out. The following paragraphs go into the details of these issues which are related to the organization and the labeling of the data on the one hand and to the specific research goals of this thesis on the other hand.

Macken et al. refer to the lack of text type balance in numerous corpora and how they wanted to solve this issue for Dutch by building the DPC as a text-type balanced corpus (Macken et al., 2011). However, one of the main issues with regard to the DPC concerns the limited background information on the genres in the corpus and how they were defined. Obviously, there is some information about the sub-genres in the DPC Manual[8], where it is stated that "the labels for the subtypes were chosen from cognitively tangible categories, most of them are encountered in everyday use" (p 3). Additionally,

[8] https://www.kuleuven-kulak.be/dpc/manual/DPC.pdf

Macken et al. (2011, p. 378) mention subdividing the main genres according to the prototype approach suggested by Lee (2001). Lee, suggests for

> "genres to be treated as basic-level categories which are characterised by (provisionally) a set of seven attributes: *domain* (e.g., art, science, religion, government), *medium* (e.g., spoken, written, electronic), *content* (topics, themes), *form* (e.g., generic superstructures, à la van Dijk (1985), or other text-structural patterns), *function* (e.g., informative, persuasive, instructive), *type* (the rhetorical categories of "narrative," "argumentation," "description," and "exposition") and *language* (linguistic characteristics: register/style[?])."(2001, p. 49; see also Steen 1999)

However, no further information with regard to how the main genres were chosen and/or defined is provided. For example, the corpus metadata contain information for only three of the seven attributes mentioned by Lee, i.e. *domain, medium* and *contents* (by means of keywords). Although the two-level typology resulted in five text types the contents of which seem fairly straightforward, a thorough manual analysis revealed that the actual contents are at times somewhat problematic (see also Section 2.2.3).

An additional genre-related problem in the DPC concerns the absent or erroneous labeling of some of the sub-genres. Both Macken (2011) and the Dutch Parallel Corpus User Manual give a detailed overview of the genres and their sub-genres, or, basic-level categories, by means of a table (Macken et al., 2011, p. 378). However, based on the information in the fields <TextType></TextType> and <TextSubType></ TextSubType > in the actual corpus it was shown that there are four genres which contain texts from a sub-genre which does not belong to the genre according to the official manual and Macken et al. For example, the actual corpus showed that the genre External Communication contains three 'yearly reports' which, based on the suggested typology, should be attributed to the genre Administrative Texts. Moreover , both Administrative Texts and External Communication contain texts without sub-genre labels, which may be problematic for studies on (sub-)genre variation.

A third issue arises when we look at the metadata on translation status as these do not always provide exact information. For nearly 9% of the texts, for example, it is unknown whether the source language is Dutch, English or French. This issue is of particular importance for corpus-based research on translated versus non-translated texts as knowing the translations status of a text is essential for this type of research.

In addition to the main shortcomings mentioned above, there are a few issues with regard to the DPC which are particularly relevant for our research goals and which prevented us from using all the available material in the corpus.

As the research focus lies on investigating norm-conformity in Belgian Dutch, we could only use the Belgian Dutch sub-corpus in the DPC, a representation of which is shown in Table 8. This sub-corpus no longer contains the texts which received the label

'NL-NL' or 'NL' in the core files. The 'NL-NL' label was given to those texts which were written in or translated into Netherlandic Dutch. The 'NL' label found in the actual corpus is not mentioned in the DPC manual or Macken et al. (2011), and the texts which carry this label could therefore not be used for the research described in this thesis, as it remains unclear whether this label was used for Belgian Dutch or Netherlandic Dutch. In order to counter the (source) language related problems with the DPC, we extracted a sub-corpus with Belgian Dutch as the central language (as opposed to Netherlandic Dutch) and which only contains texts whose metadata on (source) language and translation status is complete, which resulted in a Belgian Dutch sub-corpus of 3,808,603 tokens and 2101 texts.

Table 8 Survey of the Belgian Dutch sub-corpus in DPC

	Original Belgian Dutch		Belgian Dutch translated from French		Belgian Dutch translated from English	
	N° tokens	N° texts	N° tokens	N° texts	N° tokens	N° texts
Fictional literature	0	0	116178	5	0	0
Non-fictional literature	412712	15	96688	6	0	0
Journalistic Texts	485876	356	272429	240	295039	253
Instructive texts	106640	27	45371	20	0	0
Administrative texts	428391	229	339826	177	237579	25
External communication	372256	255	261640	116	337978	377
Total	1805875	882	1132132	564	870596	655

The overview of the Belgian Dutch sub-corpus represented in Table 8 reveals an additional problem with regard to the DPC corpus concerning our main research questions. The table shows that there are no data available for non-translated Fictional Literature, for Fictional Literature translated from English, for Non-Fictional Literature translated from English and for Instructive Texts translated from English. These structural zeroes prevent us from comparing various sub-varieties with one another,

e.g. Instructive Texts that were translated from English cannot be compared with Instructive Texts that were translated from French, and therefore the influence of the source language on the linguistic differences for this particular genre cannot be investigated. Additionally, should our analyses reveal similarities or differences between Instructive Texts translated from French and non-translated Instructive Texts, it cannot be determined whether these similarities or differences are caused by the factor translation status or by the factor source language, as we cannot compare the results with a second source language variety, i.e. Instructive Texts translated from English.

Given the importance of the factor genre in this thesis on the one hand and the lack of detailed information provided by the DPC on the other, verifying how it was defined and interpreted in the DPC was of the utmost importance. Therefore, we conducted a thorough inspection of the genres in the Belgian Dutch section DPC which resulted in a number of surprising observations which are summarized in the following paragraphs.

The DPC only contains Belgian Dutch **Fiction** translated from French (Fiction<FR) and consists of only four different novels: *Gelukkige dagen, De schreeuw, De man die op reis ging* and *Een vrouweljke Odysseus, N'zid.* Although these novels fit in the genre Fiction and cause no difficulties as far as the genre definition is concerned, the fact that there is only text material available which was translated from French prevents us from investigating two of our research goals, i.e. to determine the influence of translation status on the one hand and source language on the other hand.

For Belgian Dutch **Non-Fiction**, the corpus includes only non-translated text material and text material translated from French. This genre consists of twenty-one different texts from three different publishers: *Ons Erfdeel*, *Lannoo* and *The Flemish Government.* Similar to Fiction, there are no genre -related problems here, however, the factor source language cannot be investigated for this genre either. A second issue with regard to the contents of this genre arises from the three text providers. More particularly, these providers are quite different in nature: *Ons Erfdeel* is a cultural institution whose main aim is to promote cultural cooperation between the Dutch-speaking communities. *Lannoo* is a Belgian publishing house which "wants to be a leading, creative and flexible knowledge enterprise pursuing both cultural and economic value"[9]. *The Flemish Government* is a completely different type of provider, and whereas the texts from *Lannoo* and *Ons Erfdeel* are mainly essayistic in nature, the texts from *the Flemish Government* are essentially 'expository works of a general nature' on cultural and historical topics.

The sub-corpus for Belgian Dutch **Journalistic Texts** is rather large and contains text material from six providers: *KULeuven* (a university), *De Morgen* (a newspaper), *ING* (a

[9] http://www.lannoo.be/international

bank), *Fortis* (a bank), *Roularta* (a publishing house) and *De Standaard* (a newspaper). As mentioned above the DPC Manual does not provide details with regard to the exact contents per genre and no additional information as to why certain texts were assigned to this genre could be found. Moreover, and contrary to Fiction and Non-Fiction, the text material in Journalistic Texts is highly diverse and a rigorous examination of the texts revealed some potential problem areas. Various texts from *Roularta*, for example, are clearly Instructive Texts such as step-by-step instructions on how to make curtains (e.g. dpc-rou-003283-nl-tei.xml), felt accessories (e.g. dpc-rou-003309-nl-tei.xml), cushions (dpc-rou-003339-nl-tei.xml); various recipes (e.g. dpc-rou-003288-nl-tei.xml, dpc-rou-003289-nl-tei.xml, etc.); and step-by-step instructions on how to prepare a picnic (dpc-rou-003359-nl-tei.xml). Furthermore, and similar to Non-Fiction, the texts in this genre were provided by a number of providers which are fairly different in nature, an aspect which is not unlikely to result in a less homogeneous genre.

The text material in the Belgian Dutch sub-corpus which belongs to the genre **Instructive Texts** was obtained from two text providers: the *FOD Sociale Zekerheid* and the *RIZIV*, both governmental institutions. Various texts from the *FOD Sociale Zekerheid*, the Federal Public Service of Social Security, are texts of a clearly legislative nature such as Royal Decrees (e.g. dpc-fsz-000466-nl-tei.xml); employment contracts (e.g. dpc-fsz-001959-nl-tei.xml); rules and regulations (e.g. dpc-fsz-001965-nl-tei.xml, dpc-fsz-001038-nl-tei.xml, dpc-fsz-002036-nl-tei.xml); and agreements (e.g. dpc-fsz-000540-nl-tei.xml). In addition to legislative texts, this genre contains texts of a rather administrative nature such as the written report of a workshop (dpc-fsz-000474-nl-tei.xml); an informative document with regard to job regulations for students (dpc-fsz-000469-nl-tei.xml) and assignment specifications (dpc-fsz-000438-nl-tei.xml). The texts provided by the *RIZIV,* the Belgian National Service for Medical and Disablement Insurance, are often procedural manuals (e.g. dpc-riz-001579-nl-tei.xml, dpc-riz-001479-nl-tei.xml, dpc-riz-001463-nl-tei.xml); agreements (e.g. dpc-riz-001548-nl-tei.xml, dpc-riz-001528-nl-tei.xml, dpc-riz-001505-nl-tei.xml); highly specific documents regarding the nomenclature of medical services (e.g. dpc-riz-001569-nl-tei.xml); or technical documents regarding budgetary goals (e.g. dpc-riz-001400-nl-tei.xml, dpc-riz-001406-nl-tei.xml). In other words, many of these texts might not fit the intuitive idea one has of the genre Instructive Texts, as Instructive Texts are thought to be directive and to enable the addressee to do something (see e.g. Werlich, 1982).

Investigating the texts which received the label **Administrative Texts** in the Belgian Dutch sub-corpus, showed that there are thirteen text providers which belong either to the private sector (e.g. Barco, Bekaert and Melexis) or the public sector (e.g. the Federal Public Service of Social Security, Federal Public Service of Justice, the European Parliament and the Belgian Chamber of Representatives). Remarkably, within this Belgian Dutch sub-corpus, Administrative Texts is the only genre which contains material whose outcome was "written to be spoken" or "written reproduction of spoken

language". Similar to Instructive Texts, Administrative Texts also contain technical documents regarding budgetary goals (e.g. dpc-riz-001391-nl-tei.xml, dpc-riz-001416-nl-tei.xml); documents regarding the nomenclature of medical services (e.g. dpc-riz-001417-nl-tei.xml, dpc-riz-001432-nl-tei.xml); rules and regulations (e.g. dpc-riz-001549-nl-tei.xml); and agreements (e.g. dpc-riz-001550-nl-tei.xml). As these highly specific texts occur both in Instructive Texts and in Administrative Texts, this raises the question as to whether we can consider these two genres as proper genres as their contents are clearly overlapping.

The previous paragraphs discussed a number of DPC-related issues which have profound consequences for the research presented in this thesis. In addition to some issues with regard to the aspect *language*, such as the unclear source language or empty matrix cells for specific language-genre combinations, the main shortcomings arise when we look at the aspect *genre*. For example, the DPC provides little information with regard to how the contents for the specific genres were defined. Moreover, a manual analysis revealed that in some cases the contents overlap, while in other cases they are highly heterogeneous.

2.2.3 Reorganization of the Dutch Parallel Corpus

The observations described in the previous paragraphs showed that for this particular thesis and its specific research goals a reorganization of the genres in DPC was called for. Similar to the typology suggested by Lee 2001, we wanted to divide the corpus into various genres by looking at a number of objective attributes or characteristics. Moreover, we wanted the ability to provide details on how the various genres were obtained and what the exact contents are. Therefore, we decided to divide the corpus into genres using Biber & Conrad's methodology (2009, p. 40). We selected various situational, non-linguistic characteristics for which we could annotate the texts in the corpus based on the available metadata in DPC. We opted for these situational characteristics because they allow us to redefine the genres without interfering with our research questions or influencing the answers to them. The features for which we annotated the corpus were: addressor, addressee, channel and communicative purpose.

2.2.3.1 Inter-annotator agreement

So as to be able to assess whether a set of preliminary annotation guidelines could be applied in a consistent and objective manner, an inter-annotator agreement was calculated. A set of 200 texts from 31 unique providers and 7 unique (original) DPC registers was annotated by two human annotators who annotated this sub-set, independently from one another. To calculate the inter-annotator agreement between

these two annotators, we applied Cohen's kappa statistic. The kappa coefficient, or κ, is widely used to measure pairwise agreement between two annotators who make category judgments while correcting for expected chance agreement (Carletta, 1996). κ is zero when there is no agreement between the annotators other than that which would be expected by chance alone while κ is one when there is total agreement.

For our first annotation task, an average kappa score of 0.58 was obtained, which is relatively low. As a result, we adapted the annotation guidelines so as to improve the annotation task's reliability, which was tested in a second inter-annotator agreement experiment. The second test set which was annotated consisted of 43 texts from 17 unique providers and resulted in an average kappa score of 0.86, which can be interpreted as good reliability ($\kappa > 0.8$). The results of the inter-annotator agreement on the one hand, and the analysis and discussion that followed led to more detailed annotation guidelines, so as to avoid subjective annotating as much as possible. The following paragraphs provide more details with regard to the final annotation guidelines per situational feature.

2.2.3.2 Addressor

The first characteristic, addressor, refers to the person or the institution who produced the text. As is often the case for non-literary texts in corpora, it can be difficult to determine who was the actual author (or translator) of a given text. Biber & Conrad, for example, mention the example of a university catalog, which contains the "official description of services and requirements with no indication of who produced the text" (2009, p. 40). They mention the possible influence of the addressor's age, sex, level of education, occupation and/or social class on his or her output as a potentially important determinant of linguistic variation and as "part of the larger situational context for a register" (p. 40). However, the DPC metadata do not provide such detailed information so we worked with values on a higher, more general level. Based on the metadata provided by the DPC, the possible values which were assigned to the characteristic *addressor* are*:* "(Commercial) Company", "Research and Education", "Public Service", "Media" and "Public Enterprise". The following guidelines were used to annotate the corpus for the characteristic *addressor*:

- (Commercial) Company: a private enterprise whose main purpose is to produce or to trade certain goods or services.

  For example: Inbev, a brewing company.

- Media: publications in newspapers or magazines that were published by a publishing house with commercial intent. The articles cover topics which are not related to the publishing organization.

  For example: De Standaard, a newspaper.

- Research and education: publications from educational or research institutions which discuss the institution's own activities, products or organization.
  For example: University College Ghent.
- Public service: a governmental organization which does not engage in educational or commercial activities. The texts discuss the organization's own activities, products or organization.
  For example: RIZIV, the Belgian National Service for Medical and Disablement Insurance
- Public enterprise: a governmental organization which does engage in commercial activities. The texts discuss the organization's own activities, products or organization.
  For example: De Post, a postal company.

2.2.3.3 Addressee

The second characteristic, *addressee*, is the message's intended reader or listener. The addressee can either be an individual or a group of multiple individuals. Furthermore, it cannot always be determined who these individuals are exactly. Biber and Conrad mention a novel as an example as this can "exist physically for decades and even centuries" (2009, p. 41) and it cannot be determined who exactly will read the book over time. Based on the metadata provided by DPC, it is impossible to determine the exact individual addressee per text, but the metadata allowed us to distinguish between the following labels for this particular feature: "internal target audience", "broad external audience" and "specialized external audience". The following guidelines were used to annotate the corpus for *addressee*:

- Internal target audience: in order to be able to understand the text's message, the addressee must have prior and/or expert knowledge. The information is confidential and intended for a limited audience only and is not supposed to be shared outside the organization. The available metadata are consulted to determine whether this label applies.
  For example: staff of the RIZIV, the Belgian National Service for Medical and Disablement Insurance.
- Broad external target audience: in order to be able to understand the text's message, a rather broad and/or general knowledge suffices.
  For example: newspaper readers.

- Specialized external audience: in order to be able to understand the text's message, the addressee must have prior and/or expert knowledge, but the information is not confidential.

  For example: shareholders in a company.

2.2.3.4 Channel

The third characteristic for which the corpus was annotated is *channel* or *mode* and focuses on the differences between written and spoken material. Biber and Conrad mention the fact that these differences between the spoken and the written mode often interact with other situational characteristics such as addressor (e.g. the specific addressor in the case of oral speeches) and addressee (e.g. the less identifiable target audience of, say, a blog). In addition to addressor and addressee, the purpose of a text also intertwines with its spoken or written mode: speeches, for example, are often persuasive. Biber and Conrad refer to the distinction between the written and the spoken mode as "one of the most important situational parameters for the linguistic description of registers" (2009, p. 43). In the Dutch Parallel Corpus, however, the spoken text material was not considered a separate genre and can be found in the genre Administrative Texts, which also contains written text material. The specific labels used in this thesis to annotate the DPC for the characteristic *mode* were "written to be read", "written to be spoken" and "written reproduction of spoken language". The metadata in the DPC were consulted in order to annotate the corpus for mode based on the following guidelines:

- Written to be read: the text was produced to be read.

  For example: a newspaper article.

- Written to be spoken: the text was produced to be delivered orally.

  For example: a political speech.

- Written reproduction of spoken language: the original message was delivered orally and written down afterwards.

  For example: interviews.

2.2.3.5 Communicative purpose

The final characteristic for which the corpus was annotated is the text's main *communicative purpose*. In addition to investigating the *addressor* and *addressee,* which are two characteristics from the 'participants' category, and *mode* from the 'channel' category, one should also take into account why a certain text was produced. The communicative purpose of a text, according to Biber and Conrad (2004), is a concept

which consists of various layers: one could look at the main communicative purpose of a text, but often a combination of several communicative purposes is used. Additionally, it is possible to switch from one communicative purpose to another in the middle of a text as a speech, for example, can start by merely informing the audience about a given topic and then move on to trying to persuade the audience into taking action. Within this study, the texts were annotated for their general, basic communicative purpose, viz. "Inform", "Persuade", "Instruct", "Activate" and "Inform/Persuade". The following guidelines were used to annotate the corpus for *communicative purpose*:

- Inform: for texts whose main purpose is to inform. The addressor has no personal interest in the text and is unrelated to the text and/or its message. The label 'inform' is only chosen when the text has no other clear purpose. After reading the text, the addressee knows more about its subject and the addressor has nothing to gain from this subject.
- Persuade: the text's main purpose is to convince the target audience and to change its opinion. After reading or hearing the text, the addressee's opinion might have changed, which does not necessarily imply that the addressee has the intention to actually take action because of this new state of mind.
- Instruct: the text's main purpose is to instruct and is meant to help the addressee with regard to a certain act by means of a step-by-step guide, a manual, a recipe, etc. Vague descriptions of procedures or rules and regulations are not considered to be Instructive Texts, as these are informative rather than instructive.
- Activate: in addition to being informative, the text contains an explicit or implicit call for action. After reading or hearing the message, the addressee might actually have the intention to take action, which he or she did not intend before receiving the message. Furthermore, it is in the addressor's own interest that the addressee takes action.
- Inform/Persuade: in addition to the text's main purpose to inform, there is a second dimension, which is to persuade. Indirectly, the addressor has something to gain from the message. The addressor can either gain indirect commercial success or obtain a more positive reputation. The text provides the addressee with additional background information on a given topic and, ideally, results in the addressee having a positive image of the addressor.

2.2.3.6 New genres

All texts from the Belgian Dutch sub-corpus of the Dutch Parallel Corpus were annotated for these four situational characteristics and a thorough analysis of the annotated sub-corpus was carried out, which showed that the various potential values for the situational characteristics led to fifty unique combinations of these values. Finally, the feature analysis and the detailed investigation of the annotated text files resulted in a corpus of seven genres: Specialized Communication, Broad Commercial Texts, Journalistic Texts, Instructive Texts, Political Speeches, Legal Texts and Tourist Information.

First of all, there are four genres which could be determined based on one decisive feature option which is unique for the specific genre at hand. For example, **Instructive Texts** were determined by the feature *channel* and more specifically, by the option *instruct.* Within **Specialized Communication**, the decisive feature was *addressee*, which is always either *specialized target audience* or *internal target audience*. The most important feature for **Journalistic Texts** was the *addressor*, which was either *media* or *research and education*. **Political Speeches** are defined by their mode, which is either *written to be spoken* or *written reproduction of spoken language*.

Secondly, the remaining three genres were determined by a combination of features such as, for example, **Legal Texts**, which are determined by *communicative purpose*, which is either *activate* or *persuade* and *addressor*, which is *public service*. **Broad Commercial Texts** were mainly defined by the *addressee*, which is a *broad external target audience*. The texts which belong to **Tourist Information** are determined by *communicative purpose*, which is either *activate* or *inform/persuade*, while the *addressor* for this genre is either a *public enterprise* or the *media* and the *addressee* is always *broad external.*

In addition to determining the feature values for each text, analyzing the text material in the corpus allowed for exceptions within the strict feature-based classification system. For example, Legal Texts are mainly defined by *communicative purpose* and *addressor*, which normally are activate or persuade and public service, respectively. However, the analysis of the text material revealed that dpc-ibm-001316-nl-tei.xml, for example, which is a text from a commercial company, is in fact a copy of a Parliamentary Decree. Therefore, this text was labeled as a legal text, even though the addressor was a commercial company and not a public service. It should be noted that this manual, rather intuitive approach was only adopted in case of problematic texts the contents of which did not correspond with the information provided in the metadata. An overview of the Belgian Dutch sub-corpus of the DPC and its new genres is shown below in Table 9.

Table 9 Survey of the contents per genre and (source) language variety of the reorganized Belgian Dutch sub-corpus.

	Original Belgian Dutch		Belgian Dutch translated from French		Belgian Dutch translated from English	
	N° tokens	N° texts	N° tokens	N° texts	N° tokens	N° texts
Specialized communication	296535	145	244264	97	308489	88
Broad commercial texts	236402	146	144593	101	157690	257
Journalistic texts	756459	406	374407	242	295039	253
Instructive texts	65214	31	22330	25	0	0
Political speeches	156493	51	0	0	82114	11
Legal Texts	141118	83	149055	66	779	1
Tourist Information	152999	19	81305	28	0	0

2.2.4 Data Extraction

Upon the release of the Dutch Parallel Corpus, the accompanying search interface which was developed at KU Leuven was not ready for use yet. Therefore, it was decided to build a web interface which would allow for carrying out highly detailed searches by means of a custom developed search engine. More specifically, a graphical search engine was developed which compiles the searches into regular expressions which are then matched to the corpus, whose POS-annotated XML files have been serialized into a parsable format. The custom-built DPC engine allows for different types of searches in the various sub-corpora of the DPC, viz. the Belgian Dutch sub-corpus, the French sub-corpus and the English sub-corpus. Figure 3 shows an example of a possible query which can be carried out and more information with regard to the various fields of the search engine is provided below.

The DPC engine provides the following search options:

LANGUAGE

One can choose between querying the following sub-corpora:

EN-UK: United Kingdom English, for texts which were produced in the United Kingdom;

EN-US: United States English, for texts which were produced in the US of A;

FR-FR: French French, for texts which were produced in France;

FR-BE: Belgian French, for texts which were produced in Belgium;

NL-NL: Netherlandic Dutch, for texts which were produced in the Netherlands;
NL-BE: Belgian Dutch, for texts which were produced in Belgium;
NL: Dutch, for texts for which their origins were unknown.

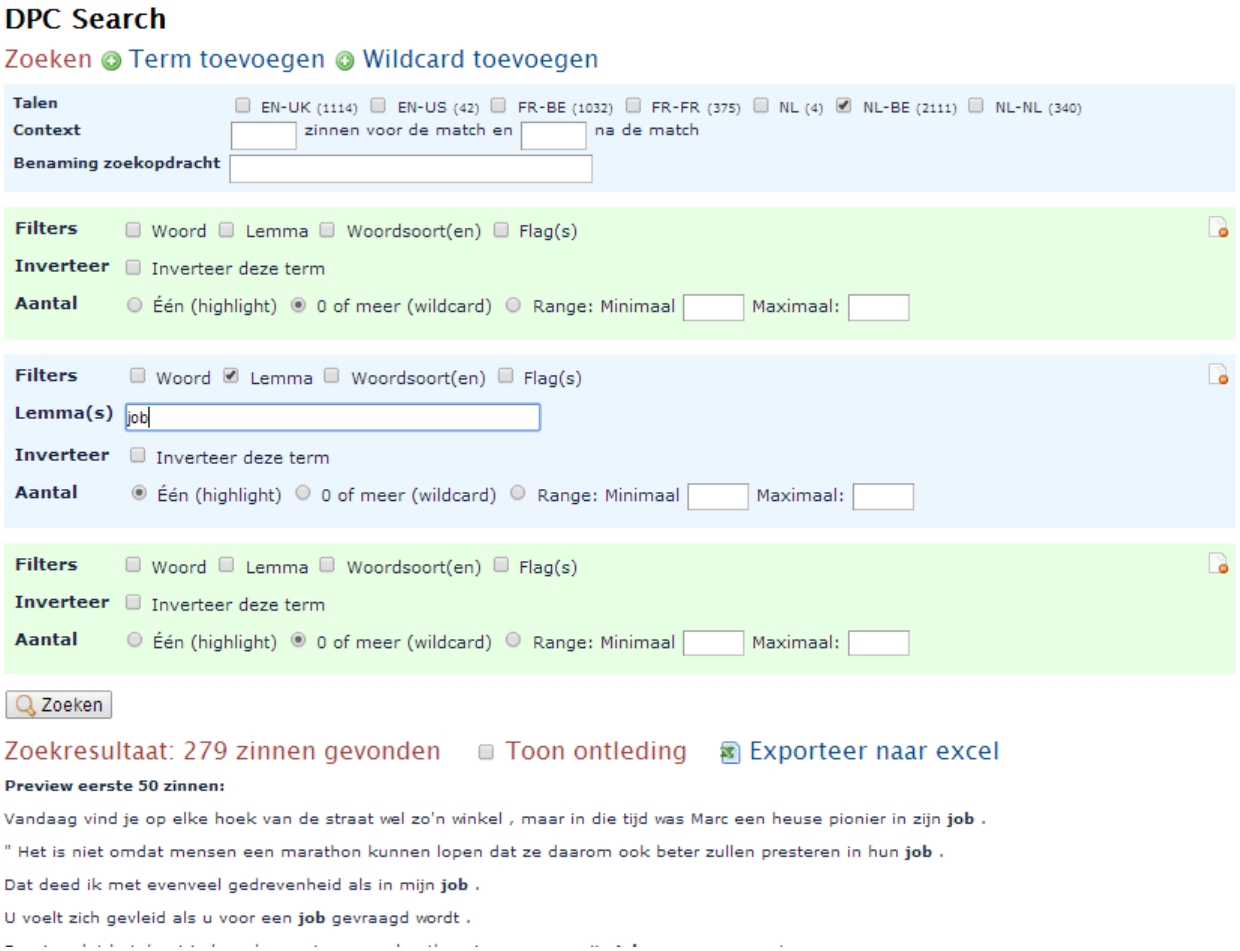

Figure 3 A query example which is carried out by means of the DPC search engine which was custom-built for the COMURE research project[10].

WORD FORM

One can search the corpus for sentences which contain a specific word, or a combination of specific words. This can be done either by explicitly entering which word one wants to find, for example "working", or by entering a combination of words one wants to find, for example "working+conditions". Additionally, the DPC engine allows for detailed searches which are based on the use of regular expressions. Applying regular expressions when querying the corpus allows for searches which cover a wide

[10] http://dpcserv.ugent.be/comure

variety of options at once. For example, say we were to query the corpus searching for sentences which contain any of the following options: “working out”; “working with”; or “working on”. This could be done by looking for the word form “working” in combination with the regular expression “^(out|with|on)$”.

LEMMA

In addition to querying the corpus based on a specific word form, the DPC engine allows for searches which are based on the linguistic annotations available in the corpus. More specifically, one can query the corpus by applying a lemma-based search. For example, say we were looking for sentences which contain either “work”, “works”, “worked” or “working”, this could be done by applying the regular expression as mentioned above. However, this can also be done by means of using the lemma as a query. More specifically, if one queries the corpus for “work” as lemma, all sentences which contain any of the four forms mentioned above will be selected. Using the lemma option is especially useful when looking for lexical items for which one wants to retrieve both the singular and the plural form, or, as shown in the example, when looking for all possible forms of a specific verb.

PART OF SPEECH

The corpus cannot only be consulted by searching for specific word forms or lemmas, one can also insert a query based on the morphosyntactic class of a given word. For example, “working” can either be a noun, e.g. the inner workings of a system, a verb, e.g. the system is working, or an adjective, e.g. a working system. Adding information with regard to the word class of the specific phenomenon one is looking for, can greatly reduce the amount of noise in the results. The DPC engine allows for detailed part of speech-based searches, the specific categories for which depend on the chosen sub-corpus (English, French or Dutch) as these have been pre-processed by means of different tools and the POS information is therefore not similar for the three languages.

ATTRIBUTES

In addition to word class, the part-of-speech tagging process provides more detailed information as well and this information can also be used when querying the corpus by means of the DPC search engine. For example, when looking for a noun, one can specify whether this noun should be singular or plural, a proper name, etc. The DPC search engine allows for adding this additional information in the field ‘flag(s)’, which leads to more detailed searches. The information which can be added to the field ‘flag(s)’ depends on which sub-corpus one wants to query (Dutch, French or English).

FREQUENCY

The DPC engine was built in such a way that for each chosen search option, one can determine how often it should occur in order to be a valid attestation. For example, say we are interested in retrieving every sentence which contains the word "working" from the corpus. We can then determine whether we want sentences in which "working" occurs exactly once, or between a given range, or not at all. The number of occurrences can be specified for each search option, e.g. lemma, word form, part of speech. Furthermore, the engine allows for negative searches as well as one can query the corpus for all sentences which do not contain a given word form. The option to extract sentences which do not contain a given word form, lemma or part of speech is especially useful when used in combination with other search options that should occur. For example, for the extraction of sentences which contain a verb which is not preceded by another verb, it is useful to use this negative search option.

It should be noted that the query results of the DPC engine depend on the quality of the preprocessing step which was applied to the Dutch Parallel Corpus and which added the information with regard to lemmas, part of speech classes, etc. More details with regard to the accuracy scores of this process and the consequences are provided in Section 2.2.1.3 above.

2.3 Manual validation

The third phase in our methodology involves manual validation of the extracted data, a time-consuming step in the process which was carried out by means of a single, strict annotation guideline. Section 2.3.1 below motivates the choice for profile-based correspondence analysis in this thesis, a technique requiring the use of variables which are grouped into linguistic profiles. These profiles consist of a series of variants for a certain linguistic function (Speelman, Grondelaers, & Geeraerts, 2003) and the annotation guideline states that only those attestations which contain a variant which can be replaced by the other variant(s) in the profile are included in the data set. In other words, if one variant cannot be replaced by the other variant(s) in the profile, the attestation is discarded. This can be due to a difference in meaning, figurative use of a given word, fixed terminology, etc. Below, some examples are provided of attestations which had to be discarded.

(4) Profile *baan-job* (CONCEPT: JOB)

Manual	Sentence in Dutch	English counterpart	Remarks

validation			
☒	Zolang het probleem Irak niet definitief van de **baan** is [...].	As long as the Iraqi problem has not been resolved [...]	*baan* cannot be replaced by *job,* because *baan* has a different meaning here
☑	In de afdeling commerciële vliegtuigen heeft hij duizenden **banen** geschrapt [...].	[...] where he cut thousands of **jobs** from the commercial aircraft division [...]	*banen* can be replaced by *jobs*

(5) Profile *uiterlijk-ten laatste* (CONCEPT: AT THE LATEST)

Manual validation	Sentence in Dutch	English counterpart	Remarks
☒	Ik moet nu meer op mijn **uiterlijk** letten.	I need to look after my **appearance** more [...].	*uiterlijk* cannot be replaced by *ten laatste* because *uiterlijk* has a different meaning
☑	Als u **uiterlijk** op 31 december langsgaat bij uw BBL-kantoor [...].	If you go to your BBL branch **by the latest** on 31 December [...].	*uiterlijk* can be replaced by *ten laatste*

In addition to discarding attestations due to the annotation guideline described above, we set a threshold with regard to two aspects. First, the total sum of the occurrences of all variants per profile should be at least 35 across all factors. For example, for Case Study III which investigates the dispersion of General Standard Dutch versus non-standard Dutch, *op het eerste zicht* vs. *op het eerste gezicht* (CONCEPT: AT FIRST SIGHT) was a profile candidate. However, across all factors, the total sum of all occurrences only amounted to 29. Therefore, this profile candidate could not be retained.

Secondly, the total sum of the occurrences of all variants should be at least 100 per factor. For example, for Case Study II which investigates the dispersion of English loanwords versus more endogenous lexemes, the total sum of all occurrences only amounted to 63 for the genre Tourist Information and this genre was therefore not

added to the analysis. As a result of this second threshold, the number of genres varies per case study.

2.3.1 Measuring linguistic distances

As the linguistic choices made in both translated and non-translated language are affected by a wide variety of factors, a multifactorial approach is needed. Or, as Gries put it: "multifactorial data must be analyzed multifactorially [...] as the complexities of linguistic data do not reveal themselves easily to the naked or to the monofactorial eye" (2010, p. 143). In addition to investigating various factors at once, the aim of the research presented in the case studies throughout this thesis was to aggregate the linguistic profiles under investigation instead of analyzing them one by one. Consequently, we wanted to be able to generate results which are more broadly applicable and which are not based on the individual behavior of a single variable.

2.3.2 Correspondence analysis

Advanced statistical methods are being applied more and more in corpus-based translation studies and the application of multivariate techniques in this research domain has proved to lead to revealing results (e.g. De Sutter, Delaere, & Plevoets, 2012; Diwersy et al., 2014; Gries, 2010; Rybicki, 2012) and is therefore greatly encouraged (see e.g. De Sutter, Goethals, Leuschner, & Vandepitte, 2012; Oakes & Ji, 2012).

As, among other aspects, we want to verify whether a given variable is used more often in one genre in comparison with another genre, we need an analytical tool to investigate whether the data points in our frequency table behave homogeneously or not such as, for example, the chi-square test. Unfortunately, this test can only determine whether (or not) the observed frequency distribution deviates from the assumed or the theoretical independent distribution and it typically does not say anything about the dependencies within the contingency table. Correspondence analysis, a technique highly suitable for the analysis of linguistic frequency data, can offer a solution to this problem as it uses the chi-square deviations between the cells (Greenacre, 2007). It results in a spatial representation of the data, which is one of the strong points of correspondence analysis, as the analysis's output "is characterized by the fact that it projects both the column and the row points into the same subspace (biplot), thus allowing us to inspect the position of both observations and variables at the same time" (Jenset & McGillivray, 2012, p. 317).

The contingency tables, or matrices, which are often used in corpus-based studies such as the one presented here can thus be interpreted as large geometric spaces with multiple dimensions. Overall, these matrices represent the frequencies of certain

features (the rows) over a number of factors (the columns) and these matrix cells can be considered coordinates of the features, which become data points in the geometric space which has as many dimensions as there are columns (and thus factors) in the matrix (Jenset & McGillivray, 2012). Moreover, the distances between these data points can provide information with regard to the similarities or differences between them. The visualized results based on our datasets can show, for example, how one genre behaves with regard to a given variable in comparison to another genre.

Although these multidimensional spaces contain valuable information with regard to the data, it is impossible to visualize any space which contains more than three dimensions on a screen or on paper. However, dimension reduction techniques can be applied so as to reduce the multidimensional space to a two-dimensional space, which can be plotted and which allows for an easier analysis and thus interpretation of the differences or similarities in a dataset. As dimension reduction technique Singular Value Decomposition was applied, a standard technique which reduces the original, multidimensional space to a new, two-dimensional space while as much as possible of the original variation is retained (Strang, 2009). Figure 4 shows a simplified representation of how (i) in an original three-dimensional space (ii) a plane is fitted so (iii) the average distance of the data points to the plane is as small as possible, which results in (iv) a two-dimensional plot which is the best possible representation of the original variation.

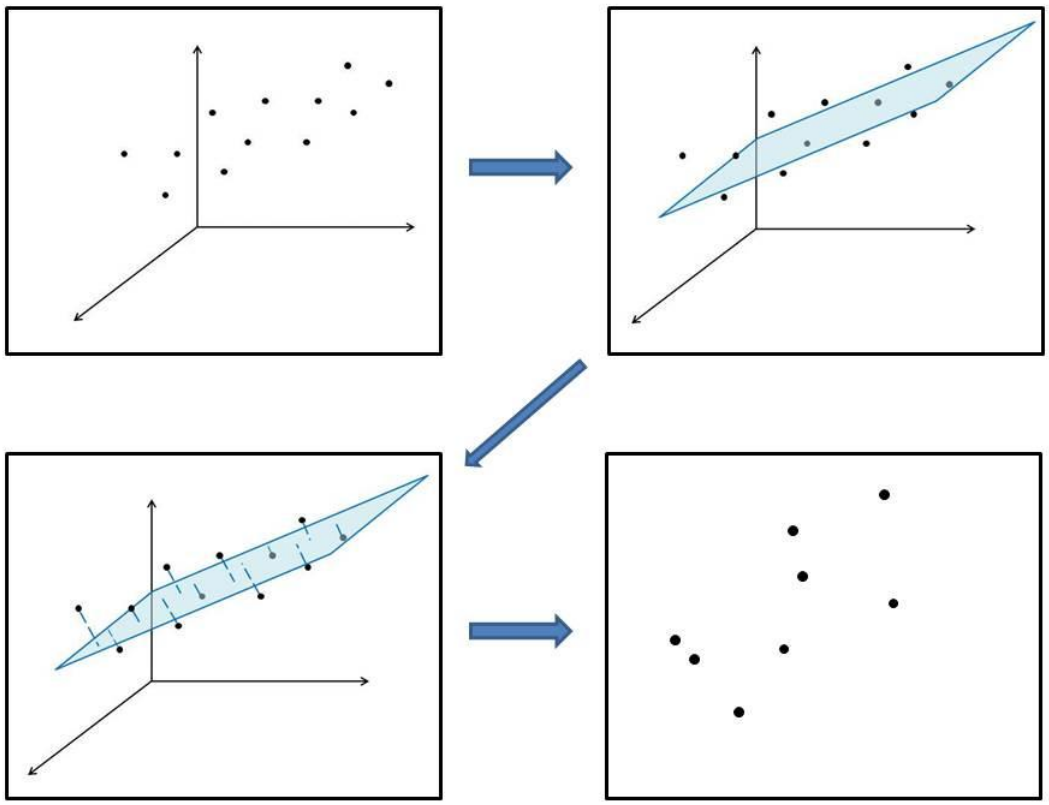

Figure 4 Schematic representation of the reduction of a three-dimensional space to a two-dimensional space, or plot.

However, calculating and visualizing the distances between the language varieties in our corpus is but the first phase as it is equally important to assess whether these distances are statistically significant or not. This is done by computing confidence ellipses for each point in the space, which give the regions of uncertainty around every point (by consequence, confidence ellipses are the two-dimensional analogues of confidence intervals) (Reiczigel, 1996). For each analysis, the confidence level was set at 95%, which means that under resampling 95% of all confidence ellipses will contain the true population position of the language variety in question. As a consequence, the distance between two factors (e.g. 'Instructive Texts' or 'Dutch translated from French') can be considered statistically significant ($p < .05$) if their ellipses do not overlap. The use of biplots allows for an exploratory analysis of how the variables are used throughout the various factors in the dataset. Moreover, it should be noted that although correspondence analysis is indeed a technique which is highly suitable for exploratory data analysis and can therefore be used to find associations and patterns in the data, making explanatory conclusions based on the results of this technique should be avoided.

2.3.3 Profile-based Correspondence analysis

For the research presented in this thesis we carried out an extended version of correspondence analysis, i.e. Profile-Based Correspondence Analysis as this has the additional advantage that the distance measurements are only sensitive to the different linguistic choices being made and not to topic or meaning. The idea of a profile-based uniformity measure was first developed by Geeraerts et al. (1999) and further developed in Speelman et al. (2003). This technique requires the use of a dataset which consists of various linguistic profiles, viz. sets of synonymous onomasiological variants that can be used to express a given concept, e.g. *car* and *automobile* (Speelman et al., 2003). One of the main advantages of the technique is described in Ruette (2012, p. 198):

> "Whereas typical (distance-based) aggregation methods regard all the individual words across the variables as the dimensions of one hyperdimensional space, the profile-based approach creates a space for every onomasiological variable. So, if we have five lexical variables with each three variants, a regular aggregation technique would interpret this as one fifteen-dimensional space (five times three), whereas the profile-based approach will consider this to be five spaces of three dimensions. The distances per variable (or space) are then merely averaged to get the aggregated distance. The reduced sensitivity to conceptual variation generates an improved sensitivity to the relevant variational dimensions."

One of the advantages of this profile-based approach is that the obtained results, or distances, are not a consequence of the topic of the (sub)corpora because the distances between the language varieties are calculated based on the frequency differences between formal variants, and not between meanings. For example, suppose we were to compare the use of General Standard Dutch versus Belgian Standard Dutch in Legal Texts versus Journalistic Texts. Hypothetically, we could investigate this matter by measuring the frequencies of a number of lexical items which are labeled as General Standard Dutch (GSD), see Table 10.

Table 10 Absolute frequencies of three lexical items in two genres.

Variable	**Concept**	**Legal**	**Journal**
Werkgelegenheid (GSD)	Employment	79	11
Uiterlijk (GSD)	At the latest	199	41
Telkens als (GSD)	Each time	11	3

According to this table, the General Standard Dutch variants *werkgelegenheid*, *uiterlijk* and *telkens als* occur more frequently in Legal Texts than in Journalistic Texts, thus one could say that in Legal Texts, General Standard Dutch is used more often than in Journalistic Texts. However, these absolute frequencies could also be a result of the fact that, in the corpus, the sub-corpus of Legal Texts is twice as large as the sub-corpus of Journalistic Texts. In order to eliminate the corpus design's influence on our results, normalized frequencies are needed (see Table 11).

Table 11 Normalized frequencies per 10,000 words for three lexical items in genres.

Variable	**Concept**	**Legal**	**Journal**
Werkgelegenheid (GSD)	Employment	0.79	0.22
Uiterlijk (GSD)	At the latest	1.98	0.80
Telkens als (GSD)	Each time	0.11	0.06

Looking at just the normalized frequencies in Table 11 for *werkgelegenheid*, which is a General Standard Dutch lexeme referring to the concept EMPLOYMENT, the higher frequency of this General Standard Dutch form in Legal Texts could either mean that (i) Legal Texts are a genre which prefers General Standard Dutch or (ii) this genre contains many EMPLOYMENT-related texts. Because these topic-related differences could have an influence on the differences between the relative frequencies of the alternatives, we cannot use this kind of observations to say something about differences or similarities in normative behavior.

Therefore, we group the alternatives in so-called "profiles", and the relative frequencies per profile sum to one. An illustration is given in Table 12.

Table 12 Profile-based, normalized frequencies per 10,000 words for three profiles in two genres.

Variable	**Concept**	**Legal**	**Journal**
Werkgelegenheid (GSD)	Employment	**0.73**	0.46
Tewerkstelling (BSD)	Employment	0.27	**0.54**
Uiterlijk (GSD)	At the latest	**0.80**	0.39
Ten laatste (BSD)	At the latest	0.20	**0.61**

Telkens als (GSD)	Each time	**0.92**	0.38
Telkens (BSD)	Each time	0.08	**0.62**

On the basis of these profiles, we can easily observe that within Legal Texts, the GSD options occur more often than the BSD options while within Journalistic Texts the BSD option is preferred. It should be noted that with this technique the observed distances are less dependent (i) on the topics involved or (ii) on the specific linguistic variables. First of all, as lexical variation is based on the idea that language users can choose between two or more formal alternatives to express a given concept, profile-based correspondence analysis is very well suited to investigate these linguistic preferences as the analysis is based on the difference in use between these formal alternatives and not the concepts behind them. Second, the technique leads to conclusions which are valid on a general level due to the objectivity of the statistical method and the substantial number of variables which can be used in each analysis, which are analyzed together instead of separately. This approach of aggregating the variables leads to results which are not based on the behavior of one single variable, for which we cannot know whether it is representative for the behavior of the other variables as well.

The tables in the example above merely illustrate how the profile-based approach applies profile-based, normalized frequencies, however, profile-based correspondence analysis makes use of profile-based chi-square deviations, as these are the basis for calculating the linguistic distances between the various language varieties.

Apart from the use of profiles, profile-based correspondence analysis also leads to mapping all distances in a lower-dimensional space, which captures the original distances as accurately as possible. This makes (profile-based) correspondence analysis similar to the Vector Space Model in information retrieval (Salton, Wong, & Yang, 1975) and/or Latent Semantic Indexing (Landauer & Dumais, 1997), where a high-dimensional space of terms and documents is reduced to a few, easy-to-handle dimensions. The difference is that (profile-based) correspondence analysis makes explicit use of the chi-square distance, as outlined above. These distances are then represented in a two-dimensional plot, for the sole practical reason that they can be graphically visualized, which results in a so-called bi-plot.

In conclusion, profile-based correspondence analysis is a multidimensional approach dealing with multiple language-internal and language-external variables simultaneously which makes it possible to visualize and measure the degree to which translated and non-translated texts conform to linguistic norms.

Chapter 3
Case Study I: visualizing formal versus neutral lexemes in search of descriptive norm conformity

3.1 Hypothesis and variable selection

3.1.1 Hypothesis

Case Study I is carried out to investigate to what degree translations display descriptive, genre-determined norm conformity by investigating the dispersion of formal lexemes versus more neutral lexemes throughout the Dutch Parallel Corpus. By doing so, we want to verify whether a given genre in translation, e.g. translated Journalistic Texts, conforms to or deviates from the prevalent norm in its corresponding non-translated genre. For example, if non-translated Journalistic Texts demonstrate a preference for neutral lexemes, we want to verify whether translated Journalistic Texts conform to this specific genre-determined norm and show a similar preference for neutral lexemes or whether they deviate from the norm by showing a preference for formal lexemes. The sub-hypothesis for this particular case study predicts the following:

> *"Genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to formal versus neutral lexemes as their corresponding non-translated genres."*

3.1.2 Variable selection

In order to investigate the norm-conformity hypothesis with regard to genre-determined formality, we ultimately selected eight profiles (16 variants) of lexical items that consist of at least one formal and one neutral alternative. As is the case for each profile used throughout this thesis, the linguistic alternatives need to be

interchangeable and thus replaceable by one another at all times. We consulted[1] the following four sources to build a profile set: (i) Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013), (ii) *De Taalgids,* a reference guide on language usage (van der Horst, 1999), (iii) *VRT Taal*[2], a database on language usage issued by the Flemish public broadcasting company, and (iv) *Taaladvies*[3], a database on language usage issued by the Dutch Language Union. The profile set was built based on the following selection criteria:

(i) at least one of these four recent sources, viz. dictionaries or language advice literature, recognizes one variant to be formal and the other to be neutral;

(ii) there is no conflicting information to be found in the remaining sources;
For example, if a candidate term received the label "formal" in *De Taalgids*, but was labeled "neutral" according to *Taaladvies*, the candidate was rejected from the profile set. Or, if a candidate (e.g. *alsmede*) was labeled "formal" in Van Dale, but according to *Taaladvies* this formal alternative is not entirely interchangeable with the neutral alternative (e.g. *evenals*) due to a certain difference in meaning, the profile could not be included in the profile set.

(iii) the only difference between the variants within one profile is a difference in formality and not, for example a regional difference.
For example, if a candidate was labeled "formal" in Van Dale, but according to *Taaladvies* this candidate was labeled "Belgian Standard Dutch", the candidate could not be included in the profile set.

In a next phase, all instances for these profile candidates were extracted from the corpus so as to verify whether the formal variants could be replaced by their neutral counterparts given the specific context they were used in (and vice versa). This manual validation step in combination with the strict profile selection guidelines above and a

[1] Last consulted on January 31 2014

[2] http://www.vrt.be/taal/taaldatabank

[3] http://taaladvies.net/

minimal frequency measure per profile and per factor resulted in a final set of 8 profiles, all of which consist of one formal variant and one neutral variant. It should be noted that, where necessary, both the singular and the plural forms of the variants were taken into account and all corpus searches were carried out in a case-insensitive manner. Table 13 provides an overview of the final profile set (n= 26015) for this case study.

Table 13 Frequencies of the formal and the neutral alternatives per genre and (source) language variety.

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL	TOURIST	INSTRUCT		DU_orig	DU<EN	DU<FR		TOTAL
maar	NEUTRAL	BUT	143	971	814	980	5778	487	154		5033	2081	2213		9327
echter	FORMAL		72	310	237	114	676	35	8		716	328	408		1452
proberen	NEUTRAL	TO TRY	6	24	32	69	378	26	17		303	137	112		552
trachten	FORMAL		6	24	17	41	39	6	0		75	26	32		133
nu	NEUTRAL	NOW	36	294	392	493	1528	109	50		1452	924	526		2902
thans	FORMAL		12	92	20	30	15	1	0		36	30	104		170
als	NEUTRAL	IF	49	101	77	35	342	28	88		419	173	128		720
indien	FORMAL		41	75	9	16	9	2	11		73	28	62		163
in	NEUTRAL	IN	38	385	167	535	1730	321	32		1691	1062	455		3208
te	FORMAL		16	75	4	43	91	35	0		137	52	75		264
al	NEUTRAL	ALREADY	80	380	316	380	2323	196	99		1982	860	932		3774
reeds	FORMAL		67	251	87	181	162	12	18		436	186	156		778
moeten + inf	NEUTRAL	TO HAVE TO	115	238	182	200	465	15	23		511	252	475		1238
dienen + inf	FORMAL		209	231	131	41	128	6	80		545	124	157		826
numeral + *keer*	NEUTRAL	TIMES	23	58	15	34	257	11	13		200	93	118		411
numeral + *maal*	FORMAL		13	44	6	10	22	1	1		49	12	36		97

In the following paragraphs, a description of the individual profiles and an example is provided. The Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013) was used for the concept descriptions. The example sentences were extracted from the Dutch Parallel Corpus and their origin is presented by means of the file number. Each Dutch example sentence contains one of the profile's variants which is matched by its translation or source text sentence in English.

Maar – echter

Concept: BUT - "adversative conjunction"
Word class: conjunction
Example: DU: [...] er zal **echter** steeds behoefte blijven aan [...].
EN: [...]**But** there will always be a need for [...].
dpc-bco-002536-nl

Proberen – trachten

Concept: TO TRY - “to attempt to achieve something”
Word class: verb
Example: DU: [...] **probeert** een aantal lidstaten stiekem [...].
EN: [...] some Member States have been quietly **attempting** to [...].
dpc-erp-000444-nl

Nu – thans

Concept: NOW - “the present moment”
Word class: adverb
Example: DU: [...] **nu** als museum te bezichtigen [...].
EN: [...] a castle that may **now** be visited as a museum [...].
dpc-ons-000476-nl

Als – indien

Concept: IF (conditional) - "subordinate conjunction of condition"
Word class: conjunction
Example: DU: **Als** ik deel zou uitmaken van de strijdkrachten die in Afghanistan dienen [...].
EN: [...]**if** I were a member of the forces serving currently in Afghanistan [...].
dpc-erp-000448-nl

In – te

Concept: IN - “preposition of place”
Word class: preposition
Example: DU: De overeenkomst werd op 9 oktober plechtig ondertekend **in** Montevideo [...].
EN: The signing ceremony of the contract took place on 9 October **in** Montevideo [...].
dpc-bco-002536-nl

Al – reeds

Concept: ALREADY - “earlier than one might have expected”
Word class: adverb
Example: DU: In dat opzicht hebben we **al** goed nieuws te melden.
EN: In this respect we **already** have some good news.
dpc-bco-002536-nl

Moeten + infinitief – dienen + infinitief

Concept: TO HAVE TO (obligation) - “being necessary”
Word class: verb
Example: DU: Europa **moet** dan ook alles op alles zetten om [...].
EN: Europe **should** therefore pull out all the stops to [...].
dpc-erp-000399-nl

Numeral + keer – numeral + maal
Concept: TIMES - “occasion”
Word class: cardinal number
Example: DU: [...]is het mogelijk om vliegtuigen te produceren die zes **maal** minder lawaai produceren [...].
EN: [...] it is possible to produce aircraft which produce six **times** less noise [...].
dpc-erp-000401-nl

3.2 Results

Although the main purpose of this case study is to verify whether translated genres conform to the prevalent norm of their corresponding non-translated genres, we will first discuss the global effects of genre and translation in Section 3.2.1, which will reveal what the linguistic distances are between the various genres and (source) language varieties in our corpus and which allow us to paint the bigger picture with regard to the various factors in our dataset.

Secondly, we will focus on the interaction results (Section 3.2.1.1), which show how the factor genre and source language variety interact with one another and which are crucial to this first case study as they provide information with regard to the differences or similarities between translated and the non-translated genres.

3.2.1 Global effects of genre and translation

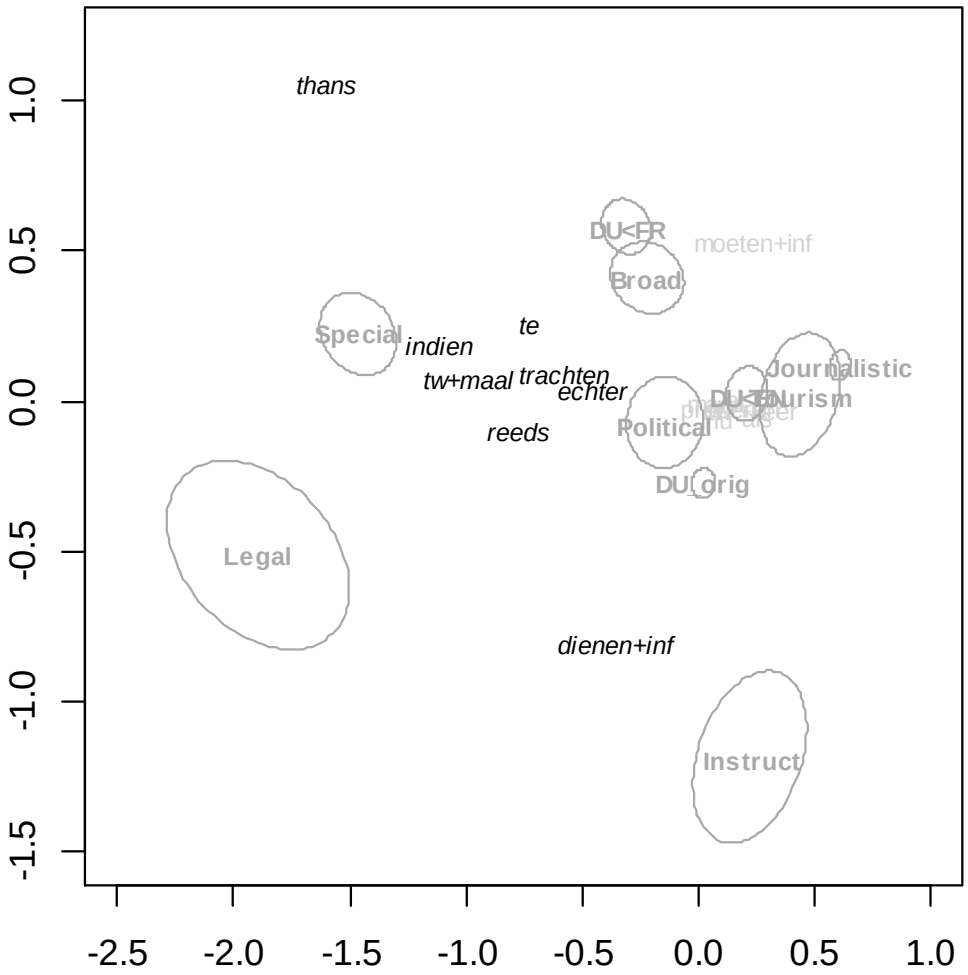

Figure 5 Case Study I: Biplot of the global results with the formal lexemes (black italics), the neutral lexemes (light grey), the genres and the (source) language varieties.

Figure 5 shows the results of the profile-based correspondence analysis that was carried out on our dataset. It is a visual representation of the distances between (i) the linguistic variants, viz. the formal (black italics) versus the neutral (light grey) alternatives ; (ii) the genres, viz. Legal Texts, Specialized Communication, Political Speeches, Broad Commercial Texts, Journalistic Texts, Tourist Information, and Instructive Texts; and (iii) the (source) language varieties, viz. non-translated Belgian Dutch, Belgian Dutch translated from French and Belgian Dutch translated from English. This biplot visualizes the global effects, which means it contains information with regard to every factor under investigation. As there are 10 possible varieties in our dataset, 10 ellipses can be seen in Figure 5, which makes it rather complex.

If we look at the linguistic profiles in this plot two clusters can easily be identified. The formal lexemes, which are visualized as black italic data points, are situated

towards the left side of the plot, whereas their neutral counterparts, which are visualized as light grey data points, are situated towards the right side of the plot. Furthermore, this plot reveals that the neutral lexemes behave more homogeneously than the formal lexemes, as the dispersion of the former is remarkable smaller than that of the latter.

Figure 5 clearly shows that there are significant distances between all three source language varieties, as their confidence ellipses do not overlap. More particularly, this plot visualizes that (i) there is a difference between translated (DU<FR and DU<EN) and non-translated texts (DU_orig), whereas (ii) translated texts do not behave uniformly, as the ellipses for translations from English (DU<EN) and French (DU<FR) do not overlap. Furthermore, the position of the various genres and (source) language varieties with respect to the data points reveal that both DU<EN and DU_orig make more use of neutral alternatives, as the distance to these lexemes is smaller. DU<FR on the other hand is situated closer to the formal lexemes, which shows there is a tendency to make more use of these formal alternatives.

These results show that source language has an influence on whether formal or neutral variants are used more frequently in translated language; and the preference for formal language in texts translated from French cannot merely be assigned to the assumed conservative behavior of translators as such, as translators who translated from English appear to prefer the neutral variants (for more details with regard to the conservatism hypothesis, see 1.4.1).

In addition to the main effects with regard to translation status, Figure 5 also shows the results for the various genres in the dataset, viz. Legal Texts (Legal), Specialized Communication (Special), Broad Commercial Texts (Broad), Journalistic Texts (Journalistic), Tourist Information (Tourism), Political Speeches (Political) and Instructive Texts (Instruct). As can be seen in the biplot, the linguistic distances between all genres are significant, except for the distance between Journalistic Texts and Tourist Information. As mentioned above there is a left-right formality division along the X-axis, with the formal lexemes situated towards the left-hand side and the neutral lexemes towards the right-hand side.

If we look at the position of the genres with regard to this left-right division, we can see that Legal Texts and Specialized Communication appear to make more use of the formal options in comparison the other genres, whereas Journalistic Texts, Political Speeches and Tourist Information on the other hand appear to make more use of the neutral alternatives. Finally, Broad Commercial Texts and Instructive Texts reveal no clear preference for either formal or neutral lexemes. The position of Instructive Texts, for example, can be explained though the fact that, in general, Instructive Texts mostly prefer neutral lexemes, but where the profile *dienen+inf - moeten+ inf* (concept: TO HAVE

TO) is concerned, the genre makes substantially more use of the formal *dienen+inf* in comparison with the other genres.

These observations with regard to the various genres and their preference for either formal or neutral language are not entirely surprising. Legal Texts and Specialized Communication, for example, are genres which tend to be formal in comparison with Journalistic Texts and Tourist Information, for example, which are more neutral genres (see e.g. Böttner, 1995; Permentier, 2003). Overall, the genre-related results clearly show that, as a factor, genre has an impact on whether a formal or a neutral lexeme is used.

3.2.1.1 Interaction effects between genre and translation

The observations which were described and discussed in the previous sections resulted from analyzing the main effects only. However, these results did not allow us to investigate genre-determined norm conformity, i.e. the extent to which a given genre in translation behaves similar to or different from that same non-translated genre. Therefore, we calculated the interaction effects between genre and source language which, moreover, allowed us to investigate the influence of an additional factor, i.e. source language by comparing Legal Texts translated from French with Legal Texts translated from English.

It should be noted that these sub-plots are based on the main analysis, but only give a visualization of the data for one genre at a time, as plots with a lower information density allow us to interpret the results more easily. Although these sub-plots were scaled differently for representational reasons it should be noted that the in-between distances between the data points etc. are proportionally the same. In other words, the original positions of the linguistic variants on the one hand and the various language varieties on the other hand and hence their in-between distances are not altered.

3.2.1.2 Broad Commercial Texts

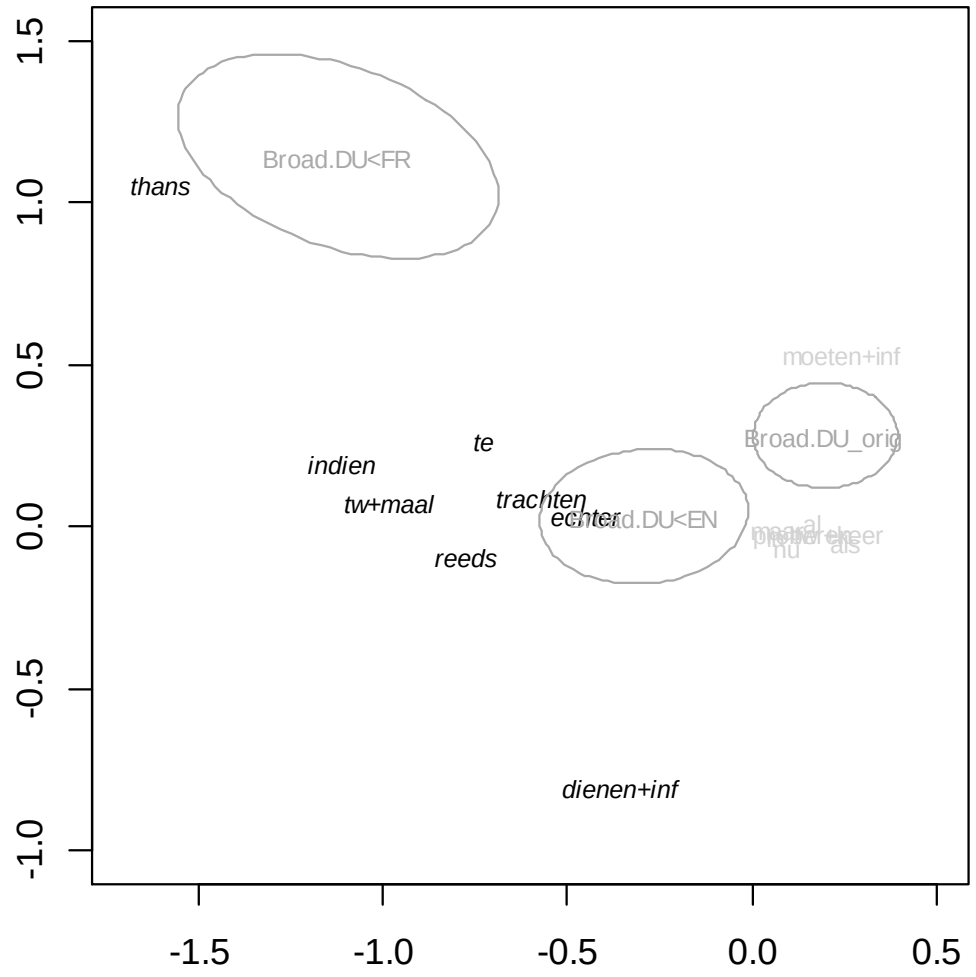

Figure 6 Case Study I: Biplot of the interaction between Broad Commercial Texts; and source language and translation status.

Figure 6 shows the interaction effect between Broad Commercial Texts on the one hand and source language and translation status on the other hand. The plot reveals significant distances between all three sub-varieties, viz. Broad Commercial Texts translated from French (`Broad.DU<FR`), Broad Commercial Texts translated from English (`Broad.DU<EN`) and non-translated Broad Commercial Texts (`Broad.DU_orig`).

These results show that translated Broad Commercial Texts do not conform to the norm established in `Broad.DU_orig`, which show a clear preference for neutral lexemes. More particularly, Figure 6 shows that `Broad.DU<FR` show a preference for formal lexemes and, more particularly, for the lexeme *thans* (concept: NOW) in comparison with the other sub-varieties of this genre. `Broad.DU<EN` on the other

hand appears to have no clear preference for either formal or neutral lexemes and wavers somewhere in between the two clusters.

Summarizing the interaction effects for Broad Commercial Texts revealed that this genre does not behave uniformly. Not only do translated Broad Commercial Texts not conform to the norm established by non-translated Broad Commercial Texts, the source language seems to have a strong effect on the preference for either formal or neutral lexemes.

3.2.1.3 Specialized Communication

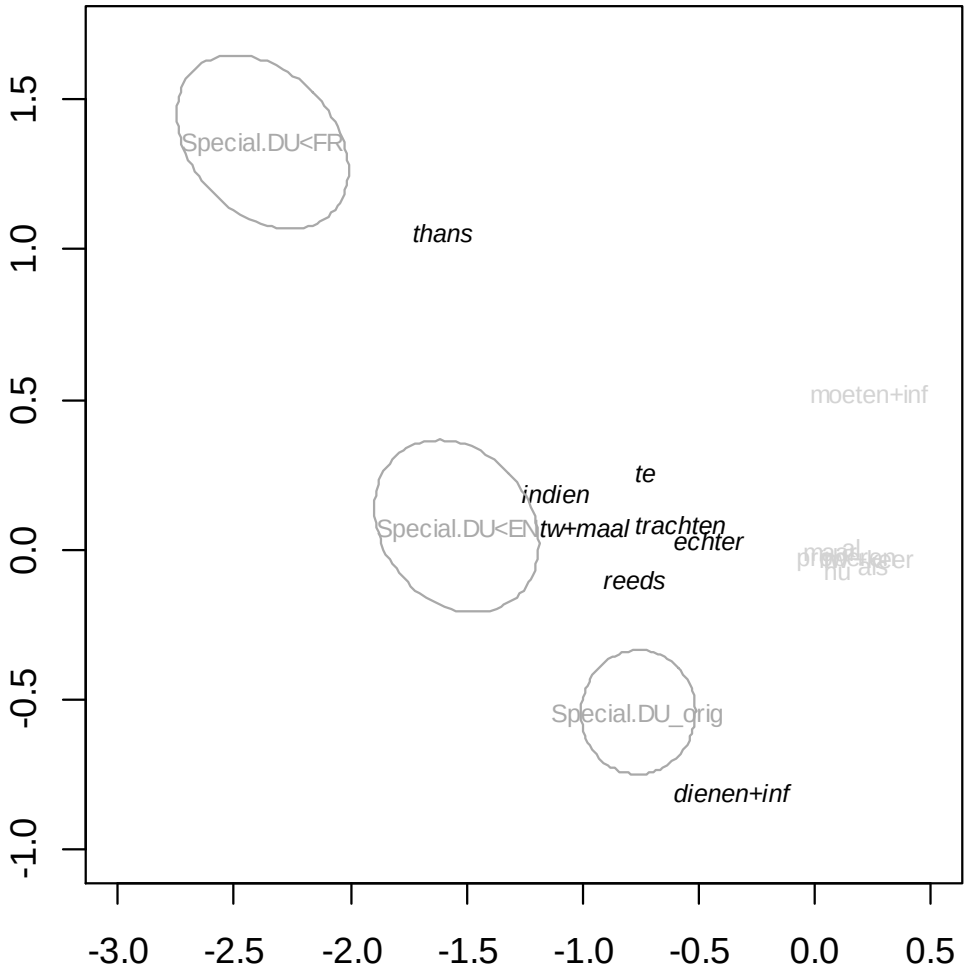

Figure 7 Case Study I: Biplot of the interaction between Specialized Communication; and source language and translation status

The biplot in Figure 7 shows that there are significant differences between all three sub-varieties of Specialized Communication, viz. Specialized Communication translated

from French (Special.DU<FR), Specialized Communication translated from English (Special.DU<EN) and non-translated Specialized Communication (Special.DU_orig). Although all three sub-varieties seem to make more use of formal lexemes instead of neutral lexemes, which indicates that both Special.DU<FR and Special.DU<EN, conform to the norm determined by Special.DU_orig, this is especially the case for Special.DU<FR, as the distance between its ellipse and the neutral cluster is larger than for the other sub-varieties. Futhermore, Special.DU<FR seems heavily influenced by the use of formal lexeme *thans*, from the profile *thans-nu* (concept: NOW).

In other words, although the three sub-varieties of Specialized Communication are significantly different from one another, this genre does behave homogeneously with regard to its overall preference for formal lexemes and the translated texts conform to the norm determined by the non-translated texts.

3.2.1.4 Journalistic Texts

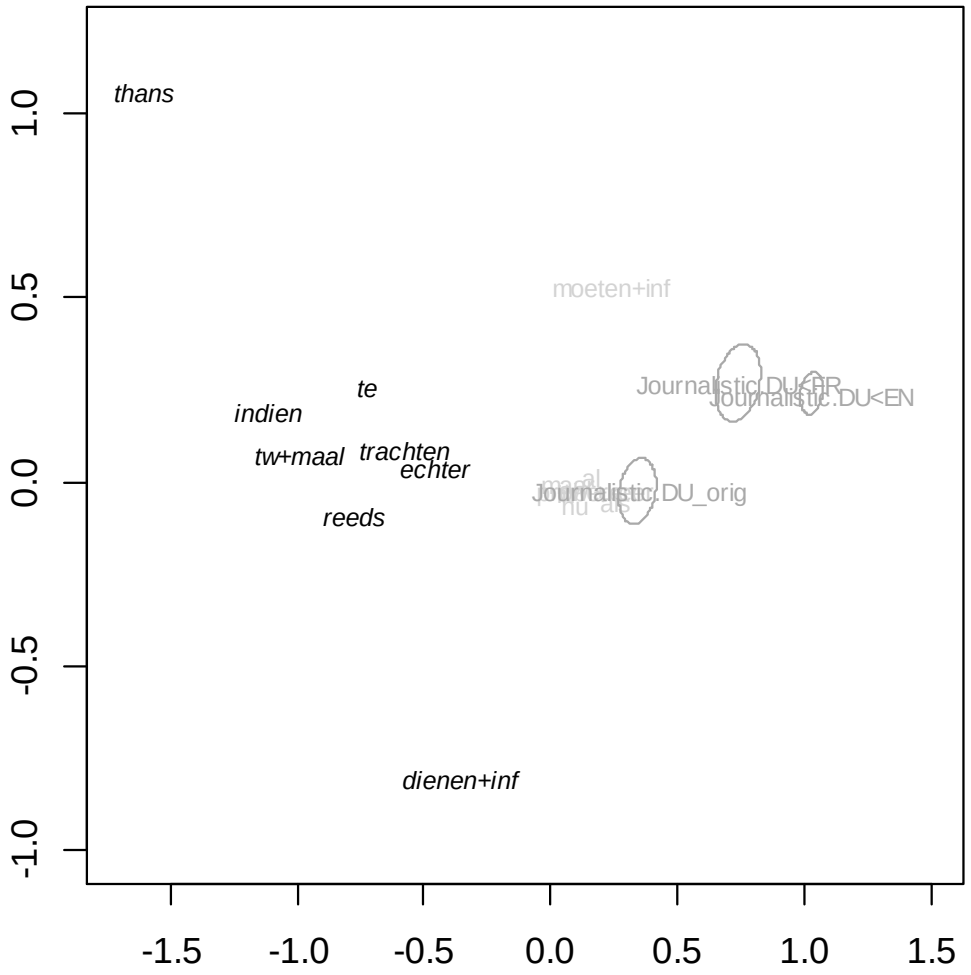

Figure 8 Case Study I: Biplot of the interaction between Journalistic Texts; and source language and translation status

Looking at the biplot presented in Figure 8, we can see the ellipses for Journalistic Texts translated from French (`Journalistic.DU<FR`), Journalistic Texts translated from English (`Journalistic.DU<EN`) and non-translated Journalistic Texts (`Journalistic.DU_orig`). Although all three sub-varieties of Journalistic Texts are significantly different from one another, they appear to prefer using neutral lexemes instead of formal lexemes. The difference between the various sub-varieties can therefore be explained in terms of specific neutral lexemes, and not in terms of a preference for formal or for neutral variants. This shows that, with regard to the choice between formal or neutral lexemes, translated Journalistic Texts conform to the norm

determined by non-translated Journalistic Texts. Moreover, this plot shows that there is no source language effect at play here, as both source language varieties show a preference for neutral lexemes.

Additionally, because of the overall preference for neutral lexemes by the three sub-varieties, it seems plausible to assume that within Journalistic Texts, there is a genre effect at play as these results show that neither source language nor translation status appear to have an influence on the linguistic preferences within this genre. These results might be due to the fact that this is a genre (i) which is mostly written by journalists, who often share a similar educational background and (ii) whose texts are often submitted to an editorial control phase. An alternative explanation for the homogeneous behavior of this genre is that authors or translators of Journalistic Texts often conform to implicit norms as "after a certain period of time, you know what is expected of you when you write a text for a given newspaper" (Böttner, 1995, p. 79 - own translation), although additional research is needed to investigate this further.

3.2.1.5 Tourist Information

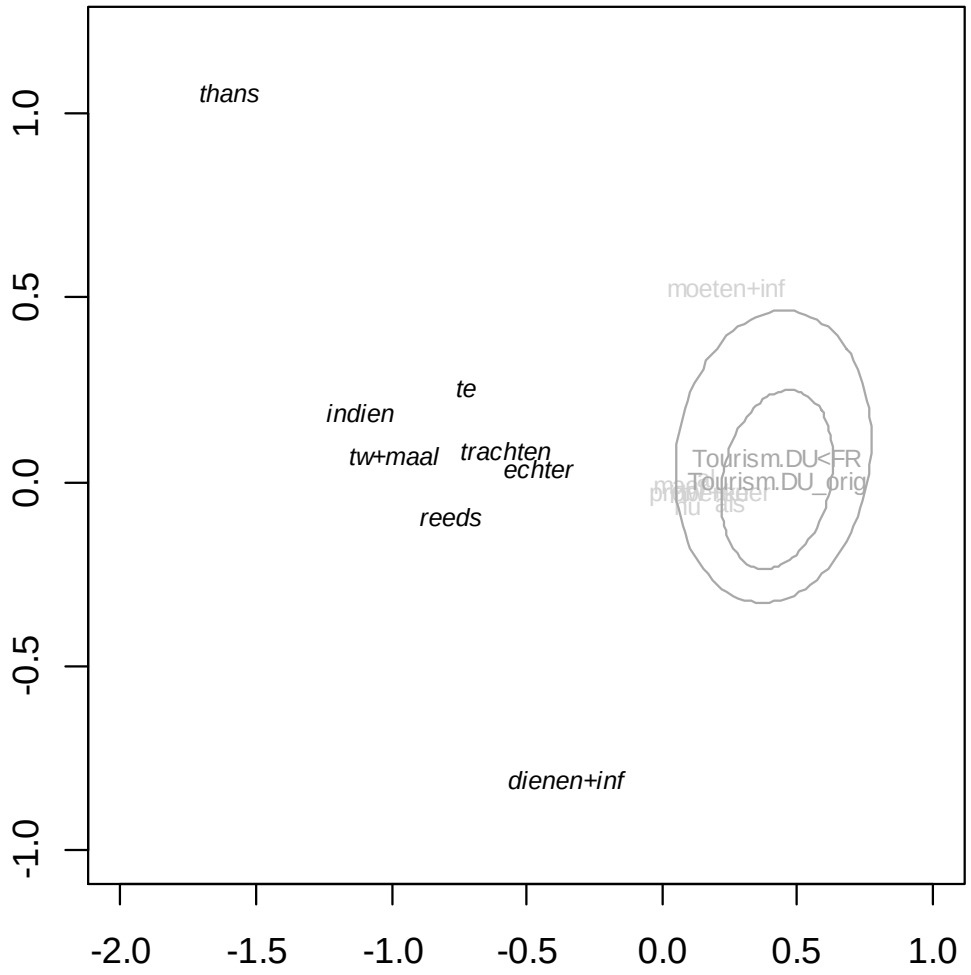

Figure 9 Case Study I: Biplot of the interaction between Tourist Information; and source language and translation status

Figure 9 shows the results of the interaction effect between Tourist Information and source language and translation status. As there are no data available in the corpus for Tourist Information translated from English, no results can be provided for this section and no ellipse can be calculated. The plot does show an interesting result for the remaining two sub-varieties, however, as there is a clear overlap between the ellipses for Tourist Information translated from French (`Tourism.DU<FR`) and non-translated Tourist Information (`Tourism.DU_orig`), which means these sub-varieties are not significantly different from one another. With regard to the norm-conformity hypothesis, this plot shows that translated Tourist Information conforms to the norm

determined by non-translated Tourist Information, which results in a clear preference for neutral lexemes.

3.2.1.6 Political Speeches

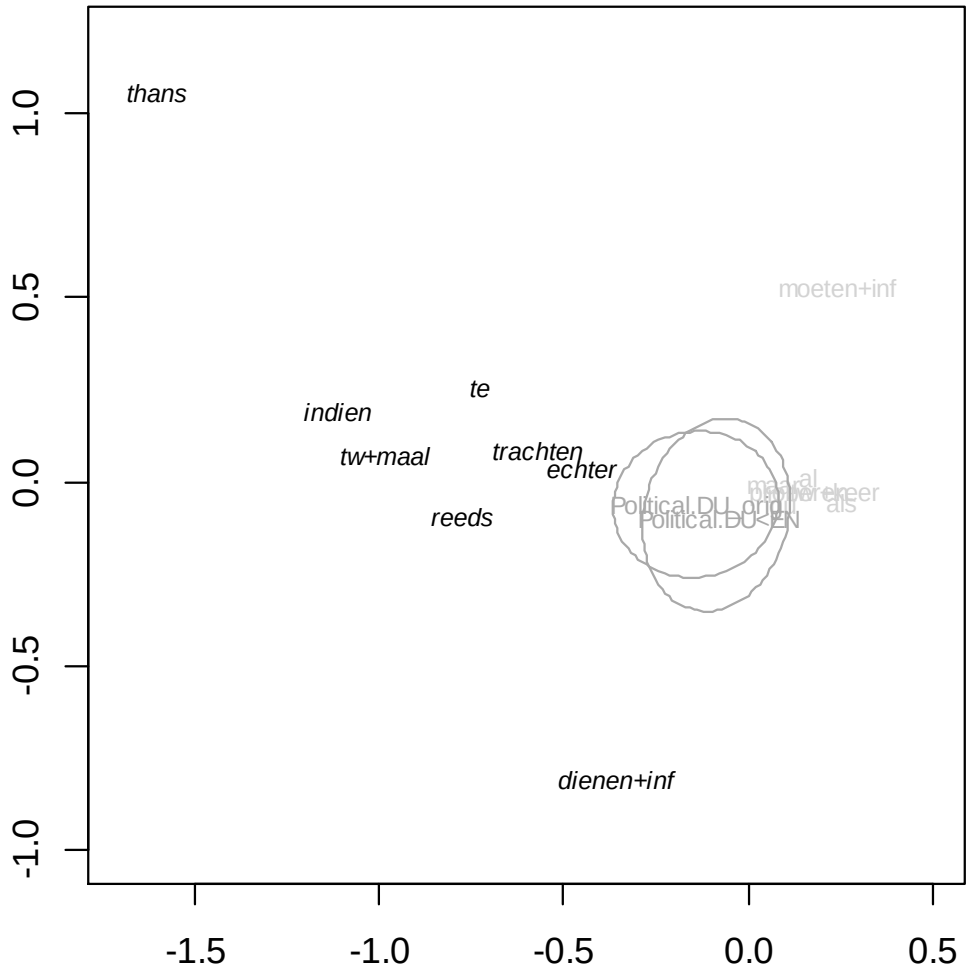

Figure 10 Case Study I: Biplot of the interaction between Political Speeches; and source language and translation status

Figure 10 shows a biplot of the interaction effects between Political Speeches on the one hand, and translation status on the other hand. As the Dutch Parallel Corpus contains no text material for Political Speeches translated from French, we cannot provide results regarding this sub-variety. This biplot shows that there are no significant differences between non-translated Political Speeches (`Political.DU_orig`) and Political Speeches translated from English (`Political.DU<EN`), which both show a preference for neutral lexemes. This indicates that translated Political Speeches

conform to the genre-determined norm established by non-translated Political Speeches.

3.2.1.7 Legal Texts

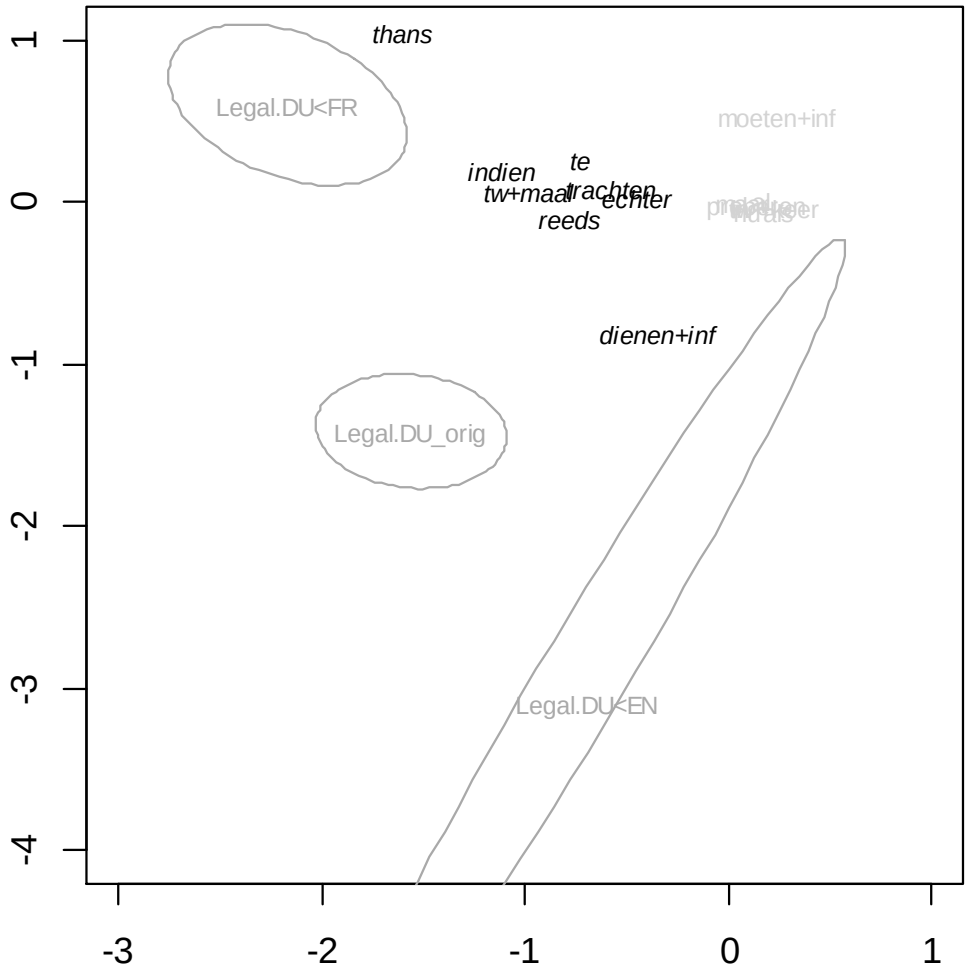

Figure 11 Case Study I: Biplot of the interaction between Legal Texts; and source language and translation status

The results presented in the biplot in Figure 11 show the interaction effect between the genre Legal Texts on the one hand, and the three (source) language varieties in our corpus on the other hand. This plot clearly shows that there are significant differences between all three sub-varieties, viz. Legal Texts translated from French (`Legal.DU<FR`), Legal Texts translated from English (`Legal.DU<EN`) and non-translated Legal Texts (`Legal.DU_orig`). The outstanding shape of Legal<EN immediately attracts attention and a closer examination of the data revealed that this

sub-variety only contains 3 data points, two of which are attestations of *dienen+inf,* hence its particular shape and position. As this prevents us from drawing truly meaningful conclusions with regard to this specific sub-variety, we will only look at `Legal.DU_orig` and `Legal.DU<FR` and it appears that translated Legal Texts conform to the genre-determined norm of using formal lexemes.

3.2.1.8 Instructive Texts

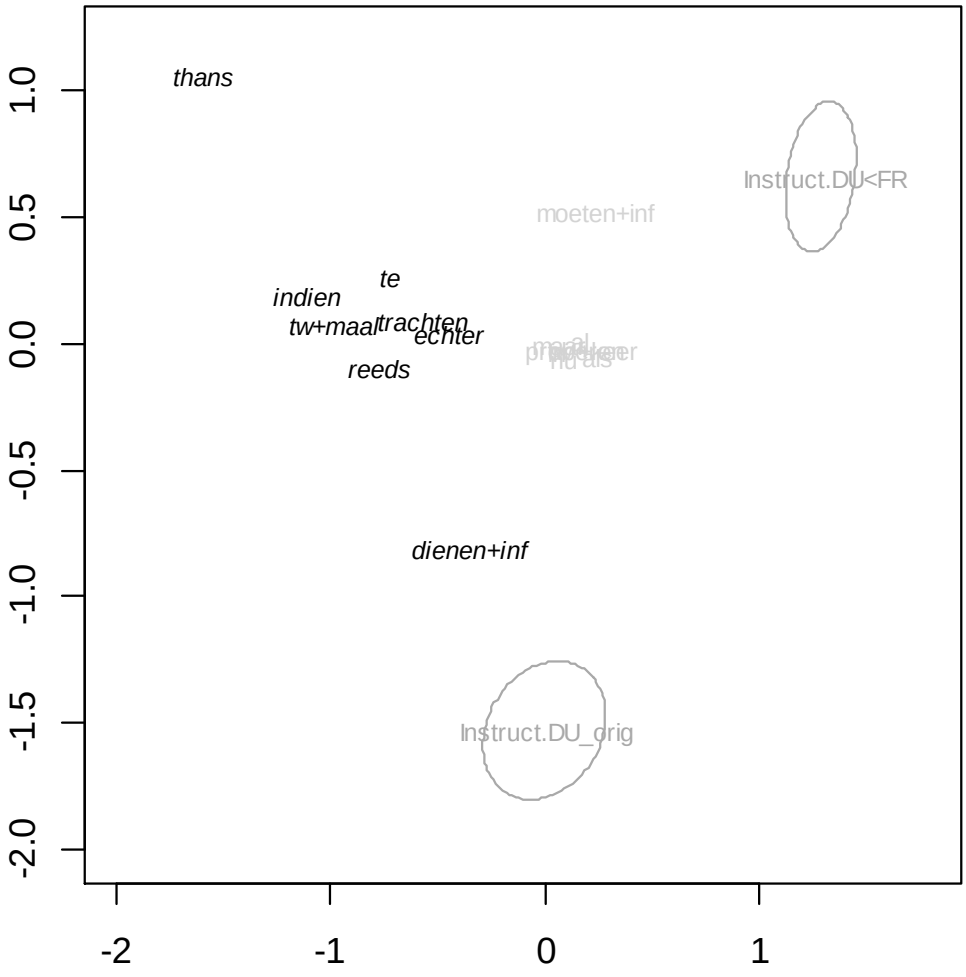

Figure 12 Case Study I: Biplot of the interaction between Instructive Texts; and source language and translation status

The interaction effects for the genre Instructive Texts are displayed in the biplot presented in Figure 12. This plots provides results for the only two sub-varieties of Instructive Texts which are available in the corpus, viz. non-translated Instructive Texts (`Instruct.DU_orig`) and Instructive Texts translated from French

(Instruct.DU<FR). The plot reveals that there is a significant difference between the two sub-varieties and the linguistic preferences. More particularly, this plot shows that whereas Instruct.DU_orig shows a less distinct preference with regard to the formal or neutral lexemes and its position seems especially determined by the lexeme *dienen+inf*, Instruct.DU<FR makes more use of the neutral lexemes. In other words, translated Instructive Texts do not (entirely) conform to the descriptive norm determined by non-translated Instructive Texts. A brief examination of the data revealed that the texts in Instruct.DU<FR come from one provider, i.e. Roularta, a publishing house. The texts in Instruct.DU_orig, on the other hand, come from a wider range of providers such as the RIZIV, the Belgian National Service for Medical and Disablement Insurance and the Federal Public Service of Social Security. It seems plausible that both provider and the corresponding specific audience have an influence on the level of formality in these texts, although further research is needed to corroborate this hypothesis.

3.2.1.9 Interaction results: summary

The interaction effect between genre on the one hand and source language and translation status on the other hand revealed a genre effect at play for the genres Political Speeches, Journalistic Texts, Tourist Information and Specialized Communication, albeit to a smaller degree for this final genre, as the translations conform to the norm established by the non-translations where the preference for neutral or formal lexemes is concerned. For the genres Broad Commercial Texts, Legal Texts and Instructive Texts, we could see that French as source language seems to have a particular effect.

3.3 Conclusion

This case study was carried out to examine the following hypothesis: "Genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to formal versus neutral lexemes as their corresponding non-translated genres".

Our results with regard to the global effect of translation showed that texts translated from English make more use of neutral lexemes than non-translated texts on the one hand and texts translated from French on the other hand. In other words: translations do not behave uniformly and the preference for either formal or neutral language depends not only on the translation status, but also on the source language. If

we look at the results on the global effect of genre, our analysis resulted in the following observations. There was a clear preference for formal language within some genres (Legal Texts and Specialized Communication) whereas other genres clearly preferred neutral language (Journalistic Texts, Political Speeches and Tourist Information). Finally, for Broad Commercial Texts and Instructive Texts the preference is less clear-cut.

Calculating the interaction effects between genre on the one hand and source language and translation status on the other allowed us to verify whether genres in translation conform to the norm established by the non-translated genres. Additionally, these analyses enabled us to determine how genre, translation status and source language affect one another. For this case study on the dispersion of formal versus neutral lexemes it was shown that (i) in some genres the hypothesis could be confirmed as translations do indeed follow the norm determined by the corresponding non-translated texts (Specialized Communication, Journalistic Texts, Tourist Information, Political Speeches, and Legal Texts), whereas (ii) in other genres in translation this was not the case (Broad Commercial Texts and Instructive Texts). Scanning the data showed that for Instructive Texts these results might be due to different types of text providers and/or target audience. It therefore appears that these two genres are still quite diverse which might explain why they behave differently depending on the sub-variety under investigation. For example, the sub-variety Instructive Texts translated from French originated from one particular provider, i.e. Roularta, a publishing house. The sub-variety non-translated Instructive texts originated from a wider variety of providers such as the RIZIV, the Belgian National Service for Medical and Disablement Insurance and the Federal Public Service of Social Security. As mentioned above, it seems plausible that this has an influence on the level of formality in these texts, although further research is needed to corroborate this hypothesis.

An additional, remarkable aspect concerns the effect of French as source language: in Broad Commercial Texts, Specialized Communication and Legal Texts this leads to an increase in formality whereas in Instructive Texts it leads to a decrease. The present study provides no easy explanation for this result, however, it seems reasonable to assume that in a given genre there are language-specific differences in formality. For example, Legal Texts in French might be a more formal genre than Legal Texts in Dutch. However, more in-depth research is needed to test this hypothesis.

Chapter 4
Case Study II: visualizing English loanwords versus more endogenous lexemes in search of descriptive norm conformity

4.1 Hypothesis and variable selection

4.1.1 Hypothesis

In this second case study, the difference in usage between English loanwords on the one hand versus their more endogenous counterpart on the other is investigated. Its aim is to verify whether genres in translation, in comparison with their non-translated counterparts, conform to the descriptive genre-determined norm with regard to using English loanwords or more endogenous lexemes. One of the advantages of applying a profile-based approach to investigate this matter is that, when comparing the use of loanwords versus more endogenous words, we can be certain of the comparison's quantitative results (see e.g. Zenner, Speelman, & Geeraerts, 2011). This is due to the use of profiles which ensure that when a loanword is used, there was always a more endogenous alternative available, an aspect which was verified during the manual validation step. In other words, the focus of this particular case study lies on investigating the choice for English loanwords which is not due to their untranslatability.

In line with the assumed norm-conforming behavior of translations and assuming that translations will thus tend to conform to the genre-specific norm, the following hypothesis was put forward:

> *"Genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to English loanwords and their interchangeable more endogenous lexemes as their corresponding non-translated genres."*

4.1.2 Variable selection

In order to operationalize this hypothesis we needed to find profiles which consist of a loanword on the one hand and a direct, interchangeable more endogenous counterpart on the other.

First, we must define what exactly is meant by the term 'English loanword'. The first aspect, English, is defined as a term "used to describe any word, phrase or sentence recognizable as belonging to any native or non-native variety of English" in Martin (2002, p. 9). With regard to the second aspect, loanword, we adopted the approach described in Zenner (2013) and used the etymological information which is provided in Van Dale's Great Dictionary of the Dutch Language to verify whether an item could be classified as English loanword or not (den Boon & Geeraerts, 2010-2013). Zenner suggests the following algorithm: "young loanwords such as *software* are by default considered to be English, and older loanwords (borrowed before 1945) are only considered to be loanwords if the pronunciation of the word is not what naive speakers of Dutch would anticipate based on the spelling of the word" (2013, p. 103). To illustrate this, Zenner provides the following examples: a naive pronunciation based on the Dutch spelling and pronunciation rules of the word 'manager' would sound like /mɑ'nɑːɣər/ instead of /'mɛnədʒər/, hence 'manager' is considered an English loanword. A naive pronunciation of the word 'film' in Dutch on the other hand is very close to how it is actually pronounced, and 'film' can therefore not be considered an English loanword.

We used the contents of the Dutch Parallel Corpus as a starting point for building our profile set. First, we extracted a case insensitive list of lemmas from the Belgian Dutch sub-corpus. We then analyzed this list and found that the only part-of-speech classes which seemed to contain words of English origin were nouns and verbs. Because the Dutch part of the corpus was pre-processed by a lemmatizer which was originally trained on Dutch data the lemmas it generated for the words of English origin were not always correct. Therefore, we extracted a case-insensitive list of unique words from the Belgian Dutch sub-corpus as well. We then thoroughly examined both lists of nouns and verbs and extracted the items (n=132) which appeared to be of English origin. As our technique is based on the use of profiles, only English terms which have a more endogenous counterpart could be considered for the profile set. Therefore, we verified whether this was the case for the terms in this basic list, which resulted in a list of 98 profile candidates.

The next phase consisted of extracting all instances for these remaining candidate profiles from the corpus so as to verify (i) how they were used, and (ii) whether, in general, the loanwords could be replaced by their more endogenous counterparts given the context they were used in (and vice versa). Additionally, we verified the English origin of the loanwords according to the algorithm mentioned above by means of Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013).

This first, rudimental sifting process resulted in a data set of 17 profiles, the attestations of which had to be manually validated in order to ensure the context-dependent interchangeability between the alternatives. Finally, the manual validation process in combination with a minimal frequency measure per profile and per factor resulted in a data set of 7 profiles or 14 alternatives (n=2331) which could be used for this case study, an overview of which is provided in Table 14. It should be noted that for the alternatives within the profiles, both the singular and the plural forms were taken into account. Furthermore, the corpus queries were carried out in a case-insensitive manner and for the profile "RESEARCH AND DEVELOPMENT" both the abbreviations (r&d; *o&o*) and the full forms (research and/& development; *onderzoek en ontwikkeling*) were taken into account.

Table 14 Frequencies of the loanwords and the more endogenous alternatives per genre and (source) language variety

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL		DU_orig	DU<EN	DU<FR		TOTAL
ploeg team	endogenous term loanword	TEAM	2 63	5 115	0 2	3 148	130 270		70 268	10 103	60 227		140 598
dienst-verlening service	endogenous term loanword	SERVICE	18 0	75 14	9 1	117 57	28 18		111 33	43 46	93 11		247 90
baan job	endogenous term loanword	JOB	0 6	4 13	35 45	21 24	74 144		57 142	44 25	33 65		134 232
afdeling unit	endogenous term loanword	DIVISION	23 0	76 2	3 0	62 9	172 11		208 12	65 10	63 0		336 22
o&o r&d	endogenous term loanword	"RESEARCH & DEVELOP-MENT"	1 0	84 15	11 0	23 28	15 15		44 25	90 32	0 1		134 58
partnerschap partnership	endogenous term loanword	PARTNERSHIP	1 0	15 10	7 1	105 59	3 19		75 24	44 50	12 15		131 89
hulpmiddel tool	endogenous term loanword	TOOL	12 0	30 20	1 0	15 11	10 21		33 18	9 12	26 22		68 52
TOTAL			126	478	115	682	930		1120	583	628		2331

In the following paragraphs, a description of the individual profiles and an example is provided. The Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013) was used for the concept descriptions. The example sentences

were extracted from the Dutch Parallel Corpus and their origin is presented by means of the file number. Each Dutch sentence contains one of the profile's variants which is matched by its translation or source text sentence in English.

Ploeg – team

Concept: TEAM - "a group of people who are or belong together for a certain goal"
Word class: noun
Example: DU: De **ploeg** met het schrijnendste verhaal moet die uit Liberia zijn.
EN: The **team** with the most poignant story must be the one from Liberia [...].
dpc-ind-001653-nl

Dienstverlening – service

Concept: SERVICE - "the offering of services"
Word class: noun
Example: DU: Ten eerste wilden we controleren hoe tevreden onze Europese industriële klanten zijn over onze producten en **dienstverlening**.
EN: First, we wanted to evaluate satisfaction levels among our European industry customers with regard to our products and **services**.
dpc-arc-002042-nl

Baan – job

Concept: JOB - "a post, a position"
Word class: noun
Example: DU: Dit cijfer van 113 voltijdse **banen** houdt ook rekening met [...].
EN: This figure of 113 full-time **jobs** also takes into account [...].
dpc-bco-002447-nl

Afdeling – unit

Concept: DIVISION - "separate unit of people who pertain to a larger association, society or organization"
Word class: noun
Example: DU: De **afdeling** leverde LED-visualisatieoplossingen aan klanten [...].
EN: As such the **division** has provided LED visual solutions to clients [...].
dpc-bco-002433-nl

Onderzoek en ontwikkeling – research and development

Concept: RESEARCH AND DEVELOPMENT – "research and (product) development"

Word class: noun
Example: DU: [...] zullen beide bedrijven alle kosten voor **onderzoek en ontwikkeling** in gelijke mate dragen.
EN: [...] both companies will equally share all **research and development** costs.
dpc-aby-002285-nl

Partnerschap – partnership
Concept: PARTNERSHIP - "the aspect of being partners"
Word class: noun
Example: DU: Een **partnerschap** gebaseerd op totaal vertrouwen, is [...].
EN: A **partnership** like this, based on total mutual confidence, is [...].
dpc-arc-002047-nl

Hulpmiddel – tool
Concept: TOOL - "means which facilitate reaching a goal"
Word class: noun
Example: DU: Aanvankelijk was twinning een **hulpmiddel** om de integratie te ondersteunen [...].
EN: Twinning began as a **tool** to support integration [...].
dpc-arc-002046-nl

4.2 Results

The main goal of this second case study is to investigate descriptive, genre-determined norm conformity, which is done by verifying whether genres in translation show a similar preference for either loanwords or more endogenous lexemes as their corresponding non-translated genres.

A second aspect, which seems equally interesting to investigate is the global effect of the factors genre and translation, as the linguistic distances between the various genres and (source) language varieties in our corpus and the position of these factors with respect to the linguistic variants will reveal the dominant preference for either more endogenous lexemes or loanwords within that specific (source) language variety or genre in comparison with the other varieties and genres. These global effects will be discussed in Section 4.2.1.

So as to verify whether genres in translation demonstrate similar linguistic preferences in comparison with their non-translated genres we calculated the interaction effects between genre and translation, the results of which are discussed in Section 4.2.2

4.2.1 Global effects of genre and translation

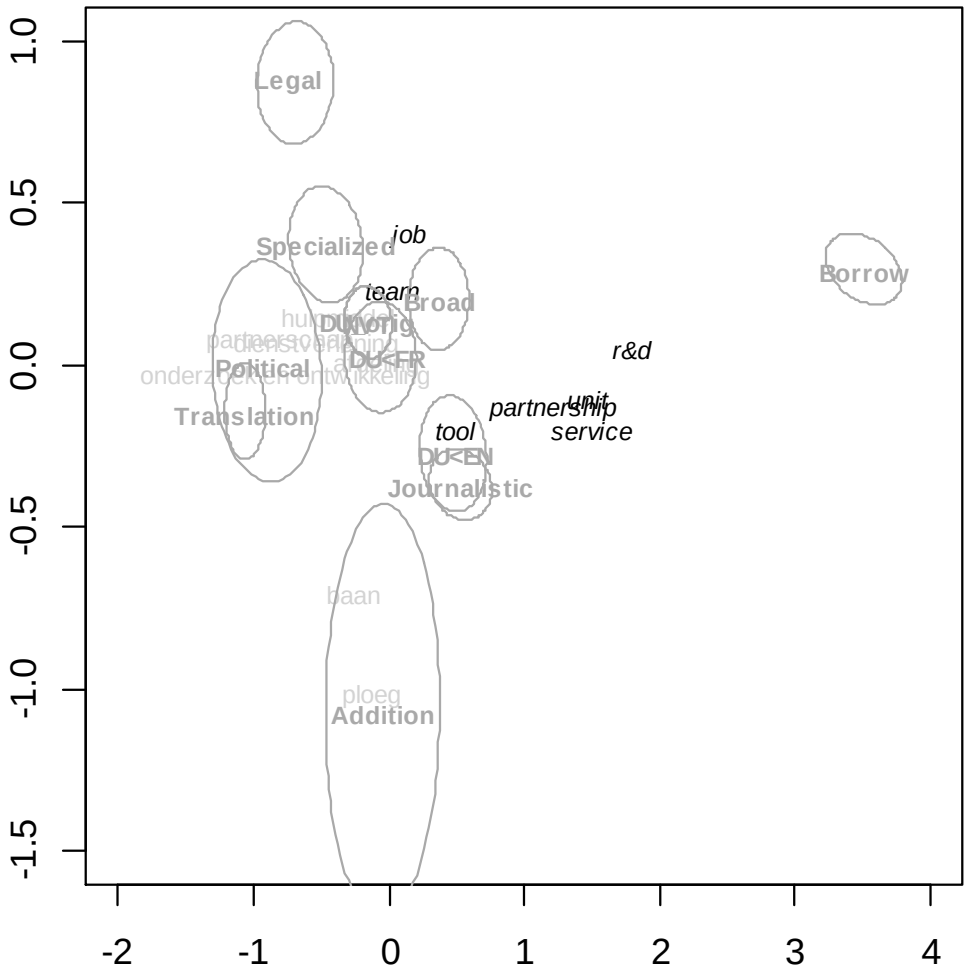

Figure 13 Biplot of the global results with the more endogenous lexemes (light grey), the loanwords (black italics), the genres, the (source) language varieties. and the translation strategies

Figure 13 is a biplot which shows the global results and hence contains information with regard to the linguistic profiles, genre, source language, translation status, and translation strategy.

This biplot shows the dispersion of our linguistic profiles which consist of more endogenous lexemes (in light grey) and loanwords (in black italics), the position of which was determined by their profile-based frequencies in the various (source) language varieties and genres in our corpus. The dispersion of these linguistic variants clearly reveals that the most important dimension of our plot, i.e. the X-axis, is defined by the opposition between more endogenous lexemes on the left-hand side of the plot and English loanwords on the right-hand side of the plot.

Throughout the following paragraphs of this results section the plot in Figure 13 will be subdivided into various subplots, as plots with a lower information density allow us to interpret the results more easily. The original positions of the data and the in-between distances are not altered, we merely zoomed in or out on the main plot to create these sub-plots as this allowed for a closer examination and revealed various interesting aspects more easily.

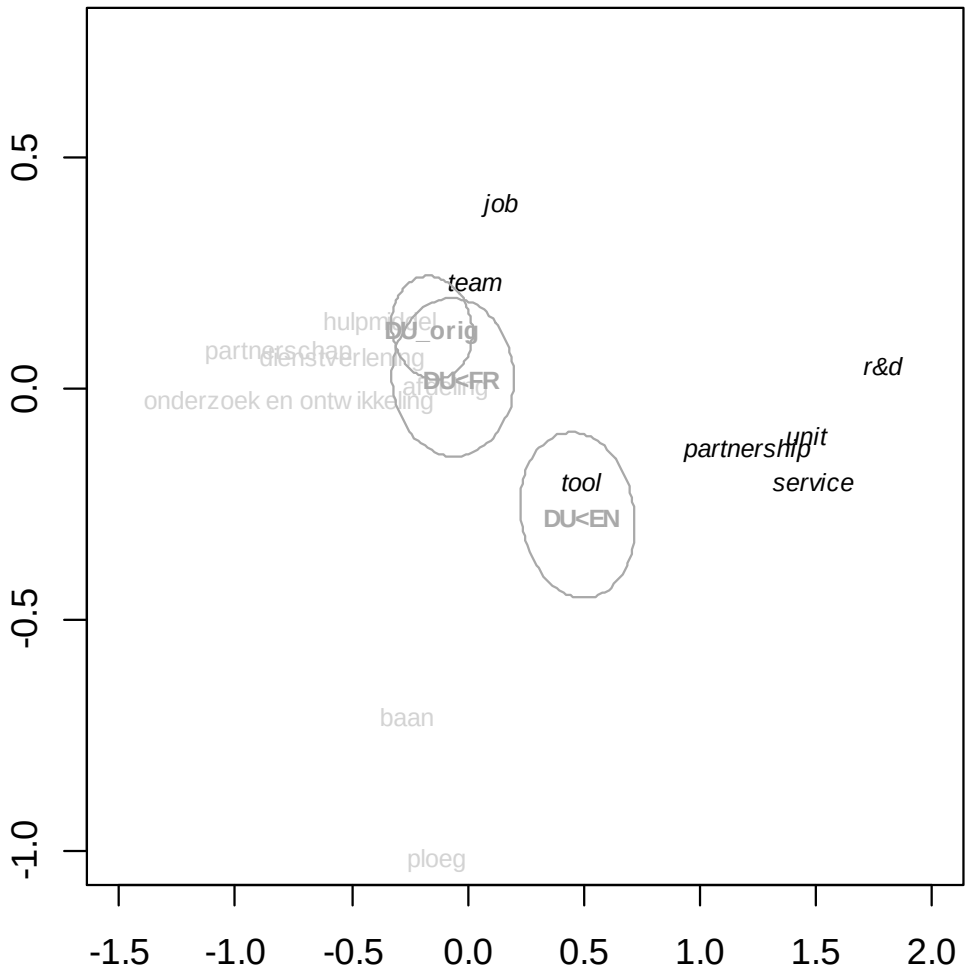

Figure 14 Biplot of the (source) language varieties, the more endogenous lexemes (light grey) and the loanwords (black italics)

As mentioned in the previous paragraphs, Figure 14 is proportionally the same as Figure 13, however, it focuses on the source language effect and translation status effect only. Figure 14 shows that there are significant differences between DU<EN and DU_orig and between DU<EN and DU<FR, but that there is no significant difference between DU_orig and DU<FR. In other words, this plot reveals no significant effect caused by translation status, as it is not the case that original Dutch is significantly different from translated Dutch as a whole. Moreover, the plot shows that non-translated Dutch and Dutch translated from French appear to make more use of more endogenous lexemes while Dutch translated from English makes more use of English loanwords. These results add an extra dimension to previous research on the use of English loanwords which found that there are indeed "contrasting tendencies in terms

of the use of Anglicisms on the part of translators and authors originally writing in Italian" and that translators make less use of Anglicisms than authors of comparable, non-translated texts (Bernardini & Ferraresi, 2011, p. 230). Similarly, Laviosa found that a corpus-driven teaching approach applied by translation trainees made her students aware of the fact that "in translational language there seems to be a preference for native equivalents" (2006, p. 272). While these two studies only investigated translations from one source language, i.e. English, the present study shows that the specific source language does seem to have an influence on the linguistic preferences, which is not surprising as we investigated the use of loanwords from English versus more endogenous lexemes.

However, the plot in Figure 14 does reveal somewhat surprising results as Dutch translated from English is situated only slightly towards the right hand side, showing only a slight preference for loanwords, where one might have expected a stronger preference, especially in comparison with Dutch translated from French or original Dutch.

In an attempt to find extra information or potential explanatory factors with regard to these surprising results, we investigated the data which, however, revealed no additional explanatory factors. For example: in the DU<EN sub-corpus the concept PARTNERSHIP occurred 94 times and was either referred to by the more endogenous lexeme (*partnerschap*) or the loanword (*partnership*). In 91 of the cases, the source word was 'partnership', which was translated almost as often by *partnerschap* (n = 43) as *partnership* (n = 48). In other words, it seems that, for this particular case study, English as source language is only slightly more likely to trigger the use of a loanword in translations into Dutch than French.

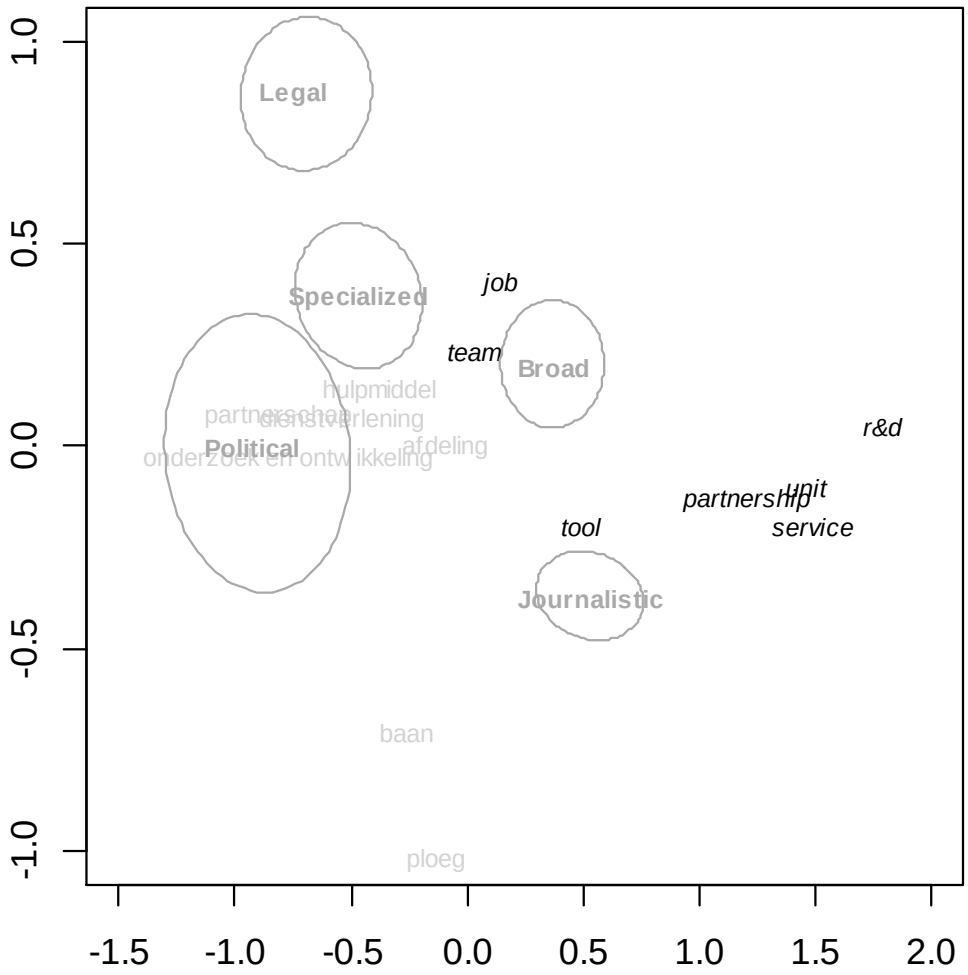

Figure 15 Biplot of the genres, the more endogenous lexemes (light grey) and the English loanwords (black italics)

Figure 15 shows only the results for the five genres in our dataset, viz. Legal Texts (`Legal`), Specialized Communication (`Specialized`), Political Speeches (`Political`), Broad Commercial Texts (`Broad`), and Journalistic Texts (`Journalistic`).

In addition to the dispersion between the more endogenous lexemes and loanwords in the horizontal direction, i.e. the X-axis, this plot shows a similar dispersion between `Political`, `Legal` and `Specialized` on the one hand and `Broad` and `Journalistic` on the other. This left-right division of the genres can be freely interpreted and one possible interpretation could be based on the level of 'specificity' per genre: while `Political`, `Legal` and `Specialized` might have more specific contents, vocabulary and target audience, this is not the case for `Broad` and

Journalistic, which are, on similar levels, broader genres. The position of these broader genres with respect to the more specific genres corresponds with the dominance of loanwords for Broad and Journalistic, whereas Political, Legal and Specialized make more use of more endogenous lexemes. Although additional research and alternative methods are needed so as to explain the observed differences, we would like to put forward the following, however speculative, explanatory hypotheses based on these observations. It seems not entirely implausible that for Legal Texts there is a more cautious attitude towards the use of English loanwords as the language used in this genre is often rather conservative and in some cases even archaic. In contrast, there might be a tendency for the language used in a genre like Journalistic Texts to be more susceptible to trends, which might explain why this genre makes more use of English loanwords.

A qualitative analysis of our dataset was carried out in an attempt to identify additional explanatory factors for the differences between the genres. However, similar to the analysis of the data for source language and translation status in the previous section, this analysis revealed no additional remarkable aspects in the data, which leads us to conclude that genre as such might be a determining factor where the difference in usage of more endogenous lexemes versus loanwords is concerned. The factor 'genre' includes various aspects which were not investigated in this study, but which might help explain the observed differences. For example, English loanwords might enjoy high status in one genre, while this is not the case for another. Similarly, while conventions with regard to the use of loanwords might exist within one genre, this is not necessarily the case for another. However, further research is needed to investigate these and other potential explanatory factors.

4.2.2 Interaction effects between genre and translation

The biplots in the previous section can only provide us information regarding the global effects and cannot give us any insight into whether a given genre in translation behaves similarly to its non-translated counterpart, thus conforming to the descriptive, genre-determined norm. Additionally, the previous plots provide no information on whether a given source language has a particular influence on the linguistic behavior of a genre in translation or not. It would be interesting, however, to find out whether Journalistic Texts, for example, behave differently when they are translated from French or translated from English. Therefore, we calculated the interaction effects between genre and translation, which are described in the following paragraphs.

4.2.2.1 Broad Commercial Texts

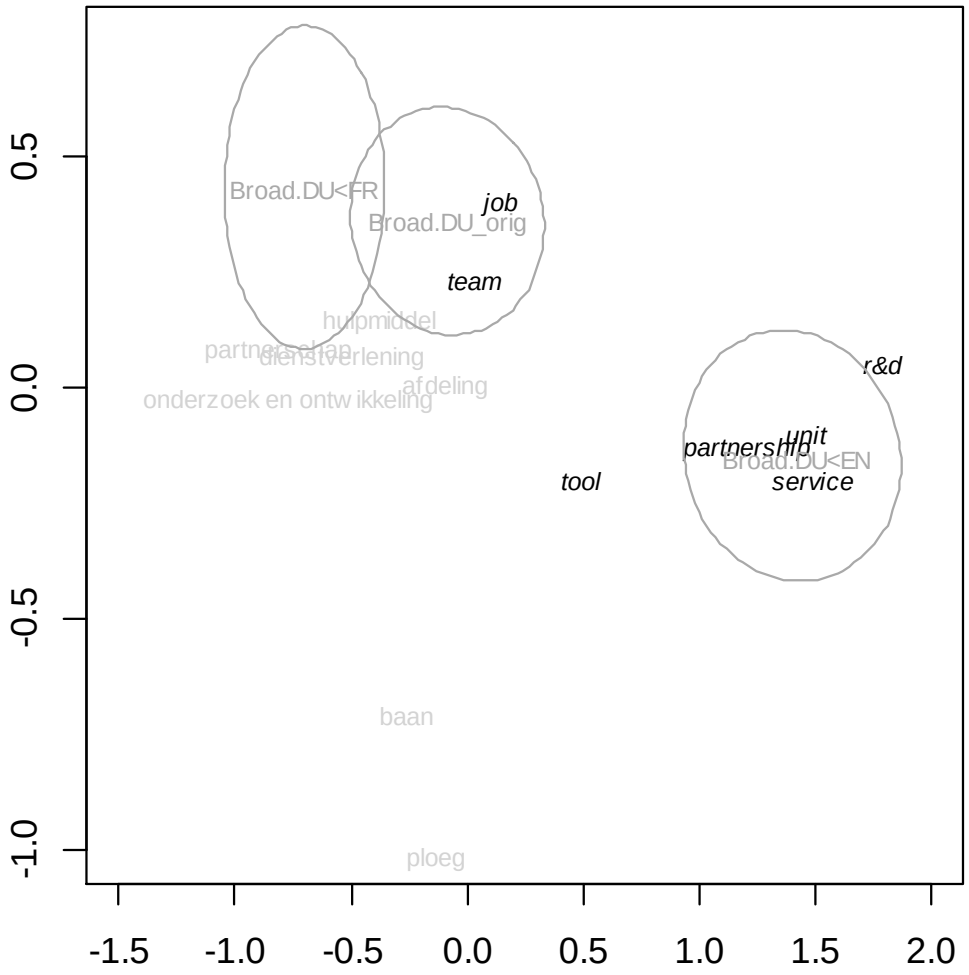

Figure 16 Case Study II: Biplot of the interaction between Broad Commercial Texts; and source language and translation status

Figure 16 shows the interaction effects between Broad Commercial Texts and the (source) language varieties, which results in the following varieties: Broad Commercial Texts translated from English (`Broad.DU<EN`), translated from French (`Broad.DU<FR`) and non-translated Broad Commercial Texts (`Broad.DU_orig`). There is no significant difference between `Broad.DU<FR` and `Broad.DU_orig`, which both seem to make more use of more endogenous lexemes, whereas there is a difference between `Broad.DU<FR` and `Broad.DU_orig` on the one hand and `Broad.DU<EN` on the other, which contains more loanwords.

Whether the established norm in non-translated texts which results in a preference for more endogenous lexemes is followed in translated Broad Commercial Texts thus depends on the source language. Moreover, this source language influence is in line with what we might have expected from an intuitive point of view as it seems logical that texts translated from English (`Broad.DU<EN`) contain more English loanwords than texts translated from French (`Broad.DU<FR`).

4.2.2.2 Specialized Communication

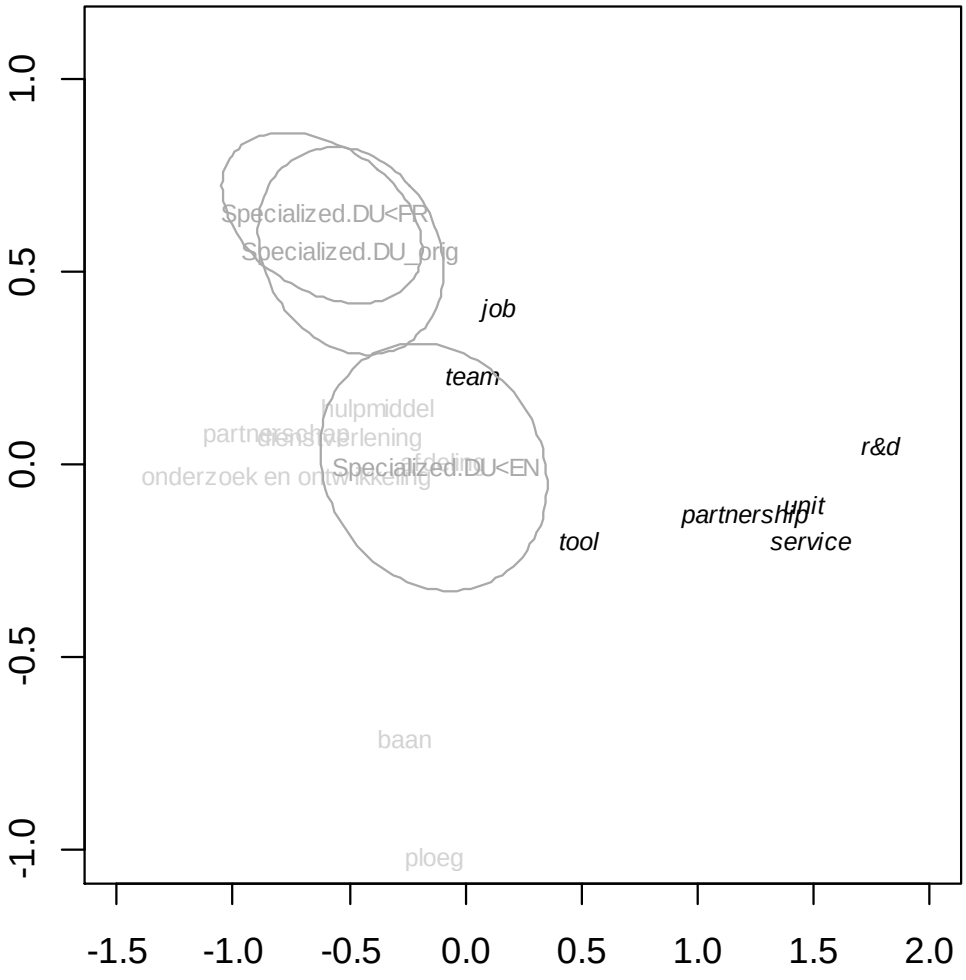

Figure 17 Case Study II: Biplot of the interaction between Specialized Communication; and source language and translation status

Figure 17 visualizes the results of the interaction effect between Specialized Communication and source language and translation status by means of confidence

ellipses for Specialized Communication translated from English (Specialized.DU<EN), Specialized Communication translated from French (Specialized.DU<FR) and non-translated Specialized Communication (Specialized.DU_orig). As can be seen, the interaction effect is similar to that in Broad Commercial Texts, as there are no significant differences between Specialized.DU<FR and Specialized.DU_orig, which are both situated close to the more endogenous lexemes. Although Specialized.DU<EN is significantly different from the other two sub-varieties, this sub-variety also appears to have a preference for more endogenous lexemes instead of loanwords.

Summarizing these results, it seems that even though there are significant differences between some of the sub-varieties, these differences are mainly based on a different preference for specific lexemes, as all sub-varieties show a preference for the more endogenous lexemes. In other words, in this particular genre translated texts conform to the norm established by non-translated texts which results in an overall preference for more endogenous lexemes.

4.2.2.3 Journalistic Texts

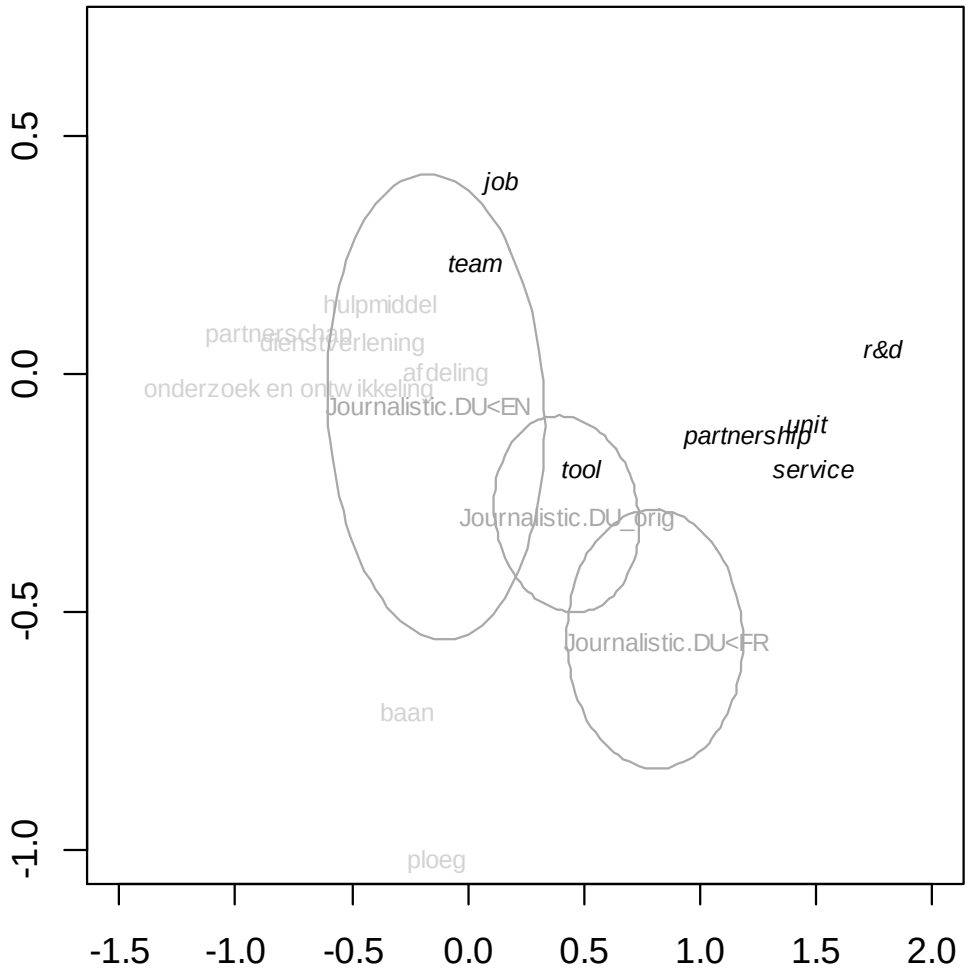

Figure 18 Case Study II: Biplot of the interaction between Journalistic Texts; and source language and translation status

Figure 18 shows the linguistic behavior for Journalistic Texts translated from English (`Journalistic.DU<EN`), Journalistic Texts translated from French (`Journalistic.DU<FR`), and non-translated Journalistic Texts (`Journalistic.DU_orig`). This plot shows that there are no significant differences between `Journalistic.DU<EN` and `Journalistic.DU_orig` on the one hand and between `Journalistic.DU<FR` and `Journalistic.DU_orig` on the other hand. More particularly, these results show that non-translated Journalistic Texts appear to make as much use of English loanwords as of more endogenous lexemes. Journalistic Texts in translation, however, show different preferences depending on the

source language and Journalistic.DU<FR appears to make slightly more use of English loanwords than Journalistic.DU<EN, which seems counterintuitive as one would expect an opposite source language effect for this particular case study.

Examining our data showed that in a few cases in DU<FR, the French source text contained an English word: '*job*' (n=5) and '*research and development*' (n=1). In addition to truly English word forms, there are various occurrences in the source texts of Journalistic.DU<FR of a French lexeme which is identical to the English trigger word for the loanword in our dataset, i.e. *service*. (n=8). All 8 occurrences of '*service*' in the source texts were translated by *service* in the target texts. Apart from this partial and potential explanation, there seemed to be no clear indication as to why English loanwords occur more often in Journalistic texts which were translated from French.

Summarizing the observations with regard to the genre Journalistic Texts, it seems that source language has a certain influence on the linguistic preferences for loanwords or more endogenous lexemes and, as a result, translated Journalistic Texts do not behave as a whole in conforming to the norm established by non-translated texts.

4.2.2.4 Political Speeches

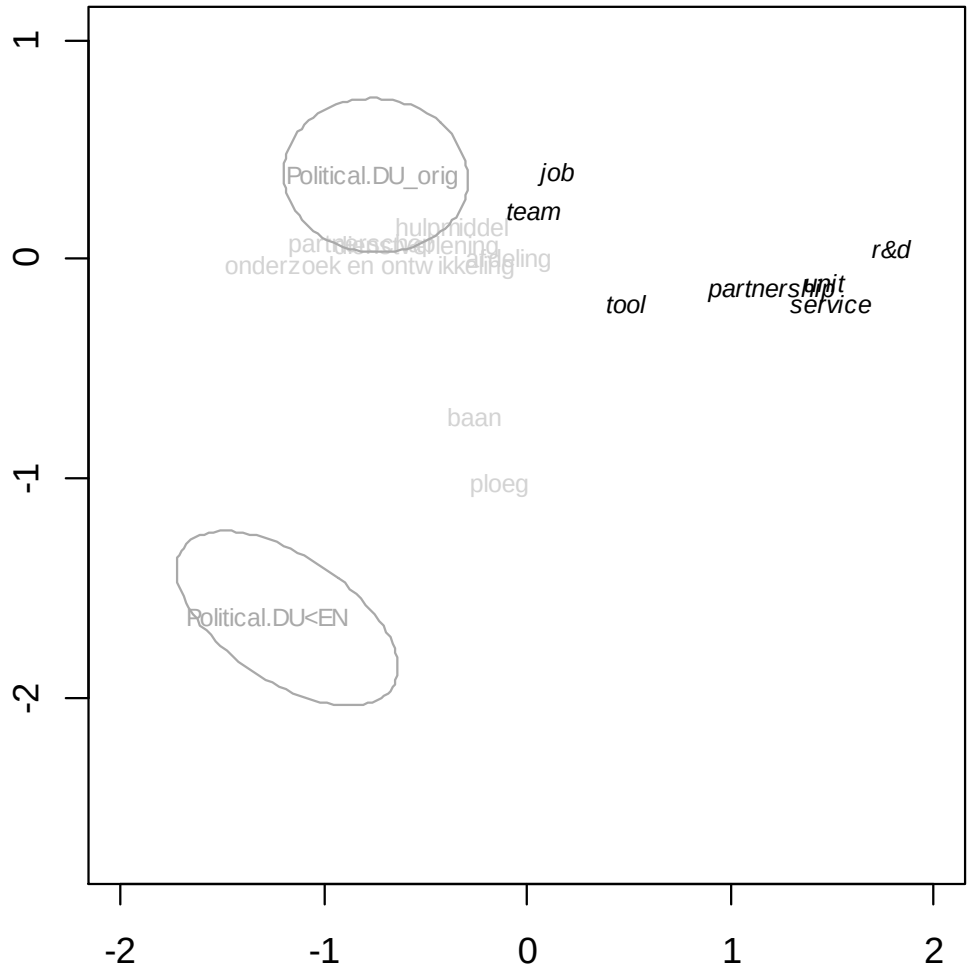

Figure 19 Case Study II: Biplot of the interaction between Political Speeches; and source language and translation status

There are only data available from one source language, i.e. English. Therefore, Figure 19 only shows the ellipses for Political Speeches translated from English (`Political.DU<EN`) and non-translated Political Speeches (`Political.DU_orig`), allowing us to verify the influence of translation status, but not of source language. As can be seen, `Political.DU_orig` is situated very close to the core of the more endogenous lexemes due to a strong preference for these lexemes. `Political.DU<EN` is situated less close to the core, but still shows a preference for the more endogenous lexemes as the distance between `Political.DU<EN` and the endogenous lexemes is smaller than the distance between `Political.DU<EN` and

the loanwords. Taking a closer look at the dataset revealed that `Political.DU<EN` is rather small in size (n=25) and heavily influenced by the profile *baan-job*, and more particularly the lexeme *baan* (n=18), which explains its specific position in the plot.

Similar to Specialized Communication, a significant difference between `Political.DU<EN` and `Political.DU_orig` could be observed, but this is mostly due to the high frequency for one or more specific lexemes as both sub-varieties show an overall preference for more endogenous lexemes. In other words, translated Political Speeches conform to the genre-determined norm and display similar behavior with regard to the choice between loanwords and more endogenous lexemes.

4.2.2.5 Legal Texts

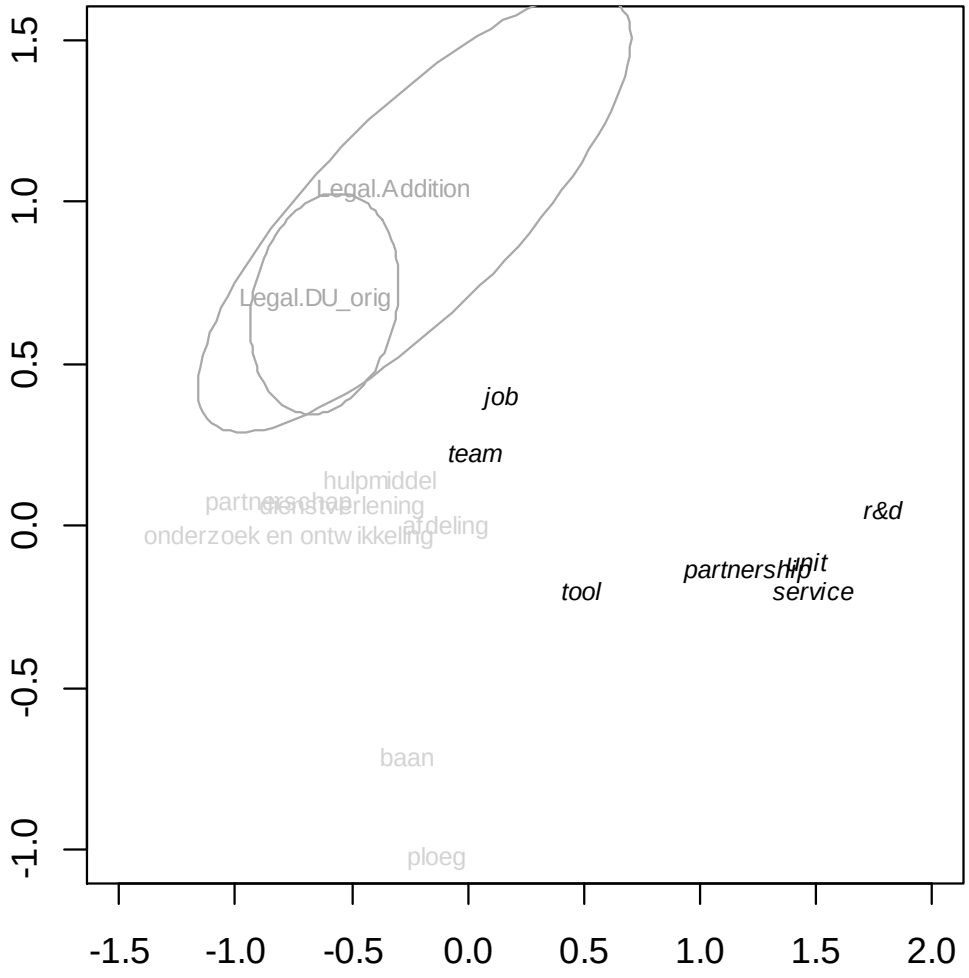

Figure 20 Case Study II: Biplot of the interaction between Legal Texts; and source language and translation status

Once again, the corpus contains only data for Legal Texts translated from one source language which prevents us from investigating the influence of source language on this genre. However, we can verify to what extent Legal Texts translated from French behave conform to the descriptive norm established in non-translated Legal Texts. As can be seen in Figure 20, there are no significant differences between these two sub-varieties, showing that translation status has no influence on the linguistic behavior of this genre. More particularly, `Legal.DU<FR` conforms to the norm established in `Legal.DU_orig` which results in a preference for more endogenous lexemes instead of loanwords.

4.2.2.6 Interaction results: conclusion

Analyzing the interaction effects for the various genres in our dataset revealed that there was a source language effect on the norm conforming behavior in translated Broad Commercial Texts as texts translated from French conform to the norm, but texts translated from English do not. Within the genre Journalistic Texts, there was a source language effect at play as well and translated Journalistic Texts do not behave as a whole in conforming to the norm established by non-translated Journalistic Texts.

For the genres Specialized Communication, Political Speeches, and Legal Texts, a genre effect could be observed, as the sub-varieties for these genres behave homogeneously and the translated genres conform to the norm established in their non-translated counterparts.

4.2.3 Additional factor: source text lexemes

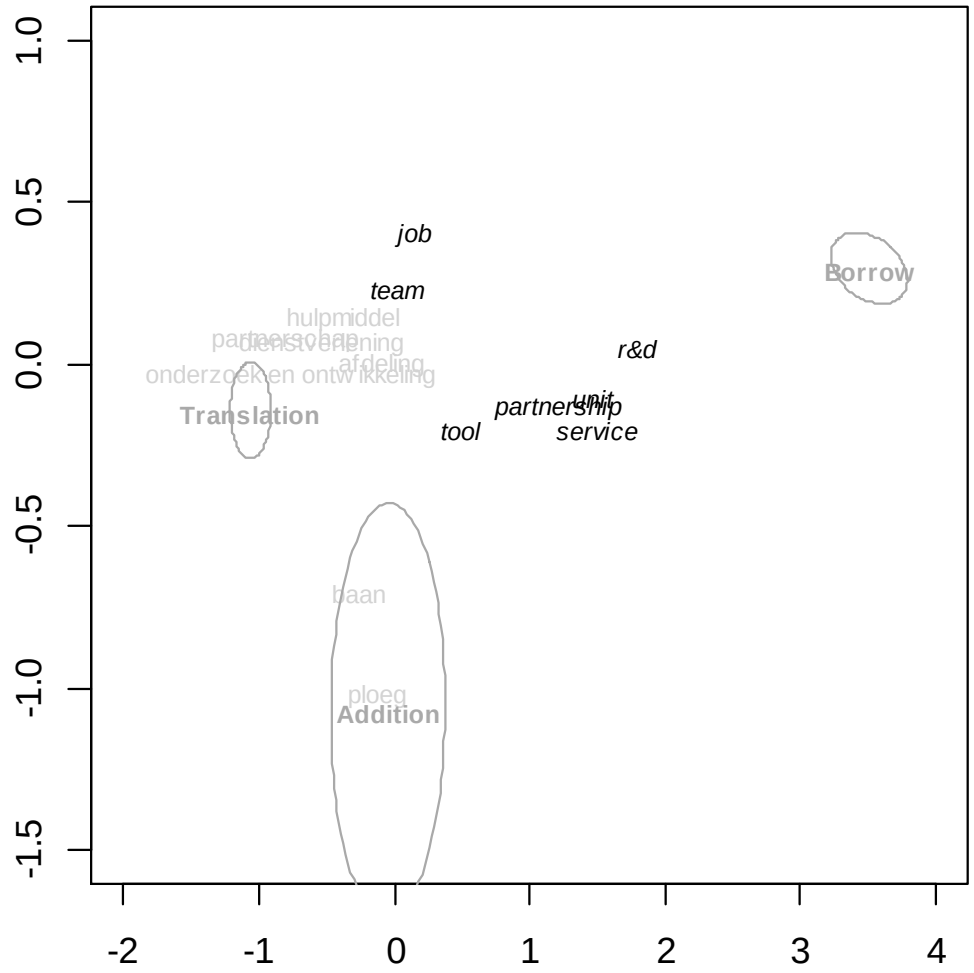

Figure 21 Case Study II: Biplot of the translation strategies, the more endogenous words (light grey) and the loan words (black italics)

Investigating the dispersion of loanwords versus more endogenous lexemes throughout the Dutch Parallel Corpus triggered the question as to whether, in addition to genre and source language, the specific lexemes in the source texts could have an influence on the choice for a loanword versus a more endogenous lexeme in the translation. For example, an English source text may contain a lexeme (e.g. 'unit') which can either be copied in the Dutch translation (e.g. *unit*) or translated as a more endogenous lexeme (e.g. *afdeling*). According to our hypothesis "genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to English loanwords and their interchangeable more endogenous lexemes as their

corresponding non-translated genres", the fact that the source text contains a lexeme which can easily be borrowed in the translation should not have an influence on the linguistic preferences in the translation.

In order to verify whether this is indeed the case, we annotated the translation sub-corpora (DU<FR and DU<EN) by adhering to the following guidelines:

- If there was no corresponding item in the source text, and the more endogenous lexeme or loanword was added to the translation, this was labeled as 'addition'.
  E.g. "technical support" => "***afdeling*** *voor technische ondersteuning*"
- If the corresponding item in the source text has a direct translation in the target text which is not a clear copy, this was labeled as 'translation'.
  E.g. "the team" => "*de* ***ploeg***"
- Finally, if the item in the target text was clearly copied from the source text, this was labeled as 'borrow'.
  E.g. "unit" => "***unit***"

Figure 21 shows the ellipses for the various translation options that were detected in the translated dataset, viz. Addition, Translation and Borrow, which were annotated based on the annotation guidelines mentioned above. This additional plot was mostly generated for the sake of completeness, as we did not expect it to reveal anything other than Translation to be situated near the more endogenous lexemes, Borrowing near the loanwords, and Addition somewhere in between. The effect of this factor was especially insightful when investigated in combination with the factor genre, which will be described below.

A closer examination of the data for Addition however, did show an interesting aspect with regard to the profile team-*ploeg*, and how *ploeg* is often used to make something explicit in the target text. For example, a large portion of the attestations for *ploeg* as Addition were found in Journalistic Texts which cover news regarding football. In the source texts, football teams are mostly referred to by merely using the name of the team, e.g. "*...Bétis Seville...*", while the target texts often add *ploeg*, e.g. "*Betis Sevilla, de ploeg...*".

This additional annotation step was taken because we were interested in finding answers to the following questions: If the source text contains a possible trigger (e.g. 'unit', 'job' or 'team'), does this automatically lead to the use of a loanword in the translation (e.g. *unit, job* or *team*) or is it changed into a more endogenous lexeme in translation (e.g. *afdeling, baan* or *ploeg*)? Moreover, what is the influence of the source text lexeme on the extent to which a given genre in translation conforms to the relative, genre-determined norm in its non-translated counterpart?

Additionally, as one of the goals of this dissertation is to determine whether genre has an influence on linguistic preferences, we wanted to investigate whether the linguistic norm-conforming behavior alters depending on the genre at hand. In order to

find answers to the questions mentioned here, the following steps were carried out. First of all, as only the English source texts could contain trigger words, we extracted the DU<EN data from the dataset. Secondly, we only maintained those attestations for which the source text contained a possible trigger word, viz. 'team', 'service', 'job', 'unit', 'research & development', 'partnership', or 'tool'. In a third step, we verified to what extent these possible trigger words were either translated, resulting in a more endogenous lexeme in the translation, or borrowed, resulting in a loanword in the translation. More specifically, we calculated the relative frequencies of each translation strategy per genre, an overview of which can be found in Table 15. As the corpus contains no Legal Texts translated from English, Table 15 provides no information with regard to this genre.

Table 15 Survey of the relative frequencies of each translation strategy per genre with a trigger word in the source text

	BORROW (loanword)	TRANSLATION (endogenous word)
BROAD COMMERCIAL TEXTS	65%	35%
JOURNALISTIC TEXTS	66%	34%
POLITICAL SPEECHES	0%	100%
SPECIALIZED COMMUNICATION	34%	66%

As can be seen in Table 15, whether a possible trigger source word is either borrowed or translated more often, does indeed depend on the genre. More specifically, within Broad Commercial Texts and Journalistic Texts, the English source word is often borrowed whereas in Political Speeches and Specialized Communication, the English source word is more often translated into a more endogenous word. As mentioned in Section 4.2.1, these results could be due to the varying status of English in a given genre, or to specific guidelines with regard to the use of loanwords, or to prevailing genre-determined trends. However, further research is needed to determine the exact causes for the linguistic preferences per genre.

In addition to this first observation, we also wanted to find the answer to the following, related question: Is there a genre-related difference between using a loanword and a more endogenous lexeme if the source text contains no possible trigger word? In other words, if the source text contains a lexeme which is not 'team', 'service', 'job', 'unit', 'research & development', 'partnership', or 'tool', how often does the target text contain one of these loanwords? The following steps were carried out in order to investigate this matter. First of all, we also started by extracting the DU<EN subset. Secondly, we removed all attestations for which the source text lexeme was one of the seven possible trigger words. Finally, we calculated the relative frequencies of the more endogenous lexemes and the loanwords per genre, which are represented in Table 16. Once again,

this table provides no results with regard to Legal Texts as there are no data available for this genre.

Table 16 Survey of the relative frequencies of the more endogenous lexemes and the loanwords per genre without a trigger word in the source text

	loanword	endogenous lexeme
BROAD COMMERCIAL TEXTS	21%	79%
JOURNALISTIC TEXTS	47%	53%
POLITICAL SPEECHES	0%	100%
SPECIALIZED COMMUNICATION	29%	71%

Table 16 shows that for Broad Commercial Texts, Political Speeches and Specialized Communication, there seems to be a strong preference for more endogenous lexemes if there is no trigger term in the source text. Journalistic Texts also contain more endogenous lexemes than loanwords, but the difference is substantially smaller (53% versus 47%). Although further research is needed to corroborate the following potential explanation, this overall preference in all genres for more endogenous lexemes could be interpreted as evidence of the fact that the more endogenous lexemes are more easily used (and more accepted) than their alternatives, viz. the loanwords.

To summarize the results with respect to this additional factor, i.e. source text lexeme, we could see that the translation strategy depends on the genre if the source text contains a trigger term. More specifically, Broad Commercial Texts and Journalistic Texts tend to borrow the source lexeme and use a loanword in the translations, whereas Political Speeches and Specialized Communication tend to use a more standardized option, which results in a more endogenous lexeme in the translation. If the source text does not contain a trigger word, we could see that for all genres, the more endogenous options are used more often than the loanwords. Journalistic Texts, however, show a tendency to use a loanword almost as often as a more endogenous lexeme, even when there is no trigger word in the source text.

In general, the results obtained by this additional step, can easily be linked to the genre-related results presented in Section 4.2.1, which showed that Specialized Communication and Political Speeches make more use of more endogenous lexemes, whereas Journalistic Texts and Broad Commercial Texts are situated much closer to the loanwords. Moreover, the fact that these last two genres are more susceptible to trigger words in the source text, could be an explanation for their overall position in the biplot shown in Figure 15.

4.3 Conclusion

This case study was carried out so as to examine the norm-conforming behavior in various genres in translation by analyzing the dispersion of more endogenous lexemes versus loanwords. More specifically, we wanted to investigate the following hypothesis: "genres in translation conform to descriptive, genre-determined norms and show a similar preference with regard to English loanwords and their interchangeable more endogenous lexemes as their corresponding non-translated genres". The results of the global effects showed that, with regard to the linguistic preference for either a more endogenous lexeme or a loanword, there is no significant difference between Dutch translated from French and non-translated Dutch. There is, however, a significant difference between Dutch translated from French and Dutch translated from English. These two observations show that source language has an effect on the linguistic behavior and that translated texts as such do not behave uniformly.

Given our focus on the choice between an English loanword and a more endogenous lexeme, it was not surprising that Dutch translated from English behaved significantly different from the other two source language varieties in our corpus. However, the distance between DU<EN and the other two source language varieties was not as large as we might have expected and loanwords were only slightly more often used in DU<EN in comparison to DU_orig and DU<FR. Examining the dataset revealed no additional information which could help explain this observation.

We also examined whether the various genres in our corpus display differences in linguistic behavior where the choice between a loanword and a more endogenous lexeme is concerned. Our results with regard to genre variation showed that Broad Commercial Texts and Journalistic Texts, which can be considered broader genres, make more use of loanwords whereas Political Speeches, Legal Texts and Specialized Communication, viz. genres with more specific content, show a preference for more endogenous lexemes.

The main focus of this case study, however, was to verify to what extent the various genres in translation in our corpus conform to the norms established in their non-translated counterparts. Additionally, we wanted to verify whether source language has an influence on this type of descriptive norm-conforming behavior. Therefore, we calculated the interaction effects between these factors, which allowed us to compare the linguistic preferences in, for example, non-translated Journalistic Texts, Journalistic Text translated from French and Journalistic Texts translated from English. The results based on these interaction effects showed that for Broad Commercial Texts, for example, the source language had an influence on the descriptive norm-conforming behavior. More specifically, the norm in non-translated Broad Commercial Texts

resulted in an overall preference for more endogenous lexemes, and was conformed to by Broad Commercial Texts translated from French. Broad Commercial Texts translated from English, however, did not seem to conform to this norm as they displayed a preference for English loanwords. Our hypothesis could therefore not be confirmed for the genre Broad Commercial Texts.

For Journalistic Texts, the non-translated genre did not show a preference for either loanwords or more endogenous lexemes as the genre wavered in between these two clusters. Investigating the influence of the source language revealed that Journalistic Texts translated from English showed a preference for more endogenous lexemes, whereas Journalistic Texts translated from French used loanwords more often. In other words, the source language has an effect on the preference for loanwords versus more endogenous lexemes within this genre which leads us to conclude that our hypothesis could not be confirmed for Journalistic Texts.

Within the remaining genres, our hypothesis could be confirmed as the translations appear to conform to the norm established in their non-translated counterparts. More specifically, Specialized Communication, Political Speeches and Legal Texts show a preference for more endogenous lexemes.

As this particular case study focused on investigating the dispersion of more endogenous lexemes versus English loanwords, it seemed interesting to investigate whether the corresponding lexeme in the source text has an influence on the linguistic preferences. Analyzing this aspect showed that for some genres, viz. Broad Commercial Texts and Journalistic Texts, the presence of a trigger term in the source text led indeed to a higher percentage of borrowing, which then resulted in a loanword in the translation. For other genres, viz. Political Speeches and Specialized Communication, a trigger in the source text did not lead to a higher percentage in loanwords, and the translations contained a higher percentage of endogenous lexemes than loanwords. In other words, the presence of a trigger word in the source text is not a deciding factor in selecting an English loanword in the target text.

Overall, the results presented in this particular case study show that genre as such as well as source language and translation status have an effect on the linguistic behavior and, more particularly, on the preference for either English loanwords or more endogenous lexemes. However, further research is needed to determine what the genre-specific aspects are which could help explain these results as a genre is determined by numerous aspects such as the author's background, an editorial process, written and unwritten stylistic guidelines, etc.

Chapter 5
Case Study III: visualizing the preferences for General Standard Dutch lexemes versus non-standard Dutch lexemes in order to investigate prescriptive norm-conforming behavior

5.1 Hypothesis and variable selection

5.1.1 Hypothesis

With this third case study, we wish to investigate norm-conforming behavior in translated versus non-translated language by verifying whether translated texts conform to prescriptive norms and make more use of standard language than non-translated texts. As mentioned in Chapter 1, it has been suggested that translations are more conventional etc. in comparison with their source texts and, by extension, with comparable non-translated texts. A translation's choice for General Standard Dutch, i.e. the more conventional option, can be interpreted as an example of the translator's hypothesized norm-conforming behavior. Assuming that norm-conforming behavior is indeed typical of translations in general led us to formulate the following hypothesis:

> "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable non-standard Dutch lexeme which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts".

5.1.2 Variable selection

As mentioned in Section 2.1.2.3, the highly specific language situation in Belgium allows for a particularly systematic approach to investigate the prescriptive norm-conforming behavior in translated and non-translated language as there are numerous expressions, lexical items, grammatical constructions, etc. which are labeled by various normative sources as "non-standard Dutch". In other words, investigating the dispersion of General Standard Dutch, i.e. the norm-conform variant, versus non-standard Dutch throughout our corpus will show which varieties conform to the norm and which do not.

A first step in building the profile set for this particular case study consisted of gathering information from various reference works which contain advice on language usage for language users such as *Stijlwijzer* (Van der Horst, 1996), the *ANS* (1997), *Stijlboek De Standaard* (Permentier, 2003), *Liever meer of juist minder?* (Hendrickx et al., 2010), and *Correct Taalgebruik* (Penninckx et al., 2001). In addition to these reference works various articles related to this topic were consulted to find possible candidates for this profile set such as Haeseryn (1996), Van Craenenbroeck (2000), Heylen (2008), and Plevoets (2013). This resulted in a list of possible profile candidates (n = 184) whose normative label was examined for contradictions in a next step.

This process of gathering profile candidates revealed that in various cases, the reference works mentioned above did not agree on whether the linguistic item should be labeled General Standard Dutch, Belgian Standard Dutch or non-standard Dutch. However, we only wanted to use those profiles which consisted of at least one alternative whose status was undoubtedly non-standard Dutch and one alternative whose status was undoubtedly General Standard Dutch. Therefore, we decided to only rely on the labels provided by *Taaladvies*, the database on language usage provided by the Dutch Language Union because of (i) its status of leading organ where advice on language usage is concerned and (ii) its highly consistent and thorough labeling approach. For more details regarding the inner workings of and the sources which are consulted by *Taaladvies*, see Section 2.1.2.3.

In addition to using a consistent methodology where the normative labels for the variants in our profile set are concerned, we wanted to build a profile set which can visualize the dispersion of standard language versus non-standard language without being distorted by other factors. Therefore, we selected profiles whose variants are different from one another only because of their status of standard language or non-standard language and not because of other factors such as their formality level, regional character etc.

Using this more strict approach of applying the *Taaladvies* labels and discarding profiles which contain additional dimensions besides the (non-)standard language status resulted in a significantly smaller set of possible candidates (n = 17). The

attestations for the linguistic alternatives within these profiles were extracted from the Dutch Parallel Corpus in order to be manually validated so as to be sure of the alternatives' interchangeability.

Finally, the manual validation step and a minimal frequency measure per profile and per factor resulted in a dataset of 7 profiles (n=2398) consisting of at least one General Standard Dutch alternative and one non-standard Dutch alternative. An overview of the profile set and the frequencies per genre and (source) language variety is provided in Table 17 below. It should be noted that for the alternatives within the profiles, both the singular and the plural forms were taken into account and the corpus searches were carried out in a case-insensitive manner. The normative label for the linguistic alternatives is either General Standard Dutch (GSD) or non-standard Dutch (NSD).

Table 17 Frequencies of the General Standard Dutch alternatives versus the non-standard Dutch alternatives per genre and (source) language variety

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL	TOURIST		DU_orig	DU<EN	DU<FR		TOTAL
een van de	GSD	ONE OF THE	39	100	45	106	362	52		313	245	146		704
één van de	NSD		37	115	16	165	72	55		202	189	69		460
pv + inf + vd	GSD	VERBAL END GROUP	7	53	11	16	21	1		32	19	58		109
vd + pv + inf	GSD		3	16	10	11	21	1		28	16	18		62
pv + vd + inf	NSD		3	15	1	2	5	1		21	3	3		27
te veel	GSD	TOO MUCH	5	10	12	22	158	4		120	36	55		211
teveel	NSD		3	5	6	14	5	0		24	4	5		33
ten(minste)_goed	GSD	AT LEAST	65	42	5	62	51	1		86	43	97		226
ten(minste)_fout	NSD		5	26	2	8	6	1		21	19	8		48
een beroep doen op	GSD	TO MAKE AN APPEAL TO	18	54	17	51	80	3		87	29	107		223
beroep doen op	NSD		9	7	3	2	7	0		18	3	7		28
zodra	GSD	AS SOON AS	17	31	12	41	103	3		79	37	91		207
van zodra	NSD		5	9	0	1	3	0		12	1	5		18
beginnen + te + inf	GSD	TO START TO	2	1	0	4	19	1		13	6	8		27
beginnen + inf	NSD		0	3	0	1	7	4		9	2	4		15
TOTAL			218	487	140	506	920	127		1065	652	681		2398

In the following paragraphs, a description of the individual profiles and an example is provided. Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013) was used for the concept descriptions. The example sentences were extracted from the Dutch Parallel Corpus and their origin is presented by means of the file number. Each Dutch sentence contains one of the profile's variants which is matched by its translation or source text sentence in English.

Een van de – één van de

Concept: ONE OF THE

Example: DU: ***Een van de*** *gevoeligste punten in de relatie tussen[....]*
EN: **One of the** most sensitive aspects of [...].
dpc-arc-002044-nl

Pv + inf + vd; - vd + pv + inf; - pv + vd + inf
Concept: ORDER WITHIN THE VERBAL END GROUP
Example: DU: *Ik denk niet dat media-educatie op school* ***zal worden gegeven.***
EN: I don't think media education **will be done** in schools.
dpc-vla-001920-nl

Te veel - teveel
Concept: TOO MUCH
Example: DU: *Er is* ***te veel*** *energie nodig om een project binnen te halen.*
EN: **Too much** energy is needed just to get a project.
dpc-vla-001920-nl

Ten minste (correctly used) - ten minste (erroneously used)
Concept: AT LEAST
Example: DU: [...] **ten minste** op korte termijn [...]
EN: [...] **at least** in the short term [...]
dpc-bco-002536-nl

Een beroep doen op – beroep doen op
Concept: MAKE AN APPEAL TO
Example: DU: *Het doet daarbij* ***een beroep op*** *de economische sectorclassificatie van MSCI [...].*
EN: For that purpose it **calls on** MSCI's economic sector classification [...].
dpc-ing-001880-nl

Zodra - van zodra
Concept: AS SOON AS
Example: DU: *Financieringskosten worden opgenomen* ***zodra*** *ze worden gemaakt.*
EN: Financing costs are charged against the income statement **as soon as** they are incurred.
dpc-gim-002266-nl

Beginnen te + inf; - beginnen + inf
Concept: TO START TO

Example: DU: *Ik zeg niet dat een sociale jetlag de enige reden is waarom mensen* ***beginnen te roken*** *[...].*
EN: I'm not saying social jet lag is the only reason people **take up smoking** [...].
dpc-ind-001797-nl

5.2 Results

5.2.1 Global effects of genre and translation

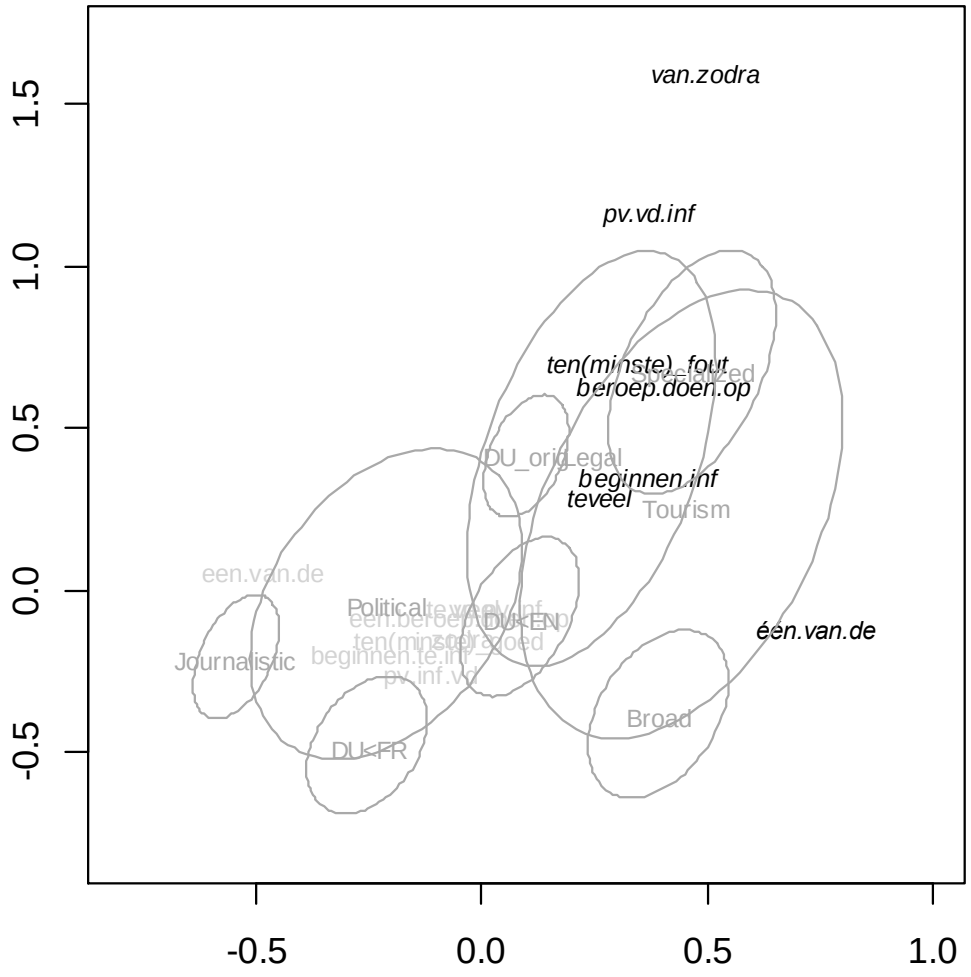

Figure 22 Biplot of the global results with the General Standard Dutch lexemes (light grey), the non-standard Dutch lexemes (black italics), the genres, and the (source) language varieties.

Figure 22 contains the overall results for this analysis and this plot shows the linguistic variants for our profile set as data points, with light grey data points (e.g. `op.het.eerste.gezicht`) which are General Standard Dutch and black italic data points (e.g. `op.het.eerste.zicht`) which are non-standard Dutch. This plot shows a clear division between the non-standard Dutch lexemes on the right to higher right side and the General Standard Dutch lexemes grouped together in the lower left corner

of the plot. This plot also reveals that the General Standard Dutch variants behave more uniformly than the non-standard Dutch alternatives, which seem to have a wider distribution over the right side of the plot.

The location of these two clouds provides us with information regarding the linguistic preferences within the various genres and (source) language varieties in our corpus. A confidence ellipse was computed for every genre and (source) language variety in the dataset and given the fact that we calculated the distances between nine language varieties (six genres, Belgian Dutch translated from French, Belgian Dutch translated from English and non-translated Belgian Dutch), there is a total of nine ellipses to be found in this plot, the relations between all of which can be observed.

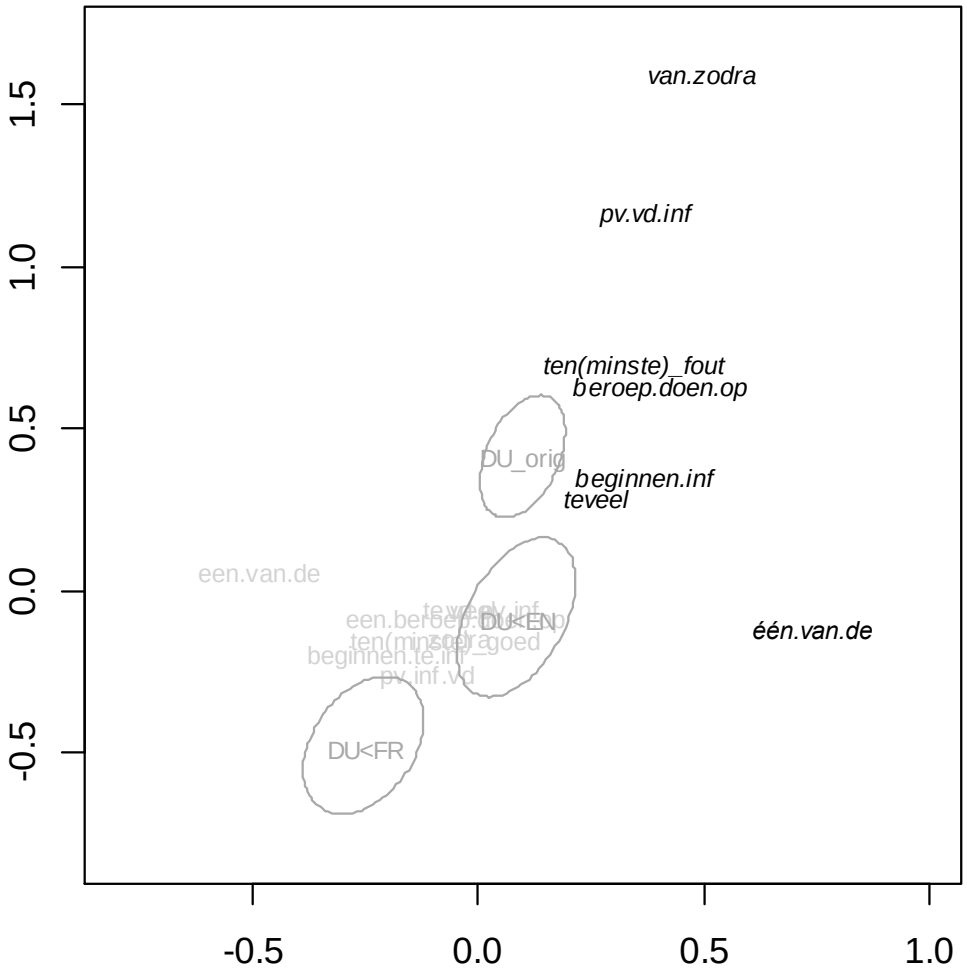

Figure 23 Biplot of the (source) language varieties, the General Standard Dutch alternatives (light grey) and the non-standard Dutch alternatives (black italics)

Figure 23 is proportionally the same as Figure 22, but it only shows the results for the (source) language varieties in our corpus, viz. Dutch translated from French (DU<FR), Dutch translated from English (DU<EN) and non-translated Dutch (DU_orig), as plots with a lower information density allow us to interpret the results more easily. This biplot clearly shows that the distance between all three varieties is statistically significant, as their ellipses do not overlap. Moreover, we can see that DU<EN and DU<FR are situated in the lower left corner whereas DU_orig leans towards the upper right corner. If we compare the position of the varieties with regard to the two clusters in the plot, this leads to the following observations: DU<FR and DU<EN show a preference for General Standard Dutch, an effect which is even stronger for DU<FR than for DU<EN, as DU<FR is situated even further away from the non-standard Dutch variants than DU<EN. DU_orig on the other hand, seems to show a preference for non-standard Dutch, as it is situated towards the non-Standard Dutch alternatives and further away from the General Standard Dutch alternatives. In other words, this plot reveals a significant difference between translated Dutch and non-translated Dutch where the use of General Standard Dutch is concerned.

We initially formulated the following hypothesis: "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable non-standard Dutch lexeme, which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts". The results shown in Figure 23 reveal an overall preference for General Standard Dutch for both DU<FR and DU<EN, which confirms our hypothesis. However, there is an additional significant difference between Dutch translated from French and Dutch translated from English, which shows that source language seems to have an influence as well.

In order to verify whether there were any particularities besides source language and translation status which may help to explain the results shown in Figure 23, we investigated the dataset. For DU<EN, examining the data revealed no additional phenomena. For example, for the profile *te veel - teveel* (CONCEPT: TOO MUCH), both *te veel* and *teveel* are translations for "too much" or "too many" fairly often, so one cannot consider "too much" or "too many" as triggers for the Standard Dutch *te veel*, which also consists of two words.

Looking at the data for DU<FR, one could suspect that for the profile *een beroep doen op - beroep doen op* (CONCEPT: TO MAKE AN APPEAL TO) the translations are biased towards the use of *beroep doen op*, as this is a literal translation from the French *faire appel à*. However, in the DU<FR sub-corpus, *faire appel à* was used 41 times in the source text and only translated as *beroep doen op* 3 times, whereas it was translated as *een beroep doen op* 38 times. As for the other profiles, no direct link was found in the source texts. The example of *faire appel à,* however, is remarkable, as this indicates that translations from French into Dutch show a clear preference for General Standard Dutch, even

though some source language alternatives could be translated as non-standard Dutch more easily. Although this would require further, in-depth research, these observations could indicate a certain linguistic awareness among translators who translate from French into Dutch and a tendency to conform to the norm and use General Standard Dutch instead of Gallicisms. Even more tentatively, this could be linked to their educational background, as they may have been explicitly trained to avoid these Gallicisms in translations.

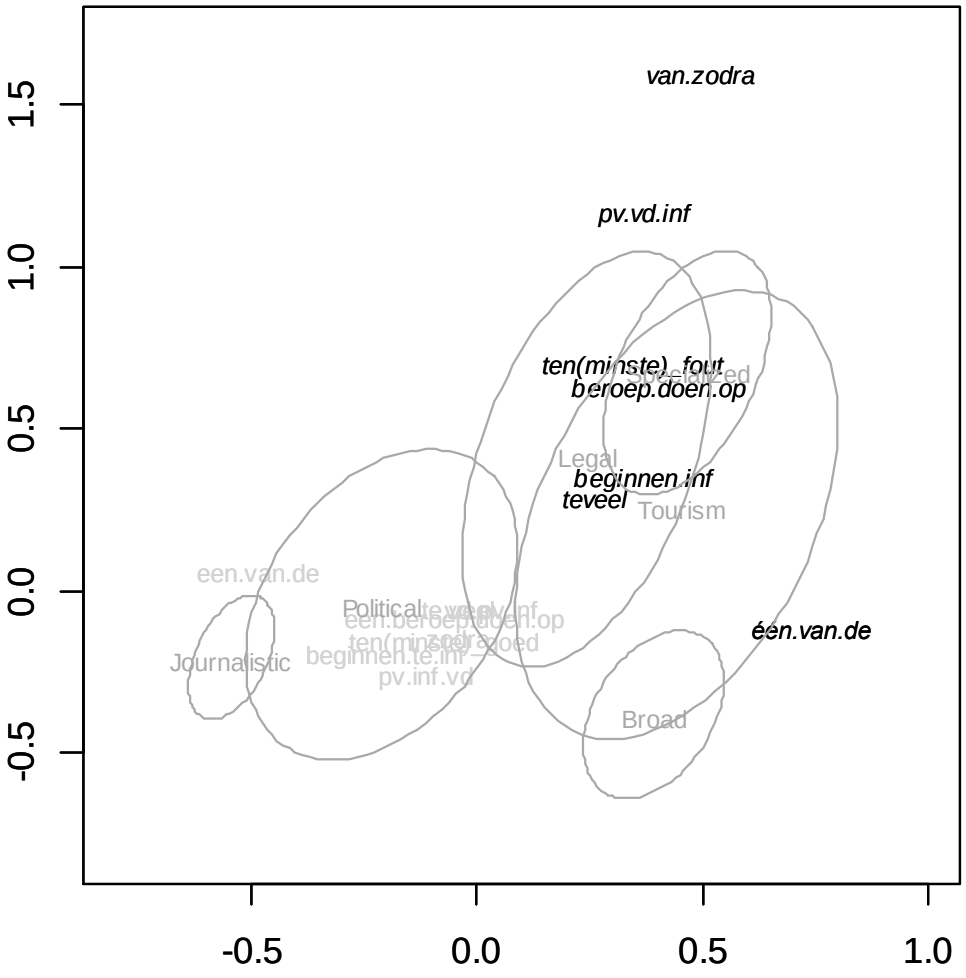

Figure 24 Biplot of the genres, the General Standard Dutch alternatives (light grey) and the non-standard Dutch alternatives (black italics)

Figure 24 only shows the results for the six genres, viz. Legal Texts (`Legal`), Specialized Communication (`Specialized`), Political Speeches (`Political`), Broad Commercial Texts (`Broad`), Journalistic Texts (`Journalistic`) and Tourist Information (`Tourism`). This plot reveals significant differences between some genres (e.g. between

Specialized Communication and Journalistic Texts), but not between others (e.g. between Broad Commercial Texts and Tourist Information). More specifically, we can observe a clear distinction between Specialized Communication, which shows a preference for non-standard Dutch, and Journalistic Texts, which tend to prefer General Standard Dutch. We can also see that the ellipses for Tourist Information, Legal Texts and Political Speeches are rather large, which is due to the fact that there are little data available for these genres, which leads to larger statistical uncertainties.

Summarizing this plot, we can see that Specialized Communication and, albeit to a lesser degree, Legal Texts and Tourist Information, contain more non-standard language in comparison with Journalistic Texts and Political Speeches, which show a preference for General Standard Dutch. Broad Commercial Texts are located somewhere in between the two clusters and show no clear preference for either standard or non-standard language.

Although additional research is needed, we would like to formulate some explanatory hypotheses with regard to these results. Journalistic Texts, for example, appear to often use General Standard Dutch, which is not entirely surprising as Journalistic Texts are written by journalists, who often received language-related training at some point. Furthermore, this is a genre which is likely to contain an editorial step in its production process often with the goal of publishing text material which is linguistically homogeneous, a goal which can partially be achieved by using standard language. Kruger refers to one of these editorial steps as copyediting which she describes as "the correction of a manuscript to ensure that it conforms to certain rules, which include grammar and spelling rules, usage rules, and the publisher's house style" (to appear). For example, the Belgian newspaper *De Standaard* which is one of the text providers for the Dutch Parallel Corpus published its own style guide, *Stijlboek De Standaard* (Permentier, 2003), which provides advice with regard to language usage. The remark with regard to an editorial step also applies to Political Speeches, which might explain this genre's preference for standard language as well. As the genre's descriptive title suggests, Broad Commercial Texts is a rather broad genre and it contains a high number of text providers (n=19). Contrary to Journalistic Texts, which is more homogeneous genre, the high number and variety of providers and the corresponding high variety of texts may explain why the more heterogeneous genre Broad Commercial Texts wavers somewhere in between General Standard Dutch and non-standard Dutch.

To summarize, analyzing the data did not reveal any additional factors which could explain the differences between the various genres in the dataset and apart from the possible influence of the editorial process which applies for Journalistic Texts and Political Speeches, no additional plausible explanatory factors could be determined based on the data set within this type of research.

5.2.2 Interaction between source language, translation status and genre

The previous plots only show the main effects and provide no information with regard to the differences or similarities between the various translated and non-translated genres in our dataset. Therefore, we calculated the interactions between the genres and the (source) language varieties so as to establish whether translations of a specific genre are more inclined to use standard language than others. Additionally, the interaction plots which are described in the following paragraphs can provide more insight into the influence of the source language by answering questions such as "Do Journalistic Texts that were translated from English behave differently from Journalistic Texts that were translated from French?".

5.2.2.1 Broad Commercial Texts

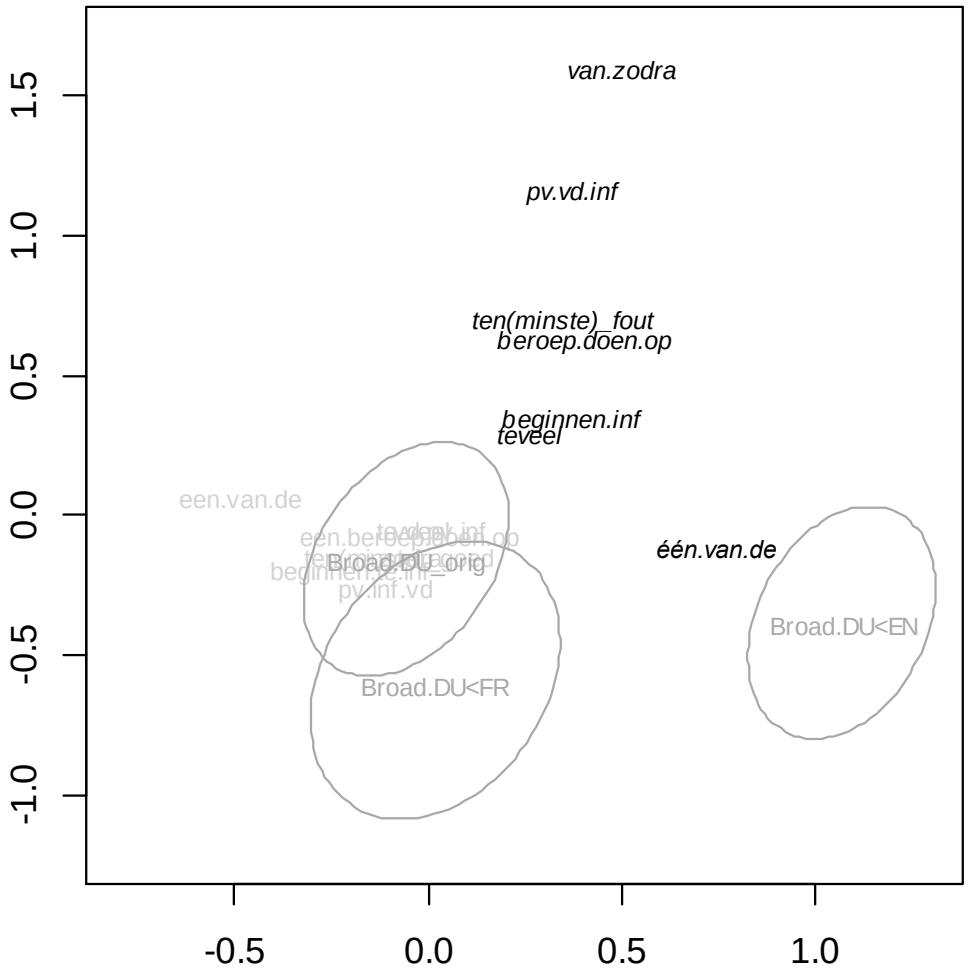

Figure 25 Case Study III: Biplot of the interaction between Broad Commercial Texts; and source language and translation status

Figure 25 shows the results of the interaction effects for Broad Commercial Texts which are visualized by means of three ellipses, viz. Broad Commercial Texts translated from English (`Broad.DU<EN`), Broad Commercial Texts translated from French (`Broad.DU<FR`) and non-translated Broad Commercial Texts (`Broad.DU_orig`). The results in the biplot show that there is a significant difference between `Broad.DU<EN` on the one hand and `Broad.DU<FR` and `Broad.DU_orig` on the other hand, but not between `Broad.DU<FR` and `Broad.DU_orig`. More specifically, this plot reveals that `Broad.DU<FR` and `Broad.DU_orig` are situated much closer to the General Standard Dutch variants whereas `Broad.DU<EN` shows a slight preference for non-

standard language variants. In other words, for this particular genre, source language seems to have an influence on the preference for either General Standard Dutch or non-standard Dutch.

5.2.2.2 Specialized Communication

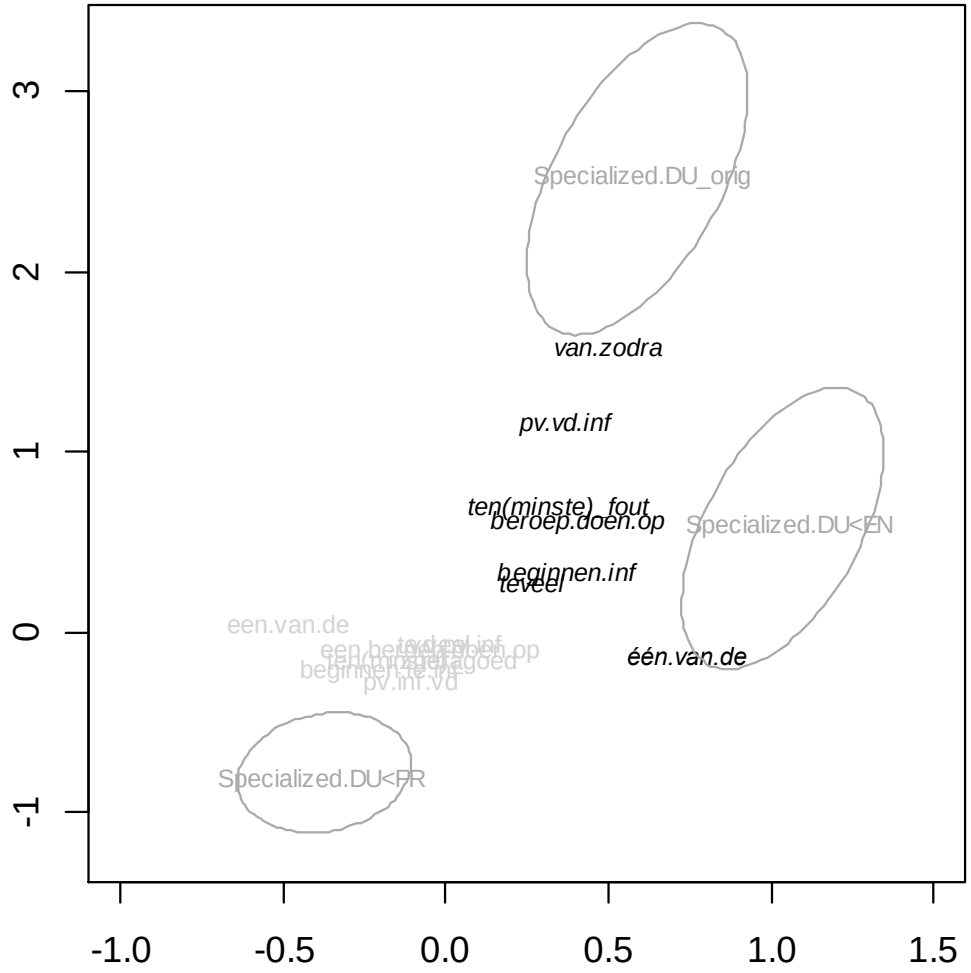

Figure 26 Case Study III: Biplot of the interaction between Specialized Communication; and source language and translation status

As can be seen in Figure 26, which shows the interaction results for Specialized Communication by means of confidence ellipses for Specialized Communication translated from French (`Specialized.DU<FR`), Specialized Communication translated from English (`Specialized.DU<EN`) and non-translated Specialized Communication (`Specialized.DU_orig`), there are significant differences between these three sub-varieties of Specialized Communication. In other words, this is a genre

which behaves differently depending not only on whether it is translated or not but also on the source language. More specifically, this plot reveals that `Specialized.DU<EN` and `Specialized.DU_orig` show a preference for non-standard Dutch variants, whereas `Specialized.DU<FR` makes more use of General Standard Dutch variants. Moreover, the position of `Specialized` in Figure 24 showed an overall preference for non-standard lexemes within this genre. It therefore seems that French as source language has a strong influence on the norm-conforming behavior within this genre which results in a strong preference for General Standard Dutch for `Specialized.DU<FR`.

5.2.2.3 Journalistic Texts

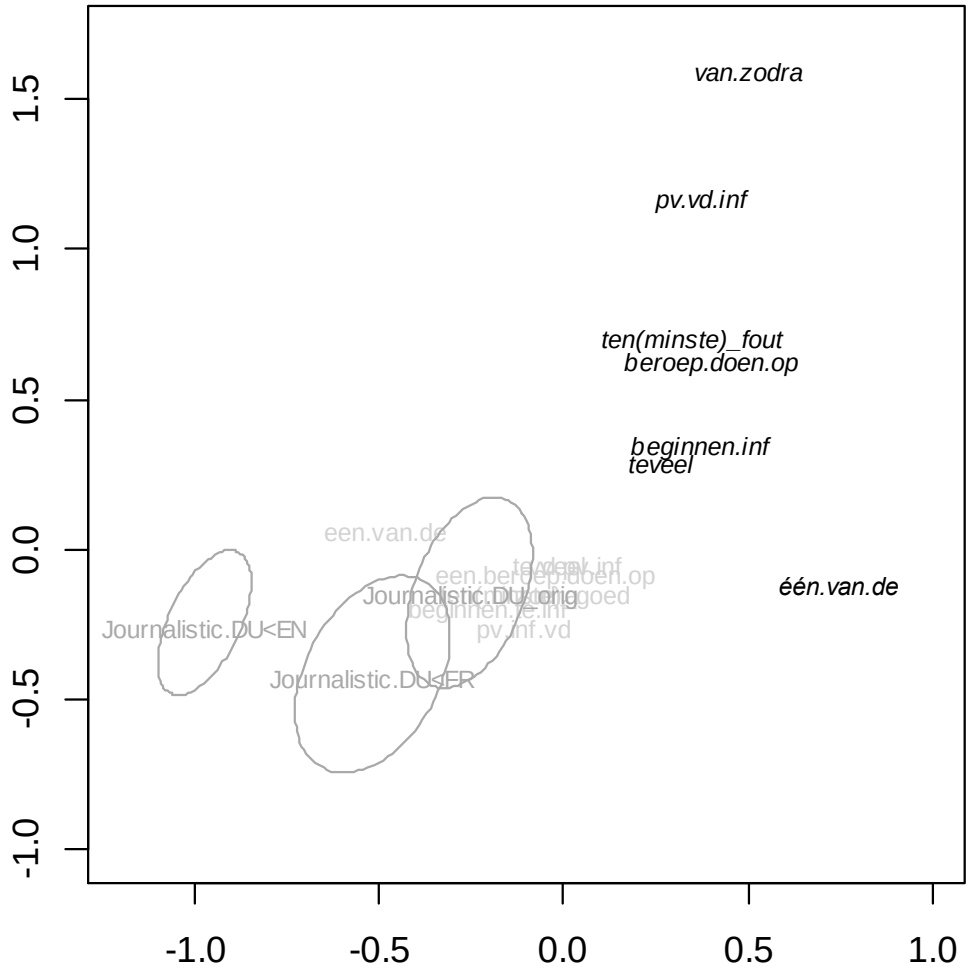

Figure 27 Case Study III: Biplot of the interaction between Journalistic Texts; and source language and translation status

Figure 27 immediately shows a strong genre effect: the various ellipses for Journalistic Texts, viz. Journalistic Texts translated from French (Journalistic.DU<FR), Journalistic Texts translated from English (Journalistic.DU<EN) and non-translated Journalistic Texts (Journalistic.DU_orig) are grouped closely together. Moreover, there is no significant difference between Journalistic.DU<FR and Journalistic.DU_orig. Put differently, this is a genre which is less susceptible to the influence of the source language or whether its texts are translated or not and this genre conforms to the norm which results in a strong, overall preference for General Standard Dutch. As mentioned in Section 5.2.1, this consistent linguistic behavior of Journalistic Texts might be due to aspects such as the editorial process or the linguistic background of individual authors and translators.

5.2.2.4 Tourist Information

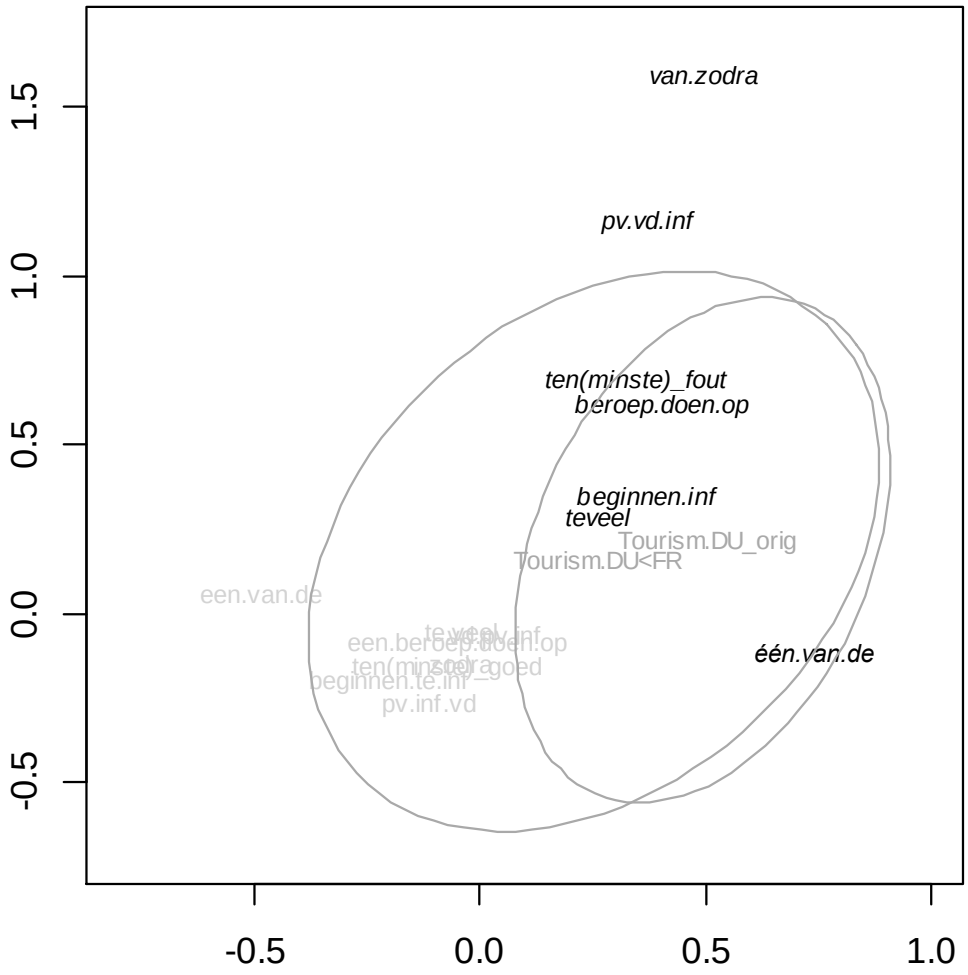

Figure 28 Case Study III: Biplot of the interaction between Tourist Information; and source language and translation status

In Figure 28, the interaction results between translation status and Tourist Information are shown in a biplot. As can be seen, there are only two ellipses available, viz. `Tourism.DU_orig` and `Tourism.DU<FR`, which is due to the fact that our corpus does not contain data for Tourist Information translated from English. This plot shows that there are no significant differences between the two sub-varieties, the ellipses for which are rather large which is due to the fact that there is little data available for these sub-varieties. Additionally, both sub-varieties appear to waver somewhere in between the standard language alternatives and the non-standard language alternatives. The results for Tourist Information thus point towards the conclusion that there is a genre effect at play as both sub-varieties behave homogeneously and translation status has no

influence on the linguistic choices that were made. More particularly and contrary to texts translated from French (see Figure 24), `Tourism.DU<FR` shows no clear preference for General Standard Dutch alternatives.

5.2.2.5 Political Speeches

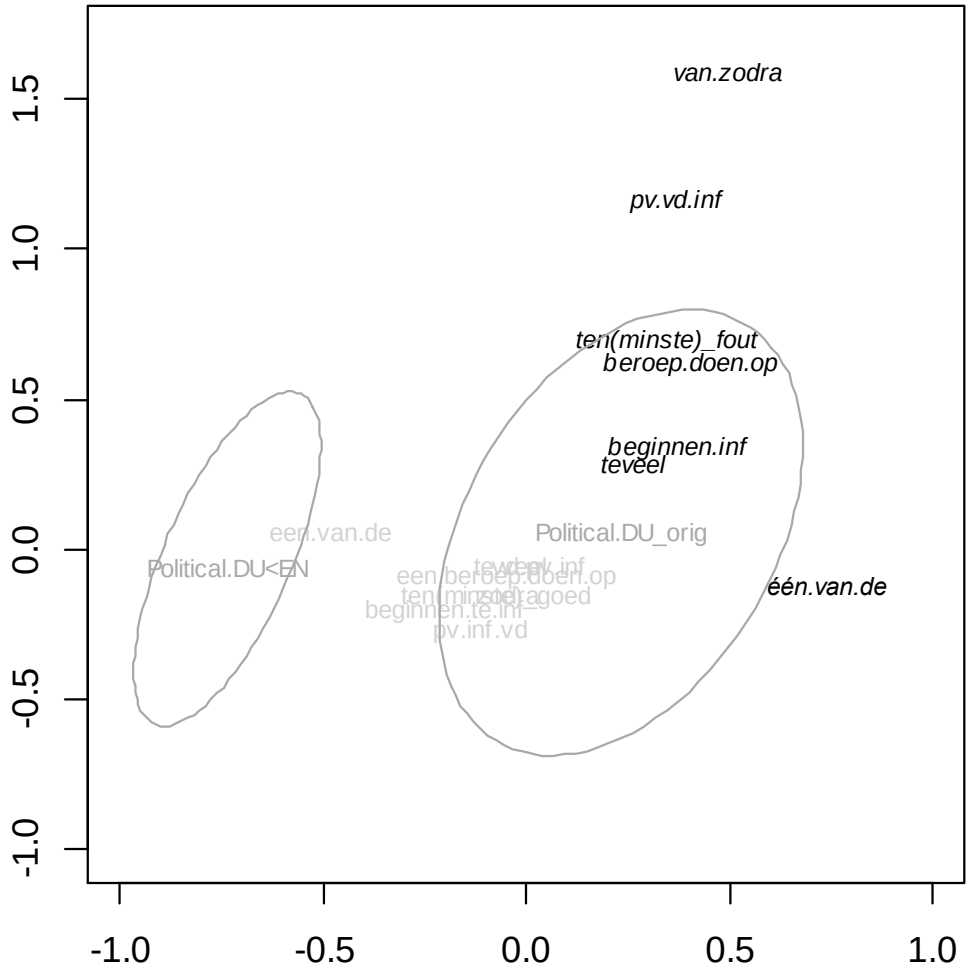

Figure 29 Case Study III: Biplot of the interaction between Political Speeches; and source language and translation status

As was the case for Tourist Information, the DPC only contains text material for two of the three possible sub-varieties of Political Speeches, viz. Political Speeches translated from English (`Political.DU<EN`) and non-translated Political Speeches (`Political.DU_orig`). The results of the interaction effects for Political Speeches are presented in Figure 29 and show that there is a significant difference between `Political.DU_orig` and `Political.DU<EN`. It appears that

Political.DU_orig show no clear preference for either non-standard Dutch or General Standard Dutch, whereas Political.DU<EN is situated much closer to the General Standard Dutch lexemes. Our hypothesis for this particular genre is confirmed as Political Speeches in translation conform to the norm and show a preference for standard language.

5.2.2.6 Legal Texts

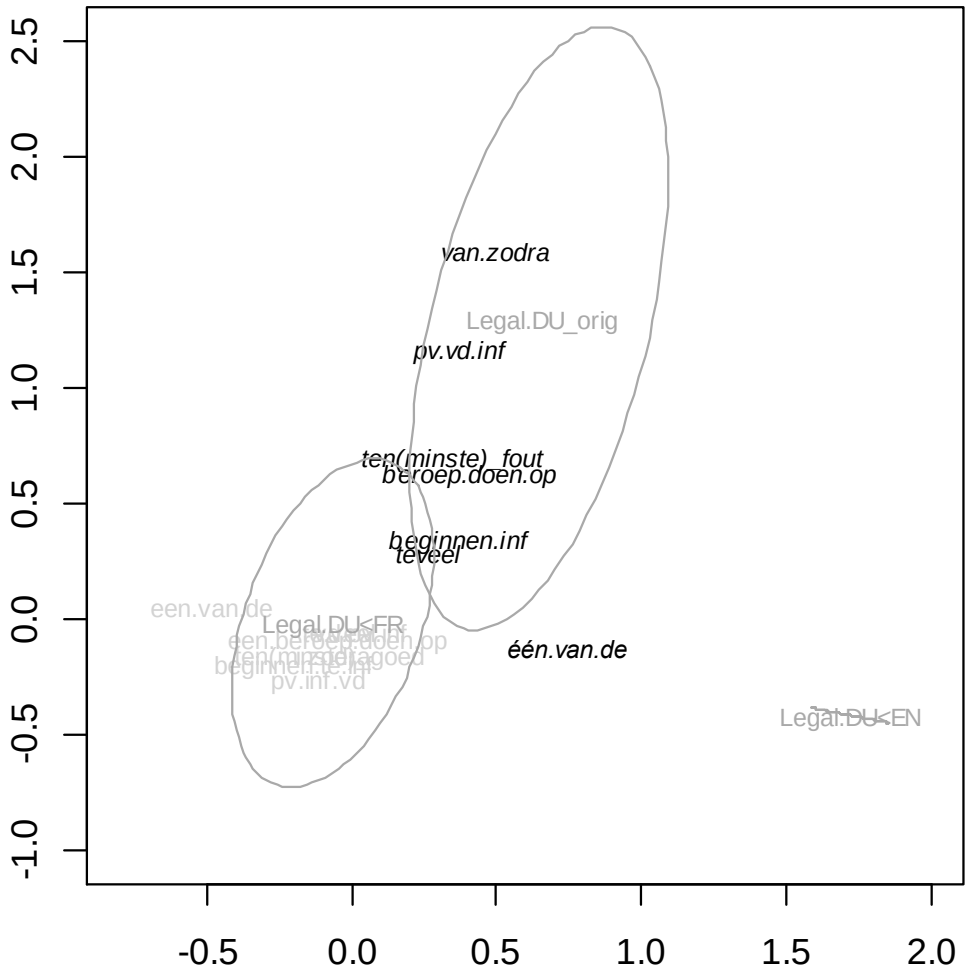

Figure 30 Case Study III: Biplot of the interaction between Legal Texts; and source language and translation status

Figure 30 shows the results of the interaction effect between Legal Texts and source language and translation status through the ellipses for Legal Texts translated from French (Legal.DU<FR), Legal Texts translated from English (Legal.DU<EN) and non-translated Legal Texts (Legal.DU_orig). The plot shows that there are no

significant differences between Legal.DU<FR and Legal.DU_orig, but there is a significant difference between Legal.DU<EN and Legal.DU<FR and Legal.DU_orig. The highly specific and remarkable position of Legal.DU<EN led us to take a closer look at the dataset, which revealed that the sub-variety Legal.DU<EN contains only one data point, i.e. *één van de.* As a single data point cannot result in any general observations, we decided not to take Legal.DU<EN into account when discussing the results the interaction effects in Legal Texts.

For the remaining sub-varieties, the following observations could be made: although the difference between Legal.DU_orig and Legal.DU<FR is not significant, Legal.DU_orig shows a preference for non-standard Dutch while Legal.DU<FR appears to make more use of General Standard Dutch. These results show that for Legal Texts the hypothesis could be confirmed as translated Legal Texts conform to the norm by displaying a preference for General Standard Dutch.

5.2.2.7 Interaction results: conclusion

The interaction plots revealed that for some genres the hypothesis could be confirmed as translated texts display a preference for General Standard Dutch (Journalistic Texts, Political Speeches and Legal Texts. For Broad Commercial Texts and Specialized Communication, there was a (partial) source language effect at play which causes these genres to behave differently depending on the source language and the hypothesis could only partially be confirmed. More specifically, when translated from English these genres show a preference for non-standard Dutch lexemes, instead of the hypothesized General Standard Dutch alternatives. Finally, for Tourist Information, the hypothesis was rejected entirely as Tourist Information in translation shows no preference for General Standard Dutch.

5.3 Conclusion

First of all, the main plot which visualized our dataset showed a separate cluster for General Standard Dutch on the one hand and a cluster for non-Standard Dutch on the other hand. The position of the various genres and (source) languages varieties with regard to these clusters reveals whether the standard or the non-standard option is preferred within that genre or (source) language variety in comparison to the other genres and varieties. Focusing on source language variety and translation status only, so abstracting from the genre influence, we could see that there is a significant difference between translated texts and non-translated texts and that translations do indeed make

more use of standard language than non-translations, which confirms our hypothesis. However, the results revealed an additional source language effect as Dutch translated from English is situated closer to the non-standard Dutch alternatives than Dutch translated from French.

With regard to the dispersion of standard versus non-standard language use in various genres, the profile-based correspondence analysis revealed that some genres prefer standard language (Journalistic Texts and Political Speeches), a result which might be caused by the fact that these two genres often undergo an editorial review, although further research is needed to confirm this hypothesis. Other genres seem to make more use of non-standard language (Specialized Communication, Tourist Information and Legal Texts), while Broad Commercial Texts is a genre which does not show a clear preference for either standard or non-standard language. These results clearly show that the factor genre, and all underlying genre-specific aspects, does indeed have an influence on the linguistic preferences of both translators and authors of non-translated language.

Because we also wanted to determine whether there are differences between, for example, Legal Texts which were translated from French, Legal Texts which were translated from English and Legal Texts which were not translated, we calculated the interaction effects between genre and source language and translation status. These interaction plots revealed that in some cases, there is a clear genre effect and the various sub-varieties are not significantly different from one another while for other genres, a source language effect could be observed. For example, Specialized Communication translated from French behaves significantly different from Specialized Communication translated from English.

In other words, and similar to the results of the other case studies, this case study shows that both genre and source language play a role in the linguistic preferences of translations and non-translations.

Chapter 6
Case Study IV: visualizing the preferences for General Standard Dutch lexemes versus Belgian Standard Dutch lexemes in order to investigate prescriptive norm-conforming behavior

6.1 Hypothesis and variable selection

6.1.1 Hypothesis

Similar to the case study presented in Chapter 5, this case study investigates whether different linguistic preferences can be detected between translated and comparable non-translated text material. Given the specific language situation and history in Belgium, we decided to investigate norm conformity by examining the dispersion between General Standard Dutch and Belgian Standard Dutch in our corpus. Chapter 2 provides more details with regard to these two language varieties and the differences between them.

In this final case study, we wanted to verify whether the assumed norm-conforming behavior which is typical of translations is equally strong when investigated across a range of genres. In addition to genre, we also investigated the influence of source language on the linguistic differences between three language varieties, viz. non-translated Belgian Dutch, Belgian Dutch translated from French and Belgian Dutch translated from English.

To summarize, the goal of this case study was to find out: (1) whether symptoms of norm conformity occur to similar degrees in these three language varieties, or whether, instead, the influence of a particular genre and/or a particular source language is

stronger than that of others; (2) whether translation status, i.e. the question of whether a text is translated or not, is a determining factor.

Combining earlier research on normalization based on Toury (1995) with the particular Belgian language situation has led us to formulate a highly specific hypothesis on the norm-conforming behavior of Belgian authors and translators:

> "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable Belgian Standard Dutch lexeme, which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts."

6.1.2 Variable selection

As the linguistic background in Belgium is particularly important to this case study, some background information is provided here, while a more detailed survey can be found in Chapter 1 and Chapter 2. Dutch was not an official language in Belgium until the 19th century and only from the 1960s onwards did the Dutch-speaking area gain political autonomy, which accelerated the standardization process of its own language as the importance of defending it against French declined. The official policy was to opt for a near complete adoption of the Netherlandic Dutch standard, which led to a language policy that sought to clear the language of Belgian variants and have them replaced by the official Standard (Netherlandic) Dutch variants.

This resulted in a series of reference works, regular features in newspapers and even television shows whose main message was "do not say x, but y" in order to 'purify' the language and to create a common standard language. However, matters are less rigid today and there is room for both Belgian Standard Dutch and Netherlandic Standard Dutch varieties alongside General Standard Dutch and advice on language usage is less discriminative towards the Belgian Dutch alternatives. So as to investigate the dispersion of these two varieties of Dutch, we required profiles that consist of at least one lexical item labeled as General Standard Dutch (e.g. *telkens als*) and one lexical item labeled as Belgian Standard Dutch (e.g. *telkens*).

Various sources were consulted to find inspiration for possible profile candidates such as (i) *Vlaams-Nederlands Woordenboek* (Bakema, 2004), a reference dictionary which summarizes the differences between Belgian Dutch and General Standard Dutch, (ii) *Referentiebestand Belgisch-Nederlands*[1], a file issued by the Dutch Language Union which contains approximately 4000 words and expressions which are typical of Belgian Dutch,

[1] http://bob.inl.nl/RBBN/

(iii) *Taaltelefoon*[2], a service offered by the Flemish government which provides advice on language usage, (iv) *VRT Taal*[3], a database on language usage issued by the Flemish public broadcasting company, and (v) *Taaladvies*[4], a database on language usage issued by the Dutch Language Union. This first step in the process resulted in a list of lexical items which received the label Belgian Standard Dutch. Additionally, we wanted to assure the correctness of the label that was given to these lexemes. Therefore, we compared the information from the various sources and ignored those possible candidates for which there was no unanimous label available. This step in the process revealed that the methodology used by *Taaladvies* was methodologically sound and highly consistent which resulted in a thorough labeling approach. *Taaladvies* is a regularly updated online database developed by the *Nederlandse Taalunie* (the Dutch Language Union) which provides linguistic information regarding lexical, morphological, and syntactical items and labels these items as General Standard Dutch, Belgian Standard Dutch, Netherlandic Standard Dutch or non-standard Dutch. The labels are decided upon and approved by the *Taaladviesoverleg*, a team of language experts both from the Netherlands and from Belgium, guaranteeing their reliability. In Chapter 2 a detailed analysis of the step-by-step procedure which is used by *Taaladvies* is provided. Because of its thorough approach and the fact that the labels are based on multiple sources, it was decided to use *Taaladvies* as a single source so as to determine the normative label per lexical item in our profile candidate collection.

In a next phase, we verified whether a General Standard Dutch counterpart existed for these Belgian Standard Dutch items as this is a key requirement for our profile-based approach. The attestations for both linguistic alternatives within these profiles were extracted from the Dutch Parallel Corpus in order to be manually validated so as to be sure of the alternatives' interchangeability. The entire process in combination with the manual validation step and a minimal frequency measure per profile and per factor resulted in a dataset of 9 profiles (n=1494), an overview of which can be found in Table 18, which shows the frequencies per genre, per (source) language variety and in total. The normative labels GSD (General Standard Dutch) and BSD (Belgian Standard Dutch) are based on the information provided by the reference site by the Dutch Language Union.

Table 18 Frequencies of the General Standard Dutch alternatives versus the Belgian Standard Dutch alternatives per genre and (source) language variety

[2] http://taaltelefoon.vlaanderen.be/nlapps/default.asp

[3] http://www.vrt.be/taal/taaldatabank

[4] http://taaladvies.net/

Option	Label	Concept	LEGAL	SPECIAL	POLITICAL	BROAD	JOURNAL		DU_orig	DU<EN	DU<FR		TOTAL
geraken	BSD	TO GET	0	7	9	12	97		65	35	25		125
raken	GSD		0	15	11	14	199		127	61	51		239
luik	BSD	PART	3	22	5	4	10		18	5	21		44
onderdeel	GSD		12	35	19	45	47		83	42	33		158
tewerkstelling	BSD	EMPLOYMENT	1	3	7	37	13		19	2	41		62
werkgelegenheid	GSD		12	30	55	23	11		60	38	33		131
ten laatste	BSD	AT THE LATEST	21	23	1	8	6		24	8	27		59
uiterlijk	GSD		86	15	1	8	1		73	4	34		111
verwittigen	BSD	TO WARN	4	6	0	0	5		8	0	7		15
waarschuwen	GSD		1	17	6	7	61		30	43	19		92
vragen + dat	BSD	TO DEMAND	5	46	7	13	13		41	7	36		84
eisen + dat	GSD		5	8	5	3	13		9	13	12		34
proper	BSD	CLEAN	4	0	1	2	12		9	3	7		19
schoon	GSD		1	1	10	9	18		23	14	2		39
telkens	BSD	EACH TIME	2	4	0	2	14		9	3	10		22
telkens als	GSD		2	7	3	1	15		10	11	7		28
eenmaal	BSD	AS SOON AS	0	9	4	1	23		20	10	7		37
zodra	GSD		11	29	12	40	103		76	37	82		195
TOTAL			170	277	156	229	661		704	336	454		1494

In the following paragraphs, a description of the individual profiles and an example is provided. The Van Dale's Great Dictionary of the Dutch Language (den Boon & Geeraerts, 2010-2013) was used for the concept descriptions. The example sentences were extracted from the Dutch Parallel Corpus and their origin is presented by means of the file number. Each Dutch sentence contains one of the profile's variants which is matched by its translation or source text sentence in English.

Geraken – raken

Concept: TO GET - "to change to a certain situation or state"
Word class: verb
Example: DU: [...] **raakt** het luchtruim steeds meer verzadigd.
EN: [...] the airspace **gets** more and more congested.
dpc-bco-002374-nl

Luik – onderdeel

Concept: PART - "a part of something"
Word class: noun
Example: DU: Twinning is nu een integraal **onderdeel** van de manier [...].
EN: Twinning is now an integral **part** of how [...].
dpc-arc-002046-nl

Tewerkstelling – werkgelegenheid

Concept: EMPLOYMENT - "employment opportunities or prospects"
Word class: noun
Example: DU: [...]voor de voedselveiligheid en de **werkgelegenheid** in de voedingsmiddelensector.
EN: [...]for food security and **employment** in the agricultural food sector.
dpc-erp-000449-nl

Ten laatste – uiterlijk

Concept: AT THE LATEST – "not later than"
Word class: adverb
Example: DU: Over **ten laatste** drie weken zullen de eerste nieuw geproduceerde projectoren kunnen worden uitgeleverd.
EN: **Within** three weeks the first newly finished projectors will be delivered.
dpc-bco-002348-nl

Verwittigen – waarschuwen

Concept: TO WARN - "to warn someone about something"
Word class: verb
Example: DU: [...]een alarmsysteem hadden geïnstalleerd om hen te **waarschuwen** [...]
EN: [...]they had arranged a warning system to **alert** them [...].
dpc-ind-001643-nl

Vragen + dat – eisen + dat

Concept: TO DEMAND - "to request something"
Word class: verb
Example: DU: [...]daarom **vragen** ze **dat** de wereld hun geaardheid op dezelfde manier erkent.
EN: [...]and they are **demanding** that the world recognises their orientation in the same way.
dpc-ind-001825-nl

Proper – schoon

Concept: CLEAN - "free of dirt"
Word class: adjective
Example: DU: [...]dat ze ergens belanden waar het **proper** en veilig is [...].
EN: [...]the place will be **clean** and safe [...].

dpc-ind-001748-nl

Telkens – telkens als

Concept: EACH TIME - “in each case”
Word class: adverb
Example: DU: **Telkens als** Cohen het werk verplaatst [...].
EN: **Every time** Cohen moves the work [...].
dpc-ind-001934-nl

Eenmaal – zodra

Concept: AS SOON AS - “immediately after”
Word class: adverb
Example: DU: **Eenmaal** uit de gevangenis ontslagen [...].
EN: **Once** out of jail [...].
dpc-ind-001643-nl

6.2 Results

6.2.1 Global effects of genre and translation

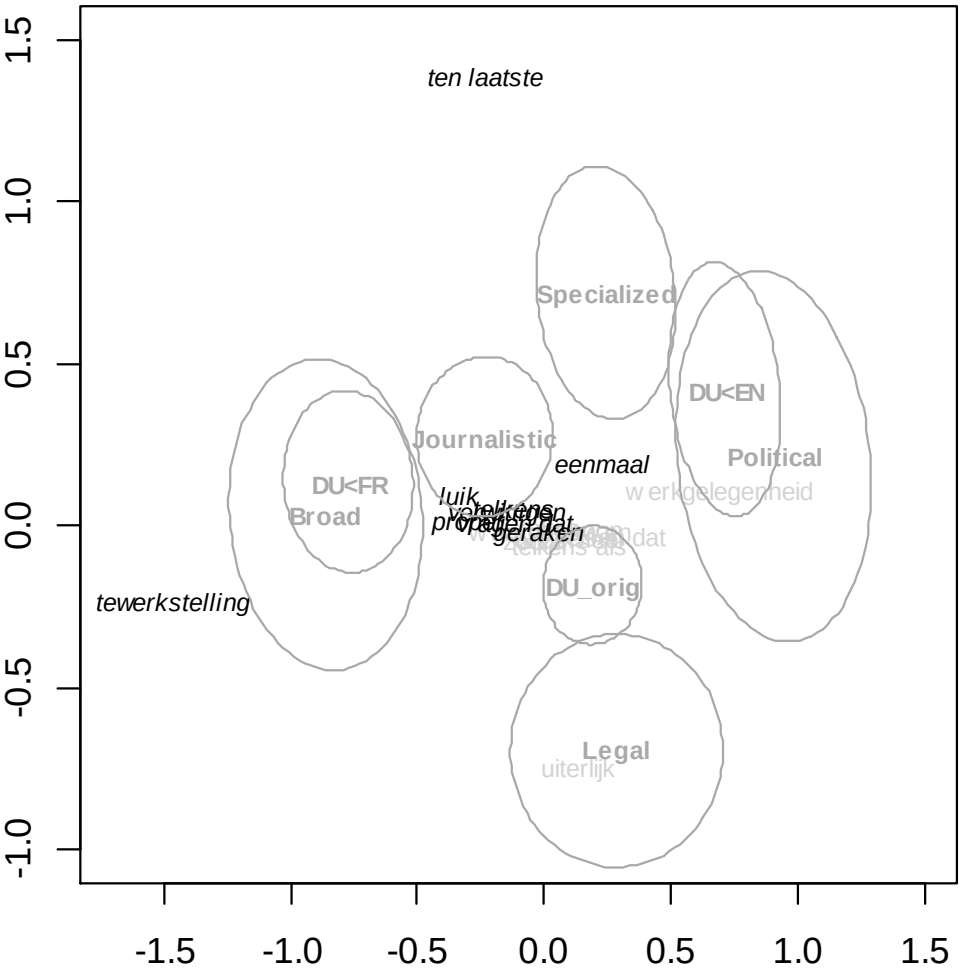

Figure 31 Biplot of the global results with the General Standard Dutch lexemes (light grey), the Belgian Standard Dutch lexemes (black italics), the genres, and the (source) language varieties.

Figure 31 shows the dispersion of Belgian Standard Dutch (BSD) and General Standard Dutch (GSD) in three language varieties on the one hand and in five genres on the other hand. In this plot, the BSD alternatives are represented as black italic data points, and the GSD alternatives are represented as light grey data points. The BSD alternatives are mostly situated towards the left side of the plot, while the GSD alternatives are situated towards the right side. The position of these two clusters provides us with information

with regard to the linguistic preferences within the different (source) language varieties and genres. In `Political`, for example, the General Standard Dutch alternatives seem to be used more often while in `Journalistic`, there seems to be a tendency to use Belgian Standard Dutch alternatives more frequently. In the following paragraphs we will first discuss the results with regard to the (source) language varieties which are displayed in Figure 32 and continue by discussing the genre-related results in Figure 33.

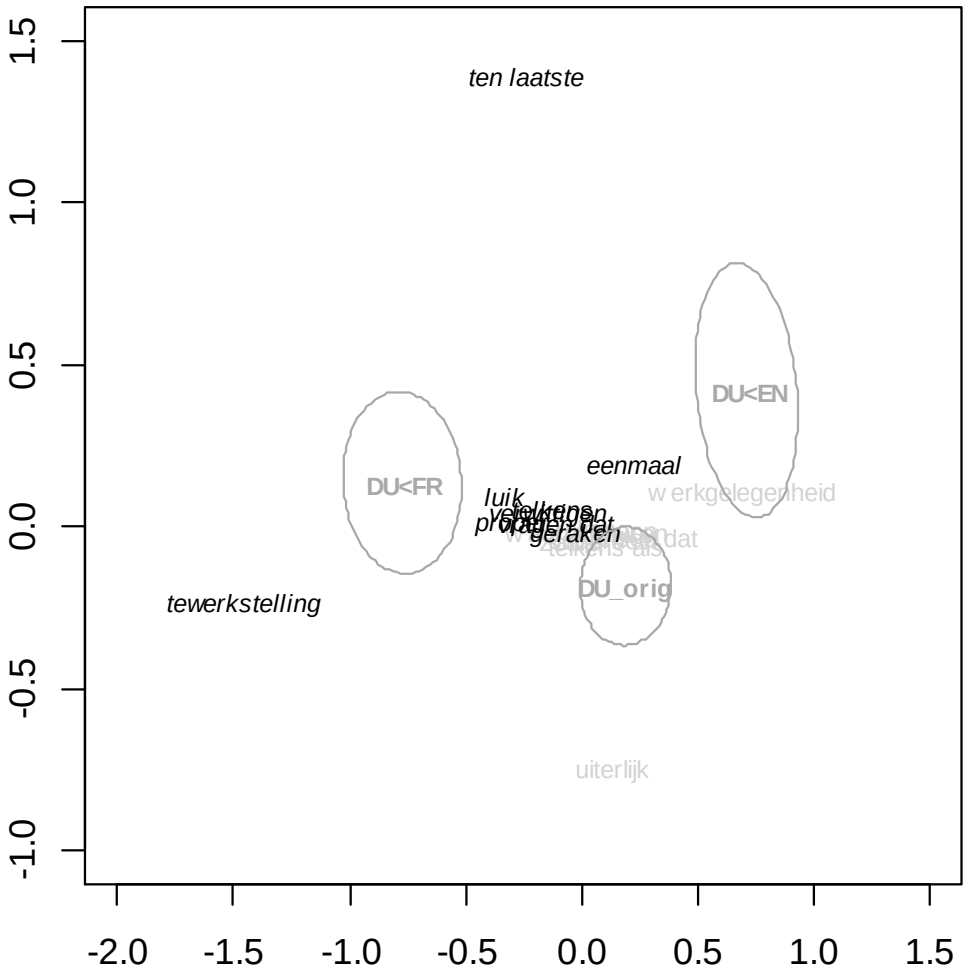

Figure 32 Biplot of the (source) language varieties, the Belgian Standard Dutch alternatives (black italics) and the General Standard Dutch alternatives (light grey)

The first aspect of the plot shown in Figure 32 that attracts attention, is the fact that the three ellipses for the various (source) language varieties are significantly different from one another. If we take a look at the positioning of the ellipses with regard to the clusters for General Standard Dutch and Belgian Standard Dutch, we can see that non-

translated Dutch (DU_orig) appears to make use of General Standard Dutch more often in comparison to the other source language varieties, while DU<FR appears to contain more Belgian Standard Dutch, and Dutch DU<EN seems to waver between General Standard Dutch and Belgian Standard Dutch. In other words, there is a significant difference between Dutch translated from French and Dutch translated from English and translated language as such does not behave uniformly. Moreover, one of the translated sub-corpora, i.e. Dutch translated from French, shows a preference for Belgian Standard Dutch, which is contrary to the assumed norm-conforming behavior of translations. These observations prevent us from confirming our hypothesis, i.e. "translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable Belgian Standard Dutch lexeme, which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts.".

A closer examination of the dataset was needed in order to verify whether there were any additional factors which could help explain the observed tendencies. As lexical items or expressions which are labeled Belgian Standard Dutch are often linked with French (e.g. Gallicisms), we thought it interesting to investigate whether there were certain triggers in the French source texts which could explain the behavior of Dutch translated from French. Within the profile *luik-onderdeel* (CONCEPT: PART), for example, the alternative *luik* is a literal translation from the French *volet*, which both mean 'shutter' or 'blind'. Both *luik* and *volet* are used to express when something is a part of a bigger whole, unit or entity. A closer examination of the data revealed that in the DU<FR sub-corpus, *luik* occurred 21 times and its counterpart in the source text was *volet* 20 times. If we compare this observation with the General Standard Dutch counterpart in the profile, *onderdeel* (n=33), the element in the source text was *volet* only 3 times. A similar example can be found with the profile *vragen dat-eisen dat* (CONCEPT: TO DEMAND), where *vragen* is a literal translation of *demander*, which both mean 'to ask'. Investigating the DU<FR dataset revealed that *vragen dat* (n=36) is a translation of *demander que* in 34 of the 36 cases, while *eisen dat* is a translation of *exiger que* in 8 of the 12 cases. Comparable results were found for the profile *proper-schoon* (CONCEPT: CLEAN), where *proper* is triggered 6 out of 7 times by its literal counterpart *propre.*

This qualitative analysis, however, also revealed an example of a possible trigger for Belgian Standard Dutch which did not result in a higher usage of this variant. In the profile *ten laatste-uiterlijk* (CONCEPT: AT THE LATEST), the alternative *ten laatste* is a literal translation of the French *au plus tard*. *Ten laatste* occurred 26 times in the DU<FR dataset and was a translation of *au plus tard* in 24 cases. The other alternative in the profile, *uiterlijk*, is not a literal translation of *au plus tard* and occurred 34 times. However, in 31 of those cases, *uiterlijk* was a translation of *au plus tard*, therefore showing that *au plus tard* in the source text was not necessarily translated by *ten laatste*.

Summarizing these results from the qualitative data analysis, we can say that, for Dutch translated from French, in some cases the choice for the BSD alternative is triggered by the French source word (e.g. *luik, proper, vragen dat*) while in other cases there is no direct link between the French source word and the fact that it is translated by a BSD alternative (e.g. *raken, tewerkstelling*) and in yet another case the French source word which could trigger a BSD alternative is not necessarily translated as such (e.g. *ten laatste*). These results show that, to some extent, the use of Belgian Standard Dutch might be triggered by the source text alternative, although this is certainly no explanation for the behavior of Dutch translated from French as a whole.

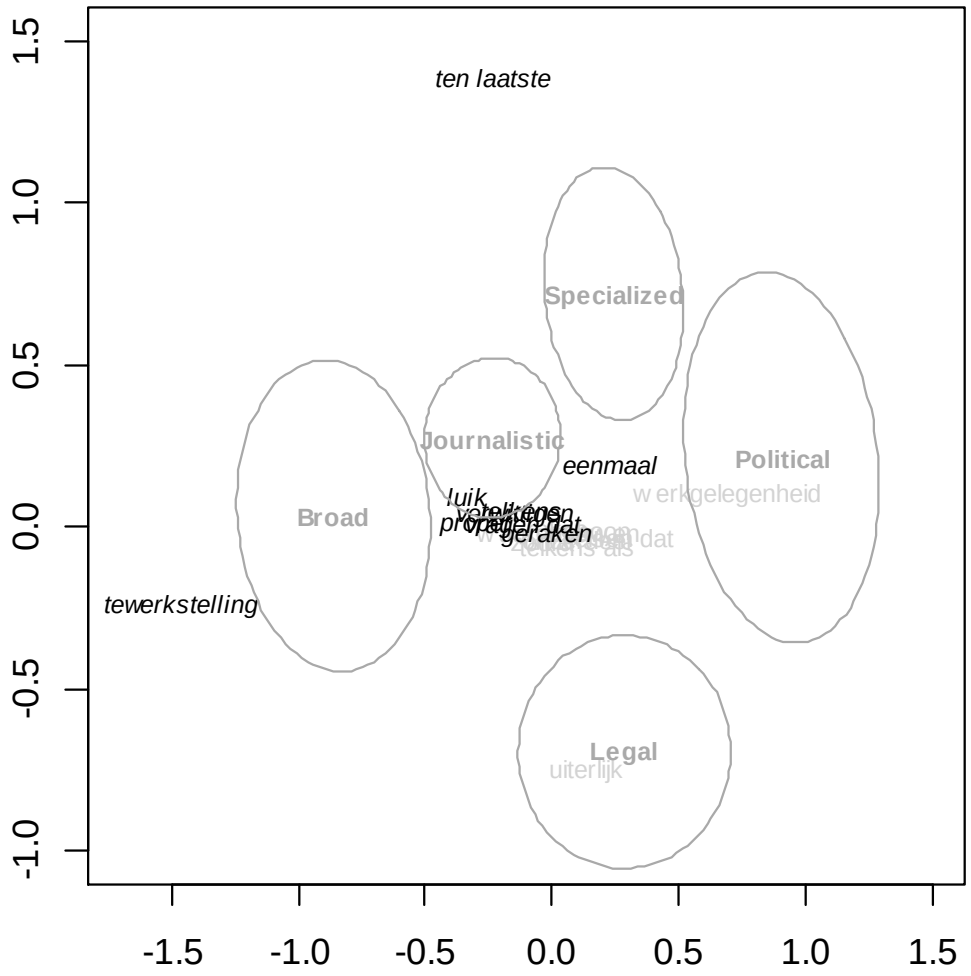

Figure 33 Biplot of the genres, the Belgian Standard Dutch alternatives (black italics) and the General Standard Dutch alternatives (light grey).

Figure 33 shows that the five genres, viz. Specialized Communication, Journalistic Texts, Political Speeches, Broad Commercial Texts and Legal Texts, are significantly different

from one another. More specifically, this plot shows that Broad Commercial Texts, Journalistic Texts and Specialized Communication appear to make more use of Belgian Standard Dutch in comparison to Political Speeches and Legal Texts, which seem to make more use of General Standard Dutch.

Taking a closer look at the contents of these five genres, we can see that the addressor for two of the genres, more particularly Political Speeches and Legal Texts, is state or government related, whereas for the other genres this is not the case. More specifically, the contents of Political Speeches show that this genre contains text material which originated from two sources: the European Parliament and former Belgian Prime Minister Guy Verhofstadt. The metadata for Legal Texts show that this genre consists of five text providers, viz. (i) the Federal Public Service of Social Security; (ii) the *RIZIV,* the Belgian National Service for Medical and Disablement Insurance; (iii) the Federal Public Service of Justice; (iv) the Belgian Chamber of Representatives and; (v) the Ministry of the Flemish Community. Considering the fact that Political Speeches and Legal Texts were (mainly) produced by Belgian institutions which have a public function, it seems not unlikely that the language used in the texts produced by these sources is normalized in order to reach a certain level of homogeneity. This, however plausible, is only a tentative attempt to explain the observed differences between the various genres in our dataset and further research is required in order to assess its validity.

6.2.2 Interaction between source language, translation status and genre

Figure 32 and Figure 33 only provide information with regard to the main effects of (source) language variety and genre. However, they do not tell us whether a given source language has a specific influence on a given genre or not. For example: does Specialized Communication translated from English behave similarly to non-translated Specialized Communication? Or: do translators who translate Specialized Communication make different linguistic choices when compared to the choices made by translators of Journalistic Texts? In order to find out whether there are (source) language varieties or genres in our corpus which behave differently from the general trend, we calculated the interactions between the genres and the (source) language varieties, which are described in the following paragraphs.

6.2.2.1 Broad Commercial Texts

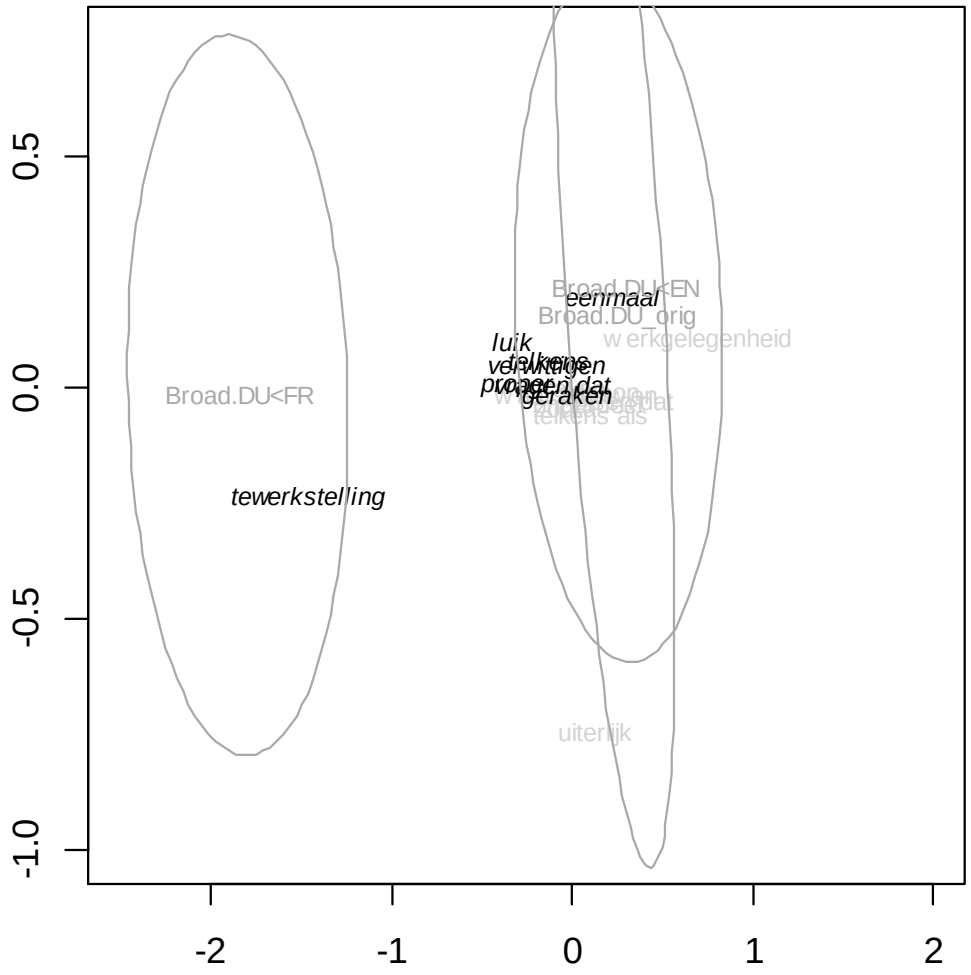

Figure 34 Case Study IV: Biplot of the interaction between Broad Commercial Texts; and source language and translation status

Figure 34 shows the results of the interaction effect between Broad Commercial Texts on the one hand and source language and translation on the other which results in ellipses for Broad Commercial Texts translated from English (`Broad.DU<EN`), Broad Commercial Texts translated from French (`Broad.DU<FR`) and non-translated Broad Commercial Texts (`Broad.DU_orig`). There is a significant difference between `Broad.DU<FR` and the other two sub-varieties, viz. `Broad.DU<EN` and `Broad.DU_orig`. As the distance between `Broad.DU<FR` and the General Standard Dutch alternatives is larger than between `Broad.DU<FR` and the Belgian Standard Dutch alternatives, it appears that `Broad.DU<FR` makes more use of Belgian Standard Dutch than of General Standard Dutch. `Broad.DU<EN` and `Broad.DU_orig` on the

other hand are situated in between the two clusters, showing no clear preference for either variant of Dutch.

To summarize the results with regard to Broad Commercial Texts, it was shown that source language has an effect and translated Broad Commercial Texts do not behave homogeneously. In other words, the assumed norm-conforming behavior of translations cannot be confirmed for this particular genre.

6.2.2.2 Specialized Communication

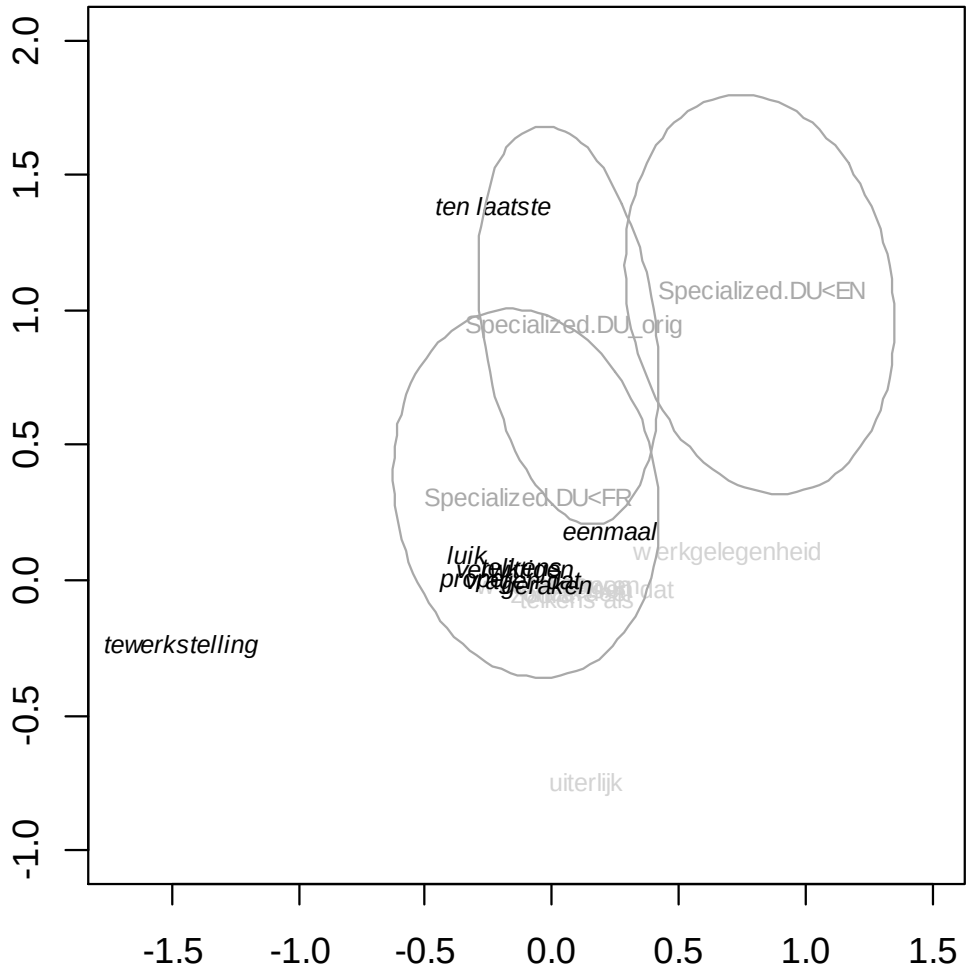

Figure 35 Case Study IV: Biplot of the interaction between Specialized Communication; and source language and translation status

Figure 35 shows the linguistic distances between Specialized Communication translated from French (`Specialized.DU<FR`), Specialized Communication translated from English (`Specialized.DU<EN`) and non-translated Specialized Communication

(Special.DU_orig). As can be seen in the biplot, there is a significant distance between Specialized.DU<FR and Specialized.DU<EN, but not between Specialized.DU<FR and Special.DU_orig or between Specialized.DU<EN and Specialized.DU_orig. In other words, for this particular genre, there is a significant distance between the two source language varieties, but not between translated versus non-translated language. With regard to the preference for Belgian Standard Dutch or General Standard Dutch, this biplot shows that none of the sub-varieties show a clear preference for either GSD or BSD and that the observed distances should be attributed to a preference for specific lexemes rather than variants of Dutch. The hypothesis can therefore not be confirmed as translated Specialized Communication displays no norm-conforming behavior.

6.2.2.3 Journalistic Texts

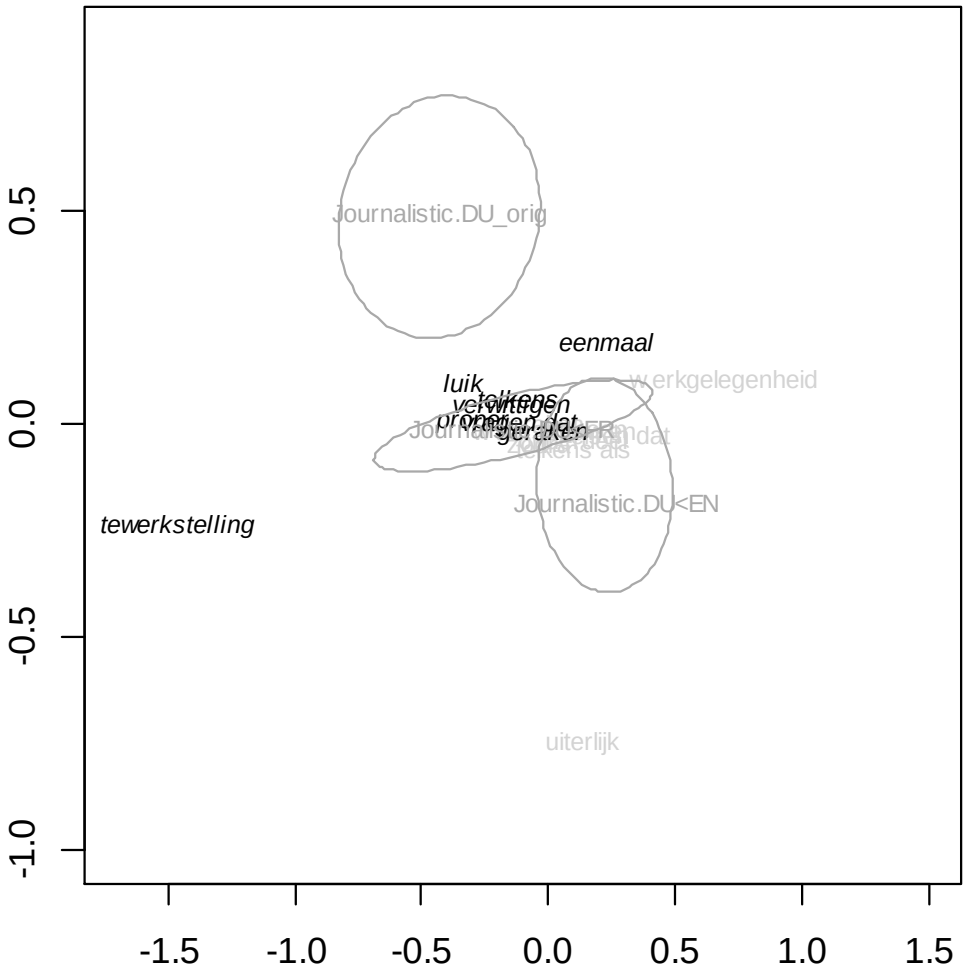

Figure 36 Case Study IV: Biplot of the interaction between Journalistic Texts; and source language and translation status

Looking at Figure 36, which shows the interaction effects for Journalistic Texts, we can see that there is a significant difference between `Journalistic.DU_orig` and `Journalistic.DU<FR` and between `Journalistic.DU_orig` and `Journalistic.DU<EN`, but not between `Journalistic.DU<FR` and `Journalistic.DU<EN`. Moreover, non-translated Journalistic Texts show a preference for BSD as the ellipse is further away from the GSD alternatives. While the distance between `Journalistic.DU<FR` and `Journalistic.DU<EN` is not significant, `Journalistic.DU<FR` appears to make more use of Belgian Standard Dutch while `Journalistic.DU<EN` shows a preference for General Standard Dutch, thus indicating that there is a source language effect, albeit only to some degree. The

hypothesized homogeneous behavior of translated Journalistic Texts and their overall preference for General Standard Dutch could therefore not be confirmed.

6.2.2.4 Legal Texts

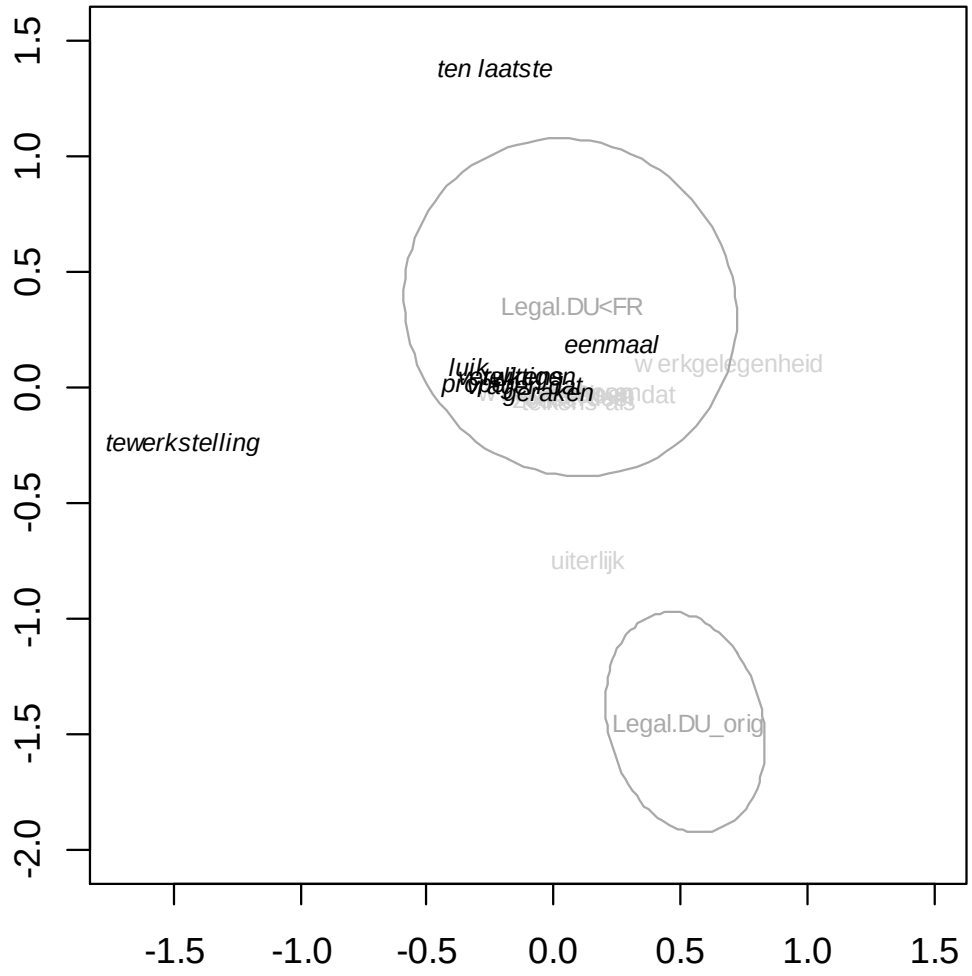

Figure 37 Case Study IV: Biplot of the interaction between Legal Texts and and translation status

Figure 37 is a visualization of the linguistic distances between Legal Texts translated from French (`Legal.DU<FR`) and non-translated Legal Texts (`Legal.DU_orig`). This biplot shows no results for Legal Texts translated from English, as there are no data available for this sub-variety. As mentioned above, this lack of data from a second source language prevents us from investigating the influence of the factor source language, but not from investigating the influence of the translation status on the linguistic preferences within this genre. This plot reveals that there is a significant

distance between `Legal.DU<FR` and `Legal.DU_orig`, which shows that translation status does indeed have an influence on the linguistic preferences as `Legal.DU_orig` make more use of General Standard Dutch whereas `Legal.DU<FR` do not display a preference for either GSD or BSD.

Overall, the norm-conformity hypothesis could not be confirmed for this genre as translated Legal Texts appear to make as much use of Belgian Standard Dutch as of General Standard Dutch.

6.2.2.5 Political Speeches

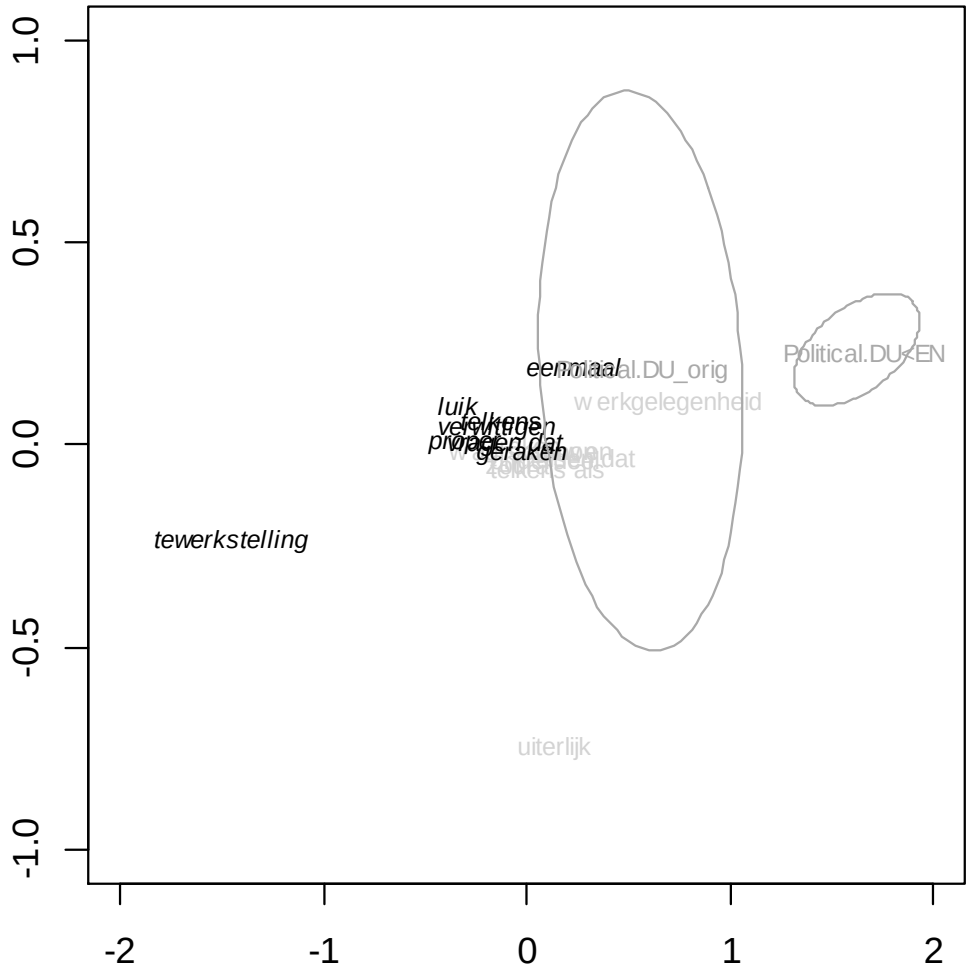

Figure 38 Case Study IV: Biplot of the interaction between Political Speeches; and source language and translation status

As the corpus only contains text material for non-translated Political Speeches and Political Speeches translated from English, Figure 38 shows the results of the interaction

effect between Political Speeches and translation status only. The plot shows that there is a significant difference between `Political.DU_orig` and `Political.DU<EN`, and while both sub-varieties show a preference for General Standard Dutch lexemes, this preference is more outspoken for `Political.DU<EN`. Although we cannot determine whether the observed difference in linguistic behavior should be attributed to the factor source language or translation status, we can confirm the norm-conformity hypothesis for this genre as translated Political Speeches do indeed conform to the prescriptive norm by making more use of General Standard Dutch.

6.2.2.6 Interaction results: conclusion

The overall results obtained by calculating the interactions showed that the norm-conformity hypothesis could only be confirmed for Political Speeches, as this is the only genre the translations of which conform to the prescriptive norm and display a preference for General Standard Dutch. Furthermore, the interaction results revealed that for Broad Commercial Texts and Journalistic Texts, the choice for either General Standard Dutch or Belgian Standard Dutch depends on the source language. More specifically, it was shown that within these two genres texts translated from French make more use of Belgian Standard Dutch than General Standard Dutch.

As shown in the case studies presented in the previous chapters, these interaction results confirm that, depending on the source language and the genre at hand, it is either the source language, the genre or a combination of both which has a stronger influence on the norm-conforming linguistic behavior. However, further research is needed as the exact causes for these differences cannot be determined with by means of our dataset and within this kind of research.

6.3 Conclusion

The results for this case study show that a preference for General Standard Dutch cannot be attributed to the translation's alleged norm-conforming behavior alone as translation status as such is not the only determining factor. Rather, the observed differences are caused by multiple factors such as genre, source language and translation status.

The overall results with respect to the (source) language varieties in our dataset revealed that our initial hypothesis could not be confirmed as translations which were translated from French behave differently from translations which were translated from English. More particularly, translations from French showed an overall preference

for Belgian Standard Dutch whereas translations from English showed an overall preference for neither Belgian nor General Standard Dutch, which showed that translations do not behave as a whole in comparison with non-translations. In other words, we could not confirm our earlier chosen hypothesis: "Translations conform to prescriptive norms and when given the choice between a General Standard Dutch lexeme and an interchangeable Belgian Standard Dutch lexeme, which is identical in meaning, translations into Belgian Dutch make more use of General Standard Dutch than Belgian Dutch non-translated texts".

Looking at the genre variation in our dataset, our results showed that Political Speeches and Legal Texts, two state or government-issued genres, make more use of General Standard Dutch whereas Broad Commercial Texts, Journalistic Texts and Specialized Communication, three non-government-issued genres, make more use of Belgian Standard Dutch. This investigation, however, could not reveal exact causes for the differences in norm-conforming behavior between these genres.

Similar to the other case studies, calculating and visualizing the interaction effects between genre and (source) language variety showed that there are various kinds of possible effects at play. For Broad Commercial Texts and Journalistic Texts, a source language effect could be observed as texts translated from French display a preference for Belgian Standard Dutch while texts translated from English make more use of General Standard Dutch. A similar source-language effect could be observed for Specialized Communication and Legal Texts, but this effect does not result in clear-cut preferences for either General Standard Dutch or Belgian Standard Dutch.

Chapter 7
Conclusion

Do translations walk the line? In Corpus-Based Translation Studies a lot of research has been carried out searching for proof of translations displaying more standardized, more conventional and more conservative linguistic choices in comparison with their source texts and, by extension, with comparable non-translations. The primary objective of this thesis was to verify whether there is indeed a fundamental difference between translated text material on the one hand and non-translated text material on the other as was often assumed in previous research in the field of CBTS. Moreover, it appears that the trends mentioned above all more or less come down to the idea that in general, translations are more conventional than non-translations, an idea which we believe can be explained by the concept of norm conformity. Therefore, a central hypothesis was put forward which predicts that translations show a general preference for norm-conforming behavior when given the choice between a norm-conform option and an interchangeable option which is identical in meaning but not norm conform. Moreover, this study focused on two types of norms, viz. prescriptive and descriptive norms, the difference between which can be summarized as follows: prescriptive norms prescribe how things should be done whereas descriptive norms describe how things are done. The prescriptive norms under investigation are determined by *Taaladvies*, a service which offers advice on language usage for Dutch, while the descriptive norms are genre dependent and based on the dominant linguistic preferences per genre.

However, and contrary to many studies in CBTS, this study set out from the assumption that translating is a multidimensional activity and that in addition to the mere fact whether a text is translated or not there are additional factors which contribute to the characteristics of the target text. In this study there are two factors under investigation aside from translation status, viz. genre and source language, and the goal was to verify how these factors affect the degree to which the linguistic behavior of a given text is norm conform. As the relationship between translated and non-translated language is not mono- but multidimensional, this study required a

technique which can capture this multidimensionality by investigating multiple factors as well as multiple variables simultaneously. Moreover, we needed a technique which can visualize the distances between the various genres and (source) language varieties in our corpus based on the linguistic variables under investigation. Therefore, we decided to apply Profile-Based Correspondence Analysis (PBCA), a technique which is particularly suited for this approach. Moreover, one of the advantages of PBCA is that the results obtained are not biased by the topic of the sub-corpora in our dataset because it does not calculate the distances between meanings but between the formal options.

In the following paragraphs, we will synthesize the empirical findings (7.1), the theoretical implications (7.2), the limitations of this study (7.3), and some directions for future research (7.4).

7.1 Empirical findings

Chapters 3 to 6 presented the results of four case studies, each of which was carried out according to a fixed methodology as described in Chapter 2 in order to examine a sub-hypothesis derived from the central norm-conformity hypothesis. Overall, we applied a corpus-based approach using advanced statistics to explore the Dutch Parallel Corpus in search of norm-conforming behavior in various genres and (source) language varieties. In the following paragraphs we will synthesize the empirical findings per chapter so as to answer this study's research questions.

Chapter 3 (Case Study I) focused on descriptive norm conformity by verifying whether genres in translation conform to the norm established by their non-translated counterparts. This was operationalized by visualizing the dispersion of formal versus neutral lexemes throughout the corpus which enabled us to investigate to what degree a given genre in translation (e.g. translated Legal Texts) displays a similar preference in comparison to its non-translated counterpart (non-translated Legal Texts). The results showed that both genre and source language are determining factors where the choice for either formal or neutral lexemes is concerned and we put forward the explanatory hypothesis that in a given genre, there are language-specific differences in formality as Legal Texts in French, for example, might be a more formal genre than Legal Texts in Dutch.

The second case study described in Chapter 4 also focused on descriptive norm conformity and investigated whether a given genre in translation shows a similar preference for either English loanwords or more endogenous lexemes as its comparable,

non-translated counterpart. The results of Case Study II were similar to those of Case Study I as the hypothesis could not be confirmed for each genre and source language appears to play a substantial role in the choice for either an English loanword or a more endogenous lexeme. Moreover, our results showed that there is no overall preference for more endogenous lexemes within translations, an observation which adds a dimension to previous studies on this topic. Given this case study's focus, it seemed sensible to investigate an additional aspect, i.e. the effect of the corresponding lexeme in the source text on the linguistic preference for more endogenous lexemes versus loanwords in the target text. However, this supplementary analysis showed that the presence of a trigger word in the source text is no deciding factor in the choice for an English loanword in the target text. Similarly, the results of this analysis showed that the absence of a trigger word did not necessarily result in a more endogenous lexeme in the translation.

Looking at the overall results for the first two case studies which focused on descriptive, genre-determined norm conformity, we observed the following trends. First of all, the genre Broad Commercial Texts seems highly susceptible to the influence of the source language, as this factor appears to determine the linguistic preferences in this genre. The behavior of the genres Specialized Communication, Tourist Information, Political Speeches and Legal Texts, however, is not susceptible to the source language as these genres in translation behave conform to the norm determined in their non-translated counterparts. Finally, for Journalistic Texts, it appears that this genre conforms to the genre-determined norm with regard to formal versus neutral lexemes, but not endogenous lexemes versus English loanwords. Surprisingly, within translated Journalistic Texts it is French as source language which appears to trigger a higher usage of English loanwords.

The third case study, which is described in Chapter 5, focused on prescriptive norm conformity by verifying whether translations are indeed more conventional than (comparable) non-translations and therefore make more use of General Standard Dutch instead of non-standard Dutch. Our results showed that the hypothesis could be confirmed as translations do indeed make more use of standard language than comparable non-translations. However, as translations from French make even more use of standard language than translations from English, it appears that there is an additional source language effect at play. The results also revealed specific, genre-related preferences for either standard or non-standard language and the hypothesis that genres which go through an editorial process are more likely to contain standard language than genres without such an editorial process was put forward as a potential explanation for the observed results.

In the final case study presented in Chapter 6, prescriptive norm conformity was investigated by analyzing the dispersion of General Standard Dutch and Belgian

Standard Dutch in translations and comparable non-translations. The results showed that the hypothesized norm-conforming behavior of translations could only be confirmed for translations from English and not from French, thus revealing a source language effect for this particular case study. The results of a qualitative analysis showed that in some cases Belgian Standard Dutch in the target text might have been triggered by the corresponding lexeme in the French source text, which could explain the results for Dutch translated from French to some extent, but not entirely.

Based on the results of Case Studies 3 and 4, which investigated prescriptive norm conformity, we can see that there is no general preference per (source) language variety for General Standard Dutch, non-standard Dutch, or Belgian Standard Dutch, respectively. Dutch translated from French, for example, shows a preference for General Standard Dutch in Case Study III, but a preference for Belgian Standard Dutch in Case Study IV. Based on the genre-related results for prescriptive norm-conformity, we can see that Political Speeches and Specialized Communication are the only genres which show a consistent preference for General Standard Dutch and Belgian Standard Dutch, respectively. However, Journalistic Texts showed a strong preference for the norm conform option in Case Study 4 which resulted in an overall use of General Standard Dutch, irrespective of translation of source language, and it could be so that within this genre Belgian Standard Dutch is considered norm conform, although this tentative hypothesis needs further research. Furthermore, it was shown that Journalistic Texts is a genre with strong linguistic preferences, which are not easily influenced by the source language (except in Case Study 2) and it seems plausible that the genre-related stylistic requirements are more important for Journalistic Texts than for Broad Commercial Texts, for example, where the style might be less important. However, this tentative explanation should be investigated in further research in order to assess its validity.

7.2 Theoretical implications

Initially, this study set out from a bipolar reality in which translations are thought to exhibit linguistic behavior which distinguishes them from non-translations. However, this concept of Translation Universals seems dated and alternative concepts have been suggested such as translational trends, tendencies and regularities which might occur in translated language, without necessarily having the status of all-or-none phenomena. One of these tendencies which was found to be highly likely to occur in translated language, is for translations to use normalized, conventional or conservative language. This tendency has been referred to by the normalization (or conservatism) hypothesis, as well as the conventionality hypothesis. As these hypotheses appear to be similar to a

certain degree and maybe even overlapping , we put forward an "umbrella" hypothesis which predicts that, in general, translations will display linguistic behavior which is norm conform. However, we added two dimensions to the analysis, viz. the genre to which a translation pertains and the source language from which a text is translated.

Our results showed that both genre and source language affect the norm-conforming behavior of translations and that the assumed norm-conforming behavior is certainly not an absolute translational law which is true without exception. However, our results revealed that neither genre nor source language show an effect which is consistent throughout the experiments which were carried out. For example, French as source language as well as Journalistic Texts showed an overall preference for General Standard Dutch in Case Study III, but not in Case Study IV. Therefore, additional research is needed in order to determine the remaining factors which affect the linguistic choices in translated and non-translated text material.

The results mentioned above raised a number of questions which were not within the scope of this thesis but which can lead to further research nonetheless. First, with regard to the differences between one (source) language variety and another, one may wonder whether these differences can be attributed to fundamental differences between one linguistic system and another, especially as we work with both Germanic and Romanic languages. This observation might be particularly relevant with regard to Case Study I as seems plausible that the concept of formality varies depending on the linguistic system under investigation.

Secondly, there is an alternative potential explanation for the observed differences between the various factors in our study which comes to mind, but which could not be confirmed based on our analysis. These differences could be due to a variety of types of language users, e.g. translators of Journalistic Texts with a background in linguistics versus authors of manuals with a more technical background.

Thirdly, with regard to the observed differences in linguistic behavior between genres, one may wonder about their possible consequences, which raises a number of questions such as: Do genre-specific linguistic phenomena improve text comprehensibility? In other words, should the specific linguistic choices be encouraged? Or is it the other way around? Do they cause difficulties to the reader? Do they make the texts, from a cognitive point of view, harder to process? These questions could lead to new interesting directions to be explored within Corpus-Based Translation Studies.

This last remark also raises the question as to how these results can be of value to translation studies from an educational point of view. In particular, it seems relevant to determine how translation training programs can benefit from the insights which were gained from our case studies. In Case Study III, for example, it was shown that texts which are translated from French display a clear preference for General Standard Dutch

and the hypothesis was put forward that a certain linguistic awareness might exist among translators who translate from French into Dutch which results in a tendency to conform to the norm and use General Standard Dutch instead of Gallicisms. It seems interesting to investigate the effect of the observed tendencies on the target audience in order to verify how the translations are perceived. More particularly, such analyses might lead to a highly specific and goal-oriented approach in translation training programs, depending on whether these tendencies are perceived as positive or negative.

7.3 Limitations of the study

In addition to the results and implications described in the previous paragraphs, this study also encountered a number of limitations, some of which might be overcome in future work. First of all, it should be noted that, as our results are based on exploratory research, we could only observe the overall effects per factor and no exact causal relations could be determined. An additional consequence of our corpus-based approach is that we study translation as a product and not as a process, which prevented us from determining where exactly in the translation process a given linguistic choice is made. In other words, the choice to conform to the norm could be made by the translator, the editor, etc. Moreover, the corpus metadata on the individual author or translator were often incomplete, which also prevented us from generating conclusions with regard to the behavior of the individual author or translator. Therefore, we would like to issue a call for corpora enriched with metadata which are detailed, accurate and complete.

Furthermore, we believe that in addition to corpora with more complete metadata, there is also a need for larger corpora, especially for the more exotic languages, as our corpus queries sometimes returned low frequencies, which resulted in inconclusive results due to statistical uncertainties which are represented by large ellipses in the biplots.

7.4 Future research

The results varied depending on the case study at hand, which may mean that there are additional factors at play aside from genre and source language and further research is needed to both determine and investigate these factors.

As mentioned in Section 7.3 the research conducted in this thesis is of an exploratory nature and although it yields a more fine-grained insight into the linguistic characteristics of translated texts, additional research is needed to determine the causes for the observed trends. For example, an explanatory analysis could be carried out while including all the available metadata under investigation as this might result in valuable insights into what aspects cause the observed trends. Ideally, both our exploratory and an explanatory product-based phase should be completed with process-based research so as to reveal when and why certain linguistic choices are made.

Although the present study suffered from a number of limitations, the empirical findings from our corpus-based multivariate approach showed that translations do not always walk the line as they do not necessarily display more norm conform behavior than comparable non-translations and that the factors genre and source language and all underlying aspects affect the degree to which a given translation conforms to the norm.

Bibliography

Allauzen, A., & Bonneau-Maynard, H. (2008). *Training and evaluation of POS taggers on the French MULTITAG corpus.* Paper presented at the Language Resource and Evaluation Conference, Marrakech.

Altenberg, B., & Aijmer, K. (2000). The English-Swedish Parallel Corpus. A resource for contrastive research and translation studies. In C. Mair & M. Hundt (Eds.), *Corpus Linguistics and Linguistic Theory* (pp. 15-33). Amsterdam: Rodopi.

Bakema, P. (Ed.) (2004). Vlaams-Nederlands woordenboek: van ambetanterik tot zwanzer Antwerpen: Het Spectrum.

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

Baker, M. (1995). Corpora in Translation Studies: An overview and Some Suggestions for Future Research. *Target, 7*(2), 223-243.

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager* (pp. 175-186). Amsterdam/Philadelphia: John Benjamins.

Baker, M. (1999). The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators. *International Journal of Corpus Linguistics, 4*(2), 281-298.

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

Baker, M., & Saldanha, G. (2008). *The Routledge Encyclopedia of Translation Studies* (2 ed.). London/New York: Routledge.

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

Becher, V. (2011). *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts.* (PhD Thesis), University of Hamburg.

Ben-Shahar, R. (1994). Translating Literary Dialogue: A Problem and its Implications for Translation into Hebrew. *Target, 6*(2), 195-212.

Bernardini, S. (2010). *Parallel corpora and the search for translation norms/universals.* Paper presented at the Methodological Advances in Translation Studies, Ghent.

Bernardini, S., & Ferraresi, A. (2011). Practice, Description and Theory Come Together: Normalization or Interference in Italian Technical Translation? *Meta, 56*(2), 226-246.

Bernardini, S., & Zanettin, F. (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals.

In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals. Do they exist?* (pp. 33-50). Amsterdam / Philadelphia: John Benjamins Publishing Company.
Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.
Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17-35). Tübingen: Gunter Narr.
Böttner, A. (1995). Journalistiek vertalen. Een impressie. *Filter, tijdschrift over vertalen, 2*(2), 77-83.
Burger, P., & De Jong, J. (1991). De vraag naar goed of fout. Normen. In P. Burger & J. De Jong (Eds.), *Onze taal! Zestig jaar strijd en liefde voor het Nederlands.* (pp. 77-89). Den Haag: SdU Uitgeverij.
Cappelle, B. (2012). English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures, 13*(2), 173-195.
Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249-254.
Catford, J. (1965). *A Linguistic Theory of Translation: An Essay in Applied Linguistics*. London: Oxford University Press.
Chesterman, A. (1997). *Memes of translation: The spread of ideas in translation theory.* Amsterdam/Philadelphia: John Benjamins.
Chesterman, A. (2004a). Beyond the Particular. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals. Do they Exist?* (pp. 33–50). Amsterdam/Philadelphia: John Benjamins.
Chesterman, A. (2004b). Hypotheses about Translation Universals. In G. Hansen, K. Malmkjaer & D. Gile (Eds.), *Claims, Challenges and Changes in translation studies*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Chesterman, A. (2010). Why study translation universals? In R. Hartama-Heinonen & P. Kukkonen (Eds.), *Kiasm. Acta Translatologica Helsingiensia* (Vol. 1, pp. 38-48). Helsingfors: Helsingfors universitet.
Claes, F. (Ed.) (1996) Verschueren Groot Encyclopedisch Woordenboek (Vol. 10). Antwerpen: Standaard.
Cockx, P. (1998). *Taalwijzer*. Leuven: Davidsfonds.
Daelemans, W., & Van Den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.
de Boer, W. T. (Ed.) (2006) Koenen Woordenboek Nederlands. Utrecht: Van Dale Lexicografie.
De Clercq, O., & Montero Perez, M. (2010). *Data collection and IPR in multilingual parallel corpora : Dutch parallel corpus.* Paper presented at the International Conference on Language Resources and Evaluation, Malta.
De Sutter, G., Delaere, I., & Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling. In M. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-based Translation Studies. A practical guide to descriptive translation research.* (pp. 325-345). Amsterdam/Philadelphia: John Benjamins Publishing Company.
De Sutter, G., Goethals, P., Leuschner, T., & Vandepitte, S. (2012). Towards methodologically more rigorous corpus-based translation studies. *Across Languages and Cultures, 13*(2), 137-143.
den Boon, C. A., & Geeraerts, D. (Eds.). (2010-2013) Van Dale: Groot Woordenboek der Nederlandse Taal (14 ed.). Utrecht: Van Dale Lexicografie.

Deygers, K. (1998). Normen voor goed Nederlands in Noord en Zuid. *Handelingen van de Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis, 52*, 77-94.

Diwersy, S., Evert, S., & Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Waelchli (Eds.), *Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech*. Berlin: De Gruyter.

Eskola, S. (2004). Untypical frequencies in translated language. A corpus-based study on a literary corpus of translated and non-translated Finnish. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals - Do They Exist?* (pp. 83-99). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Even-Zohar, I. (1978). The position of translated literature within the literary polysystem. In J. S. Holmes, J. Lambert & R. van den Broeck (Eds.), *Literature and Translation: new perspectives in Literary Studies* (pp. 117-127). Leuven: Acco.

Even-Zohar, I. (1979). Polysystem Theory. *Poetics Today, 1*(1-2 Autumn), 287-310.

Francis, N. (1992). *Language Corpora B.C.* Paper presented at the Nobel Symposium, Stockholm.

Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen.* Amsterdam: Meertens Instituut.

Greenacre, M. (2007). *Correspondence analysis in practice, Second edition.* . Boca Raton: Chapman & Hall/CRC.

Gries, S. T. (2010). Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods. In T. Harris & M. Moreno Jaén (Eds.), *Corpus linguistics in language teaching* (pp. 121-146). Frankfurt a.M.: Peter Lang.

Haeseryn, W. (1996). Grammaticale verschillen tussen het Nederlands in België en het Nederlands in Nederland: een poging tot inventarisatie. In R. van Hout & J. Kruijsen (Eds.), *Taalvariaties. Toonzettingen en modulaties op een thema.* (pp. 109-126). Dordrecht: Foris.

Haeseryn, W. e. a. (Ed.). (1997). *Algemene Nederlandse Spraakkunst: 2 Delen* (2 ed.). Groningen: Martinus Nijhoff.

Hansen-Schirra, S., Neumann, S., & Steiner, E. (2012). *Cross-Linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German* (Vol. 11). Berlin/Boston: De Gruyter.

Hansen, G. (2006). Retrospection methods in translator training and translator research. *Journal of Specialised Translation, 5*, 1-40.

Hansen, S., & Hansen-Schirra, S. (2012). Grammatical shifts in English-German noun phrases. In S. Hansen-Schirra, S. Neumann & E. Steiner (Eds.), *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German* (pp. 133-146). Berlin/Boston: De Gruyter.

Hatim, B., & Munday, J. (2004). *Translation. An advanced resource book.* New York: Routledge.

Hendrickx, E., Hendrickx, K., Martin, W., Smessaert, H., Van Belle, W., & van der Horst, J. (Eds.). (2010). *Liever meer of juist minder? Over normen en variatie in taal*. Gent: Academia Press.

Hewitt, J. A. (2005). A habit of lies: how scientists cheat.

Heylen, K., Tummers, J., & Geeraerts, D. (2008). Methodological issues in corpus-based Cognitive Linguistics. In G. Kristiansen & R. Dirven (Eds.), *Cognitive Sociolinguistics: Language variation, cultural models, social systems* (pp. 91-128). Berlin/New York: Mouton de Gruyter.

Holmes, J. S. (2000). The Name and Nature of Translation Studies. In L. Venuti (Ed.), *The Translation Studies Reader* (pp. 172-185). London/New York: Routledge.

House, J. (2008). Beyond intervention: Universals in translation? *Trans-kom. Zeitschrift für Translationswissenschaft und Fachkommunikation, 1*(1), 6-19.

Jakobson, A. L. (2005). Investigating expert translator's processing knowledge. In V. H. Dam, J. Engberg & H. Gerzymisch-Arbogast (Eds.), *Knowledge systems and translation* (pp. 173-189). Berlin: Walter de Gruyter.

Jenset, G. B., & McGillivray, B. (2012). Multivariate analyses of affix productivity in translated English. In J. Meng & M. Oakes (Eds.), *Quantitative methods in translation studies* (pp. 301-324). Amsterdam/Philadelphia: John Benjamins.

Kamp, v. d., P.H.J., & Kruyt, J. G. (2004). *Putting the Dutch PAROLE Corpus to Work.* Paper presented at the International Conference on Language Resources and Evaluation (ELRA), Paris.

Kenny, D. (1998). Creatures of Habit? What translators usually do with words. *Meta, 43*(4), 515-523.

Kenny, D. (2000). Lexical Hide-and-Seek: looking for creativity in a parallel corpus. In M. Olohan (Ed.), *Intercultural Faultlines: Research Models in Translation Studies I: Textual and Cognitive Aspects* (pp. 93-104). Manchester: St. Jerome.

Kenny, D. (2001). *Lexis and Creativity in Translation: A Corpus-based Study*. Manchester: St. Jerome Publishing.

Klaudy, K. (2001). *The Asymmetry Hypothesis. Testing the Asymmetric Relationship between Explicitations and Implicitations*. Paper presented at the the Third International Congress of EST 'Claims, Changes and Challenges in Translation Studies, Copenhagen

Koehn, P. (2005). *Europarl: a parallel corpus for statistical machine translation.* Paper presented at the Tenth Machine Translation Summit, Phuket, Thailand.

Kruger, H. (to appear). The effects of editorial intervention: implications for studies of the features of translated language. In G. De Sutter, I. Delaere & M.-A. Lefer (Eds.), *New ways of analyzing translational behavior*.

Kruger, H., & van Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures, 13*(1), 33-65.

Kruyt, J. G., & Dutilh, M. W. F. (1997). A 38 Million Words Dutch Text Corpus and its Users. *Lexikos, 7*, 229-244.

Landauer, T., & Dumais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review, 104*(2), 211-240.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta, 43*, 557-570.

Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, Findings, Applications*. Amsterdam/New York: Rodopi.

Laviosa, S. (2006). Data-Driven Learning for Translating Anglicisms in Business Communication. *IEEE transactions on professional communication, 49*(3), 267-274.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37-72.

Lefer, M.-A. (2012). Word-formation in translated language: the impact of language-pair specific features and genre variation. *Across Languages and Cultures, 13*(2), 145-172.

Lefer, M.-A., & Vogeleer, S. (2013). Interference and normalization in genre-controlled multilingual corpora. *Belgian Journal of Linguistics, 27*(1), 1-21.

Lindqvist, Y. (2010). Manipulating the matricial norms. A Comparison of the English, Swedish and French translations of La caverna de las ideas by José Carlos Somoza. In

D. Gile, G. Hansen & N. K. Pokorn (Eds.), *Why Translation Studies Matters* (pp. 69-81). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch Parallel Corpus: a Balanced Copyright-Cleared Parallel Corpus. *Meta, 56*(2), 374-390.
Malmkjaer, K. (1997). Punctuation in Hans Christian Andersen's stories and in their translations into English. In F. Poyatos (Ed.), *Nonverbal Communication and Translation*. Philadelphia: John Benjamins.
Malmkjaer, K. (2008). Norms and nature in translation studies. In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora: the linguist and the translator* (pp. 49-59). Clevedon: Multilingual Matters.
Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313-330.
Martin, E. (2002). Cultural images and different varieties of English in French television commercials. *English Today, 18*(4), 8-20.
Martin, W., & Smedts, W. (Eds.). (2010) Prisma Handwoordenboek Nederlands. Utrecht: Het Spectrum.
Mauranen, A. (2000). Strange things in translated language: A study on corpora. In M. Olohan (Ed.), *Intercultural Faultlines. Research models in translation studies 1: Textual and cognitive aspects.* (pp. 119-141). Manchester: St. Jerome.
Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals - Do They Exist?* (pp. 65-82). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Mauranen, A. (2008). Universal tendencies in translation. In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora: The Linguist and the Translator* (pp. 32-48). Clevedon: Multilingual Matters.
Mauranen, A., & Kujamäki, P. (2004). *Translation Universals: Do they exist?* Amsterdam/Philadelphia: John Benjamins.
May, R. (1997). Sensible Elocution: How Translation Works in & upon Punctuation. *The Translator, 3*(1), 1-20.
Munday, J. (2012). *Introducing Translation Studies. Theories and Applications.* (3 ed.). Oxon / New York: Routledge.
Neumann, S. (2013). *Contrastive register variation. A quantitative approach to the comparison of English and German*. Berlin: de Gruyter.
Neumann, S. (2014). Cross-linguistic register studies. Theoretical and methodological considerations. *Languages in Contrast, 14*(1), 35-57.
O'Brien, S. (2010). Eye tracking in translation process research: methodological challenges and solutions. *Copenhagen Studies In Language, 38*(251-266).
Oakes, M., & Ji, M. (Eds.). (2012). *Quantitative Methods in Corpus-Based Translation Studies* (Vol. 51). Amsterdam: John Benjamins.
Olohan, M. (2004). *Introducing corpora in Translation Studies*. London: Routledge.
Olohan, M., & Baker, M. (2000). Reporting that in Translated English: Evidence for Subconscious Processes of Explicitation? *Across Languages and Cultures, 1*(2), 141-158.
Oostdijk, N. (2000). Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde, 5*(3), 280-284.
Ordelman, R. J. F. (2003). *Dutch Speech Recognition in Multimedia Information Retrieval.* (PhD Thesis), University of Twente, Enschede.
Øverås, L. (1998). In search of the third code. An investigation of norms in literary translation. *Meta, 43*(4), 557-570.
Paroubek, P. (2000). *Language resources as by-product of evaluation: the Multitag example.* Paper presented at the Language Resources and Evaluation Conference,, Athens.

Paulussen, H. (1999). *A corpus-based contrastive analysis of English 'on/up', Dutch 'op' and French 'sur' within a cognitive framework.* (PhD), Ghent University, Gent.
Penninckx, W., Buyse, P., & Smedts, W. (2001). *Correct Taalgebruik*. Kortrijk-Heule: UGA.
Permentier, L. (2003). *Stijlboek: Onmisbaar voor wie helder wil schrijven.* Roeselaere: Roularta Books.
Plevoets, K. (2008). *Tussen spreek- en standaardtaal. Een corpusgebaseerd onderzoek naar de situationele, regionale en sociale verspreiding van enkele morfosyntactische verschijnselen uit het gesproken Belgisch-Nederlands.* (Doctoral dissertation), Katholieke Universiteit Leuven, Leuven.
Plevoets, K. (2013). De status van de Vlaamse tussentaal: een analyse van enkele socio-economische determinanten. *Tijdschrift voor Nederlandse Taal- en Letterkunde, 129*(3), 191-223.
Popovič, A. (1970). The Concept of "Shift of Expression" in Translation Analysis. In J. S. Holmes (Ed.), *The Nature of Translation: Essays on the Theory and Practice of Literary Translation* (pp. 78-87). Den Haag/Parijs: Mouton.
Prisma, & Taal, G. O. (Eds.). (2007) Het Witte Woordenboek Nederlands. Utrecht: Het Spectrum.
Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-transladet Finnish children's literature. *Literary and linguistic computing, 18*(4), 389-406.
Pym, A. (2005). Explaining Explicitation. In K. Karoly & Á. Fóris (Eds.), *New Trends in Translation Studies. In Honour of Kinga Klaudy* (pp. 29-34). Budapest: Akadémia Kiadó.
Pym, A. (2008). On Toury's laws of how translators translate. In A. Pym, M. Shlesinger & D. Simeoni (Eds.), *Beyond Descriptive Translation Studies. Investigations in Honor of Gideon Toury.* (pp. 311-328). Amsterdam/Philadelphia: Benjamins.
Reiczigel, J. (1996). Bootstrap tests in correspondence analysis. *Applied Stochastic Models And Data Analysis, 12*(2), 107-117.
Renkema, J. (1985). De Taaladviesdienst. Over regels en normen in taalgebruik. *Onze Taal, 54*(11), 138-141.
Ruette, T. (2012). *Aggregating Lexical Variation, towards large-scale lexical lectometry*. PhD Dissertation. Faculty of Arts. University of Leuven. Leuven.
Rybicki, J. (2012). The great mystery of the (almost) invisible translator. In J. Meng & M. Oakes (Eds.), *Quantitative Methods in Corpus-Based Translation Studies* (pp. 231-248). Amsterdam/Philadelphia: John Benjamins.
Salton, G., Wong, A., & Yang, c.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 18*(1), 613-620.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In D. B. Jones & H. Somers (Eds.), *New Methods in Language Processing (Studies in Computational Linguistics)* (pp. 154-163). London: Routledge.
Scott, N. (1998). *Normalisation and Readers' Expectations: A Study of Literary Translation with Reference to Lispector's A Hora da Estrela.* (PhD), University of Liverpool, Liverpool.
Shlesinger, M. (1991). *Interpreter Latitude vs. Due Process: Simultaneous and Consecutive Interpretation in Multilingual Trials.* Paper presented at the Empirical Research in Translation and Intercultural Studies.
Sinclair, J. (1996). Preliminary recommendations on Corpus Typology.
Speelman, D., Grondelaers, S., & Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities, 37*(3), 317-337.
Steiner, E. (2001). Translations English-German: investigating the relative importance of systemic contrasts and of the text type "translation". *SPRIKreports* (Vol. 7, pp. 1-49).

Stewart, D. (2000). Conventionality, Creativity, and Translated Text: The Implications of Electronic Corpora in Translation. In M. Olohan (Ed.), *Intercultural Faultlines. Research Models in Translation Studies I: Textual and Cognitive Aspects* (pp. 73-91). Manchester: St. Jerome.

Strang, G. (2009). *Introduction to linear algebra*. Wellesley: Wellesley-Cambridge Press.

Taeldeman, J. (1992). Welk Nederlands voor Vlamingen? *Nederlands van nu, 40*, 33-52.

Teich, E. (2003). *Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts.* Berlin/New York: Mouton De Gruyter.

Theissen, S., & Debrabandere, P. (Eds.). (2004) Woordenboek voor correct taalgebruik. Mechelen: Wolters Plantyn.

Tirkkonen-Condit, S. (2004). Keywords and Ideology in Translated History Texts: A Corpus-Based Analysis. In A. K. Mauranen, Pekka (Ed.), *Translation Universals. Do they exist?* (pp. 177-184). Amsterdam/Philadelphia: John Benjamins.

Toury, G. (1991). Experiments in Translation Studies: Achievements, Prospects and some Pitfalls. In S. Tirkkonen-Condit (Ed.), *Empirical Research in Translation and Intercultural Studies* (pp. 45-66). Tübingen: Guner Narr Verlag.

Toury, G. (1995). *Descriptive Translation Studies and beyond.* Amsterdam: John Benjamins.

Trosborg, A. (1997). Text typology: register, genre and text type. In A. Trosborg (Ed.), *Text Typology and Translation*. Amsterdam/Philadelphia: John Benjamins.

Tymoczko, M. (2005). Trajectories of Research in Translation Studies. *Meta, 50*(4), 1082-1097.

Uit den Boogaart, P. C. (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

Van Craenenbroeck, J. (2000). Complementerend van: een voorbeeld van syntactische variatie in het Nederlands. *Nederlandse Taalkunde, 5*, 133-163.

Van Den Bosch, A., Schuurman, I., & Vandeghinste, V. (2006). *Transferring POS tagging and lemmatization tools from spoken to written Dutch corpus development.* Paper presented at the Language Resource and Evaluation Conference, Genua.

Van der Horst, P. J. (1996). *Stijlwijzer: Praktische handleiding voor leesbaar schrijven*. Den Haag: Sdu.

van der Horst, P. J. (1999). *De taalgids : tekstverzorging van A tot Z*. Den Haag/Antwerpen: Standaard.

van der Spek, E. (1990). Zekerheid bestaat niet. Wie bepaalt wat goed taalgebruik is? *Onze Taal, 59*, 146-148.

Van der Spek, H. (2004). *Working on the future of the language. Portrait of the Nederlandse Taalunie*. Nederlandse Taalunie.

Van Eynde, F., Zavrel, J., & Daelemans, W. (2000). *Part of Speech Tagging and Lemmatisation for the Spoken Dutch Corpus.* Paper presented at the Language Resources and Evaluation Conference, Athens.

van Sterkenburg, P. (Ed.) (2002). Van Dale. Groot woordenboek hedendaags Nederlands. Van Dale Lexicografie.

Vanderauwera, R. (1985). *Dutch novels translated into English: The transformation of a "Minority" literature.* Amsterdam: Rodopi.

Vinay, J.-P., & Darbelnet, J. (1995). *Comparative Stylistics of French and English: A Methodology for Translation, translated and edited by Juan Sager and Marie-Jo Hamel.* Amsterdam/Philadelphia: John Benjamins.

Volansky, V., Ordan, N., & Wintner, S. (2013). On the features of translationese. *Literary and linguistic computing*. doi: 10.1093/llc/fqt031

Werlich, E. (1982). *A Text Grammar of English*. Heidelberg: Quelle & Meyer.

Zanettin, F. (2000a). *CEXI: Designing an English Italian Translational Corpus.* Paper presented at the Fourth International Conference on Teaching and Language Corpora, Graz.

Zanettin, F. (2000b). Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis. In M. Olohan (Ed.), *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects* (pp. 105-118). Manchester: St. Jerome Publishing.

Zenner, E. (2013). *Cognitive Contact Linguistics. The macro, meso and micro influence of English on Dutch.* (PhD dissertation), KU Leuven, Leuven.

Zenner, E., Speelman, D., & Geeraerts, D. (2011). *A Concept-Based Approach to Measuring the Success of Loanwords.* Paper presented at the Quantitative Investigations in Theoretical Linguistics, Berlin.

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target, 26*(3), 361-384.

Appendix

Abstract

This thesis set out to investigate the long-standing assumption that translated language as such differs from non-translated language by means of operationalizing an umbrella hypothesis which predicts that translated language will conform to the norm more often than non-translated language. Moreover, so as to take the multidimensional character of translations into account we not only investigated the factor translation status, but also two additional, extralinguistic factors, viz. genre and source language.

These factors were analyzed with regard to two types of norm conformity: descriptive and prescriptive norm conformity. The difference between these two types of norms can be summarized as follows: descriptive norms describe how things are done whereas prescriptive norms prescribe how things should be done. This thesis focuses on *descriptive* norm-conforming behavior by examining to what degree translations conform to genre-determined norms which describe what is prevailing language usage in a given genre and what is not. We investigate *prescriptive* norm-conforming behavior on the other hand by verifying to what extent translations and non-translations conform to the linguistic norms issued by the Dutch Language Union which prescribe whether certain lexemes, expressions, etc. belong to the standard language or not.

So as to be able to visualize the norm-conforming behavior in various genres and source language varieties we applied profile-based correspondence analysis which is an exploratory, multivariate statistical technique. As lexical variation is based on the idea that language users can choose between two or more formal alternatives to express a given concept, a profile-based approach is very well suited to investigate these linguistic preferences as the analysis is based on the difference in use between the formal alternatives and not the concepts behind them. Moreover, the technique leads to conclusions which are valid on a general level due to the objectivity of the statistical

method and the substantial number of variables which can be used in each analysis and which are analyzed together instead of separately.

In order to investigate norm-conforming behavior four sub-hypotheses were derived from the central norm conformity hypothesis, each of which was operationalized by means of a case study. More specifically, two case studies were carried out which focused on exploring descriptive norm conformity, and two case studies were carried out to investigated prescriptive norm conformity. Although the present study suffered from a number of limitations, the empirical findings from our corpus-based multivariate approach showed that translations do not always walk the line as they do not necessarily display more norm conform behavior than comparable non-translations and that the factors genre and source language and all underlying aspects affect the degree to which a given translation conforms to the norm.

Samenvatting

In dit proefschrift werd onderzocht of vertaalde teksten daadwerkelijk verschillen van niet-vertaalde teksten, een lang aangenomen veronderstelling. Dit werd onderzocht aan de hand van een overkoepelende hypothese die voorspelt dat vertaalde teksten meer normconform gedrag vertonen dan niet-vertaalde teksten. Daarenboven werd een techniek gebruikt die toelaat om het multidimensionale karakter van vertalingen te onderzoeken waardoor we niet alleen de factor vertaalstatus, maar ook twee extralinguïstische factoren konden onderzoeken, namelijk genre en brontaal.

We onderzochten deze drie factoren met betrekking tot twee soorten normconformiteit: descriptieve en prescriptieve normconformiteit. Het verschil tussen deze twee soorten normen kan worden omschreven als volgt: *descriptieve* normen beschrijven hoe de zaken eraan toegaan terwijl *prescriptieve* normen voorschrijven hoe de zaken eraan toe zouden moeten gaan. In dit proefschrift wordt normconformiteit met betrekking tot descriptieve normen onderzocht door na te gaan in welke mate vertalingen conformeren aan genre-specifieke normen die beschrijven wat gangbaar is binnen een bepaald genre en wat niet. Prescriptieve normconformiteit daarentegen wordt onderzocht door na te gaan in welke mate vertalingen en niet-vertalingen conformeren aan de normen beschreven door de Nederlandse Taalunie die aangeven wat Standaardnederlands is en wat niet.

Om de verschillende mate waarin een aantal genres en brontaalvariëteiten normconform taalgebruik hanteren te kunnen visualiseren, werd gebruik gemaakt van profielgebaseerde correspondentieanalyse, een exploratieve, multivariate statistische techniek. Lexicale variatie is gebaseerd op de idee dat taalgebruikers kunnen kiezen tussen twee of meerdere vormelijke alternatieven die een bepaald concept uitdrukken. Hierdoor is de profielgebaseerde aanpak uitermate geschikt om dit type taalkundige voorkeur te analyseren aangezien deze aanpak steunt op een verschil in gebruik tussen de vormelijke varianten en niet de concepten erachter. Daarenboven leidt deze techniek tot conclusies die algemeen bruikbaar zijn dankzij het grotere aantal variabelen die worden opgenomen in de analyse en die bovendien samen worden geanalyseerd in plaats van apart.

In totaal werden vier gevalstudies uitgevoerd waarin normconform taalgebruik werd geanalyseerd op basis van vier sub-hypotheses die werden afgeleid van de centrale normconformiteitshypothese. In de eerste twee gevalstudies werd descriptieve

normconformiteit onderzocht en in de laatste twee gevalstudies werd prescriptieve normconformiteit geanalyseerd.

Hoewel het onderzoek dat wordt voorgesteld in dit proefschrift niet vrij is van enkele beperkingen, leidden de empirische resultaten toch tot een aantal bevindingen. Onze corpus-gebaseerde, multivariate aanpak toonde aan dat vertalingen niet altijd meer normconform gedrag vertonen dan vergelijkbare niet-vertalingen en dat genre en brontaal, inclusief alle onderliggende aspecten, factoren zijn die ook een invloed hebben op de mate waarin een gegeven vertaling de norm volgt of niet.

Publications

De Sutter, G., Delaere, I., & Plevoets, K. (2012). Lexical lectometry in corpus-based translation studies. Combining profile-based correspondence analysis and logistic regression modeling. In M. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-based Translation Studies. A practical guide to descriptive translation research.* (pp. 325-345). Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Sutter, G., Plevoets, K., Delaere, I., Willems, A., Vandevoorde, L., & Prieels, L. (2014). Linguïstisch conformisme in vertalingen: een corpusgebaseerd multivariaat onderzoek naar lexicale variatie in het Belgische Nederlands Beschouwingen uit een talenhuis : opstellen over onderwijs en onderzoek in de vakgroep Vertalen, Tolken en Communicatie aangeboden aan Rita Godyns. Gent: Academia Press.

Delaere, I., & De Sutter, G. (2013). Applying a Multidimensional, Register-sensitive Approach to Visualize Normalization in Translated and Non-translated Dutch. Belgian Journal of Linguistics. *Belgian Journal of Linguistics, 27*, 43-60.

Delaere, I., De Sutter, G., & Plevoets, K. (2012). Is translated language more standardized than non-translated language? Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies., 24*(2), 203-224.

Prieels, L., Delaere, I., Plevoets, K., & De Sutter, G. A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation. *Across Languages and Cultures, 16*(1). Accepted for publication.