uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Donna A. Williams

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Translation Studies)

GRADE / DEGREE

School of Translation and Interpretation

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Recurrent Features of Translation in Canada: A Corpus-based Study

TITRE DE LA THÈSE / TITLE OF THESIS

Roda Roberts

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Lynne Bowker

Deborah Folaron

Barbara Folkart

Jean Quirion

Patricia Raymond

Gary W. Slater

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

RECURRENT FEATURES OF TRANSLATION IN CANADA: A CORPUS-BASED STUDY

by

Donna A. Williams

School of Translation and Interpretation
University of Ottawa

A dissertation submitted to
the Faculty of Graduate and Postdoctoral Studies
of the University of Ottawa
for the degree of Ph.D in Translation Studies

Your file Votre référence
ISBN: 0-494-11037-6
Our file Notre référence
ISBN: 0-494-11037-6

Approval Page

The undersigned certifies that she has read and recommends to the Faculty of Graduate and Postdoctoral Studies for acceptance, a dissertation entitled *Recurrent Features of Translation in Canada: A Corpus-based Study* submitted by Donna A. Williams in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dr. Roda P. Roberts
School of Translation and Interpretation
University of Ottawa

Date

Abstract

Based on the theory of translation universals, the general hypothesis that translated texts are distinguishable from non-translated texts by certain recurrent features of translation has been tested in recent contributions to Corpus-based Translation Studies. This hypothesis assumes that translation will leave similar traces in different languages.

Major corpus-based studies have recently investigated three specific hypothetical recurrent features of translation (normalization, explicitation, and simplification). However, each of these research projects has hypothesized only one recurrent feature of translation at a time, using mainly literary, Anglo-European corpora, and using English as the sole target language of the translated texts.

In the present study, all three of the above previously-studied recurrent features of translation are hypothesized and investigated, along with a fourth (levelling-out), which has not been the subject of previous study. Characteristics of translated and non-translated texts are compared in both English and French: appropriately for study of hypothetical "universal" features, the present research is carried out on target texts in more than one language.

Our corpora consist of texts taken from Government of Canada Web sites; they constitute a broad sample of non-literary texts. Specific techniques of analysis are adapted from the literature, and where appropriate, new techniques are devised. WordSmith (versions 3 and 4) was the primary tool used for corpus analysis.

The empirical evidence gathered in the present research supports the hypotheses of normalization and explicitation as recurrent features of translation into both English and French, but does not support the hypotheses of simplification and levelling-out. There is some indication that translated texts in both English and French tend to be more difficult to read (according to the standards of readability indices), an unexpected but interesting finding.

All of these results must be interpreted in the light of future corpus-based study of recurrent features of translation, and it is recommended that a standardized protocol for recording the attributes of future comparable corpora should be adopted.

Résumé

De récentes études en traductologie, qui s'inscrivent dans le champ des recherches basées sur le corpus, ont étudié l'hypothèse selon laquelle les textes traduits différeraient des textes non traduits en ce qu'ils présenteraient des régularités de la traduction d'une langue à l'autre.

Des études majeures se sont intéressées à trois régularités : la normalisation, l'explicitation et la simplification. Toutefois, chacune de ces études portent sur une seule régularité et s'appuient sur un corpus anglo-européen essentiellement littéraire, l'anglais étant la seule langue d'étude des textes traduits.

Dans la présente étude, nous nous proposons d'étudier de front ces trois régularités, auxquelles nous ajoutons une quatrième, le nivellement, qui n'a jusqu'ici fait l'objet d'aucune étude. Nous comparons les caractéristiques des textes traduits et non traduits en anglais et en français. Qui plus est, notre recherche porte sur plus d'une langue cible, approche nécessaire dans la mesure où nous revenons sur les « universaux de la traduction ».

Les corpus utilisés, qui se composent de textes émanant de ministères et d'agences du Gouvernement du Canada, constituent un éventail assez large de textes non littéraires. Pour les besoins de notre étude, nous nous sommes servis de techniques précises d'analyse de corpus tirées des recherches dans le domaine et, au besoin, avons élaboré de nouvelles techniques. WordSmith (versions 3 et 4) a constitué l'outil principal d'analyse de corpus.

Notre recherche empirique semble corroborer les hypothèses de normalisation et d'explicitation, régularités présentes dans les textes traduits en anglais et en français. Toutefois, elle infirmerait les hypothèses de simplification et de nivellement. Il semblerait de fait que les textes traduits, tant en anglais qu'en français, ont généralement tendance à être plus difficiles à lire (selon les indices de lisibilité), fait intéressant bien qu'inattendu.

Les résultats obtenus devront être interprétés à la lumière de corpus étiquetés beaucoup plus volumineux. Nous recommandons d'ailleurs que soit adopté un protocole standardisé, qui permettra à l'avenir de faire ressortir les caractéristiques de corpus comparables.

Acknowledgements

I am honoured to have worked with my thesis supervisor, Roda P. Roberts, and with the members of my reading committee, Lynne Bowker, and Barbara Folkart. If this work has any virtue, they must share credit for it; its flaws, however, are strictly my own. Dr. Roberts is the doctoral student's ideal champion. She was also kind enough to allow me the privilege of consulting TEXTUM.
Dr. Bowker and gave generously of her time and her innovative ideas, and offered me otherwise unobtainable copies of reading material. Dr. Folkart was a role model whose elegant theories innervate this work. The late Ingrid Meyer provided the original inspiration.

Patricia Raymond gave precious advice, particularly on the topic of readability. Jacqueline Bossé-Andrieu offered vital suggestions for measures of French. Miriam Shlesinger graciously sent me a copy of her thesis from Israel. Sara Laviosa took the time to read a chapter draft, and sent helpful comments. Timothy Stanley pointed out Canadian connections. Florence Lehmann translated the abstract with her usual finesse.

I am grateful to Dr. Ravi S. Pendakur, Department of Social Development, Government of Canada, to Dr. Gilles Lamothe, University of Ottawa, to Dr. Ian Bruce, Bell and Curve Statistics, Boston, and to Mr. Hasan Alam, Analyst, Human Resources and Skills Development Canada, for their advice in selecting appropriate tests of statistical significance, in setting up and applying those tests, and in helping with the interpretation of the results.

Without the texts contributed by many Government of Canada sources, this research would not have been possible. For going to the trouble of picking out the translated texts on their Web sites, and for providing precious information on how translation works at their government departments, special thanks to: C. Babineau, D. Barabé, J. Belanger, L. Bergeron, D. Blanchard, M. Boucher, P. Bowen, L. Brousseau, J. Bureau, S. Chartrand, N. Chow, S. Crawford, B. Dundas, S. Etco, P. Fedeski-Koundakjian, J.-M. Filion, G. Gagné, W. Gash, R. Gemme, D. Gratton, T. Harding, J. Humeniuk, E. Irwin, L. Jones, P. Kluver, V. Kohse, L. L'Heureux, M. Lloyd, T. Lessard, J. Merchant, C. Nault, S. Northrup, D. Pelchat, B. Pike, A. Poplawski , C. Purkhart, R. Quiney, S. Racine-Gibeault, A. Routliffe, D. Sarault, P. Seguin, D. Senécal, C. Tanner, D. Tardif, L. Thacker, M. Urquhart, E. Valceschini, and L. Wanczycki.

My predecessors, Michele and Andrew, were guiding lights. My father provided constant encouragement. My husband John gave unstintingly of his love and support.

To
M. J. Whistler
and
C. Baskerville Williams

Universalia notoria sunt singularibus.

List of Tables

1. Introduction

1.1 Topic

The discipline of Translation Studies has recently seen a surge of interest in **translation universals***,[1] a topic suited to the potentially immense scale of the electronic **corpus***. Earlier research in **Descriptive Translation Studies (DTS)*** emphasized the social and cultural context in which the individual translator worked. Translation was held to be governed by more or less consciously-followed behavioural **norms*** of translation that were proper to a particular literary culture or polysystem (Even-Zohar 1990: *passim*; Toury 1980: *passim*, 1981 *passim*).

This view focusses on each specific language and culture, assuming that in each, translation ranges along a behavioural continuum, from "adequacy" to "acceptability" (*ibid.*). Gradually (and perhaps prompted by the advent of electronic corpora), thinking on this subject began to shift from the specific to the general (Chesterman 2004: 33-41). The idea that there might be "general laws" of translation (Toury 1995: *passim*) began to take hold. Baker (1996) proposed a range of "translation universals" that might be accessible to observation using electronic corpora. The discipline of **Corpus-based Translation Studies*** (henceforth **CTS**) now gives researchers the opportunity for first-order observation of any "regularities" universally produced by the act of translation (Mauranen and Kujamäki 2004: 5, 9; Toury 2004: 28-29).

In other words, it is now possible to hypothesize and empirically investigate the possibility that there might be properties—general design features—common to all

[1] Certain specialized terms and abbreviations that may not be familiar to all readers are used throughout the present study. These are marked in bold face and with an asterisk (e.g. **term***) the first time they appear in the text, to indicate that they are defined in the appended glossary.

translation, regardless of the languages and cultures involved. This more recent view continues to take the specifics of language and culture into consideration; translation "universals" are rarely conceived as absolute. Furthermore, corpus-based study of "translation universals" is in practice a search for **recurrent features of translation*:** observable features that consistently recur in translated texts (Olohan 2002b). Recurrent features of translation must necessarily be considered relative, rather than absolutely universal, since they are necessarily affected to some degree by the linguistic and cultural context of translation (Tymoczko 1998: 654).

1.2 Motivation for Research

One general goal of research on "translation universals" is to accrue empirical evidence on a scale that is as wide as possible, pooling the knowledge gleaned from individual studies (Tymoczko 1998: 652). No single research project can come close to providing conclusive evidence on such a broad topic. However, it is hoped that over time, and with repeated corpus-based studies in many different languages, we can begin to understand how (or indeed whether) translated texts differ from non-translated texts. The general motivation of the present study is to contribute to this ongoing CTS research effort. By examining assumptions about translation "universals," by exploring the differences between translation and other forms of writing, and by testing the products of translation, which are observable first-hand (rather than by trying to test the translator's mental processes, which are not) we hope to add to the store of knowledge of what translation actually is.

The majority of prior research on "translation universals" has been carried out on literary texts published in Britain and Europe, and each project has been restricted to investigation of only one of the four "universals" proposed in Baker (1996). Our specific motivation for the present study is to add to the pool of research in this area, by gathering corpora of non-literary, non-European texts, and by using these corpora to investigate all recurrent features of translation for which a full-blown hypothesis has been proposed.

Four potentially "universal" recurrent features of translation (**normalization***, **explicitation***, **simplification***, and **levelling-out***) have been fully hypothesized by Baker (1996: 176-177; see also 179-185) as follows: "normalisation or conservatism" is the "tendency to conform to patterns and practices which are typical of the target language (henceforth **TL***), even to the point of exaggerating them"; explicitation is the "tendency to spell things out in translation, including in its simplest form, the practice of adding background information"; simplification is the "idea that translators subconsciously simplify the language or message or both"; levelling-out is the "tendency of translated text to gravitate around the centre of any continuum rather than move towards the fringes."[2]

While several large-scale research projects have uncovered some evidence related to these hypothesized characteristics of translation, each of these studies has been restricted to one of three recurrent features. Furthermore, as we will see in

[2] Laviosa-Braithwaite (1995: 162) also suggested "concretisation" as a possible "universal feature" of translated texts, following a personal communication with John Sinclair. Although no definition or hypothesis is given, Laviosa-Braithwaite suggests that "concretisation" could be investigated in electronic corpora by counting "the relative frequencies of abstract and concrete words" and the "relative frequencies of the concrete and abstract senses of polysemous words" (*ibid.*). However, in a personal communication with Laviosa-Braithwaite, Baker had argued that higher numbers of concrete words and senses of words would, if it were found, be "a facet of explicitation" (Laviosa-Braithwaite 1995). Laviosa-Braithwaite (1998) did not include "concretisation" among the translation universals subsequently proposed in her *Routledge Encyclopedia of Translation Studies* article, nor did she go on to study this proposed feature, perhaps because of the difficulty—and the subjectivity—involved in categorizing words as "concrete" and "abstract."

Chapter 5, the fourth hypothesized recurrent feature of translation, (levelling-out), has not yet been systematically investigated. Levelling-out will therefore be thoroughly explored in the present study. To the best of our knowledge, the present research is the first to investigate all four of Baker's (1996) hypothesized recurrent features of translation together.

The present research is "pure" (theoretical) rather than applied, and qualitative rather than quantitative. However, our methods and materials may prove useful to both translation scholars and practitioners. Corpus-based empirical research, which is fairly new to Translation Studies in Canada, may find encouragement in the present research. We hope to help refine the hypotheses and measures used in previous corpus-based research, and to develop new techniques for comparing corpora of translated and non-translated texts. In particular, we believe that adapting the measurement of **readability*** (through the use of the **readability index***) to the purposes of this study may suggest potential avenues of research toward the development of simple (but reliable and valid) methods of translation quality assessment. Finally, we would like to offer the corpora used in the present study as a potential pedagogical resource for translation instructors.

1.3 Research Objectives

The present research project has three specific objectives. The first is to achieve a better understanding of what translation actually is, by relying on first-hand observation. Second, our project seeks to widen the scope of inquiry into the nature of translation "universals," by exploring recurrent features of translation in greater

depth, by consolidating key elements of the topic into a single study, by using Canadian rather than European material, and by using non-literary rather than literary material in two languages of translation (English and French). Third, the present study aims to refine the methodology used in previous research and to devise new research techniques where appropriate, adapting those tools and techniques of **corpus linguistics*** that are particularly suited to CTS research. This approach, it is hoped, will help to clarify how "universal" the hypothesized universalia actually are.

1.4 Research Question

Our general research question is as follows: can the same recurrent features of translation be observed in two different TLs? By "recurrent features of translation," we mean specific linguistic features that have a higher frequency in translated texts than in non-translated texts, and that may eventually (in future study) prove to be independent of the characteristics of either the source or target language.

1.5 Scope of the Study

We have made our object of study and our materials as inclusive as possible in terms of the languages covered, the size of our corpora, and the number of recurrent features sought. Whereas previous corpus-based studies investigated single recurrent features of translation in only one TL, our study includes two TLs, and also consolidates all of the recurrent features that have been hypothesized previously in the major corpus-based translation studies of this topic.

However, the scope of our study is necessarily limited by practical concerns. According to Mauranen and Kujamäki (2004: 2), three "domains," the cognitive, the

social, and the linguistic, are relevant to the topic of translation "universals." We focus here almost exclusively on the third, linguistic domain.

In terms of the text characteristics studied, our research is restricted to some of the basic surface features of translated texts, such as word frequencies and sentence lengths. We do not examine more complex features, such as prosodic collocation. Wherever possible, we have included only linguistic features which appear to have counterparts in a number of languages. We have not attempted any replication of measures such as "strange strings" (Mauranen 2000), or "unique language items" (Tirkkonen-Condit 2000, 2004; Puurtinen 2003).[3] We have also categorized **coinages*** (which we deem likely to be ephemeral, as they cannot be proved to be lasting, in this synchronic study) according to two common word-formation strategies, compounding and derivation (Adams, Valerie 1973; Bauer 1983).

The TLs (languages of translation) included in the present study are French and English, the official languages of Canada. Our source of material limits the size of our corpora to some extent. Using Government of Canada Web site pages makes a large amount of text potentially available, and since these texts are in the public domain, it is not necessary to obtain permission to include them in this study. Locating texts that are comparable (having a similar range of topics, being of a similar length, and being taken from a single source) is also relatively easy. Since texts published by institutions connected with the Government of Canada must be made available in both official languages, according to the Official Languages Act (Government of Canada: 1985), it is a simple matter to find large numbers of texts in both French and English.[4]

---

[3] According to these translation scholars, "unique language items" are words that have no formal or semantic counterpart in other languages. This claim is based on Finnish-language studies only.

[4] Most, if not all, documents published by the Government of Canada will therefore necessarily be translated.

However, Government of Canada Web sites rarely note whether their texts are translated or "original." Before we could include a text in our study, it was therefore necessary to contact each department and obtain confirmation of the status of each text gathered. Because the present study was ultimately limited in terms of time as well as funding, the size of our corpus was limited to the number of texts whose status could be confirmed during the initial, corpus-gathering stage of our research (a period of about four months).

Material resources being limited, we did not have access to software that would give us a tagged or lemmatized corpus. For the same reason, the tools we used for corpus analysis (the WordSmith Tools package, Microsoft Office Word, Microsoft Excel, and a number of pen-and-paper readability formulas) were of necessity limited to those that were inexpensive and uncustomized.

1.6 Hypotheses

The present research is based on the following set of assumptions. First, we assume that translation is a type of reported speech, with the "reporting" done in a different language.[5] Second, since translation is a distinct linguistic behaviour, translated texts will have distinct characteristics. These characteristics will occur in translated texts regardless of the language involved. Furthermore, these characteristics will be observable and measurable: they will consistently recur in the surface structures of translated texts.

[5] Speech act theory has been applied to Translation Studies by Jakobson (1966), Mossop (1983 and 1998), Hatim and Mason (1990), Bell (1991), Folkart (1991), Baker (1992), Doherty (1993), Fawcett (1997), Gerzymisch-Arbogast and Mudersbach (1998), and House (2001). See Literature Review on Explicitation.

Our general hypothesis is that there are observable recurrent features of translation. To test this general hypothesis, we have formulated four specific hypotheses concerning the measurable occurrence of four proposed features of translation: normalization, explicitation, simplification, and levelling-out. Each of these specific hypotheses is discussed in the sections that follow.

It should be noted that, following Shlesinger (1989), Baker (1996), Laviosa-Braithwaite (1996), Kenny (1999b; 2001), Olohan and Baker (2000), and Olohan (2001; 2002; 2002b), but *contra* Puurtinen (2003), normalization, explicitation, and simplification are assumed to be properties which may be found in any text, whether or not it is a translation. For instance, some non-translated texts may adhere more closely than other non-translated texts to traditional rules of grammar or usage (i.e. they may be comparatively "normalized"). Some non-translated texts may have sentences whose structure is more clearly spelled out (being more "explicitated") than the sentences of other non-translated texts. Some non-translated texts may be less complex (i.e. "simplified") compared to others. When we hypothesize, below, that normalization, explicitation, and simplification are recurrent features *of translation*, we are predicting that these features will be more salient in (but not exclusive to) translated texts. Levelling-out is hypothesized to be a feature exclusive to translated texts.

1.6.1 Normalization is a recurrent feature of translation

> *Generally, normalization of vocabulary will be observable in both of the translated corpora included in the present study.*
>
> *Specifically, lower frequencies of unattested words and phrases will be found in each translated corpus than will be found in each non-translated corpus.*

Do translators tend to write more conservatively than authors, perhaps because they may be under greater pressure to produce "acceptable" (or marketable) writing?

Our general hypothesis is that the written texts produced by a population of translators will conform more closely to the norms prevailing for written texts in the language of translation, and that this will result in observably fewer instances of atypical or "non-conformist" usage in translated texts.

Our specific hypothesis is that we will find fewer examples in translated corpora of what will henceforth be referred to as "coinages": compounds and derivatives that have often (but not always) been formed according to recognizable word-formation patterns, and that are neither attested in dictionaries nor found in larger control corpora. Since it is not known whether any of these words will become part of the lexicon, we will consider the coinages discussed in the present study to be "transient" as in Valerie Adams (1973: viii) until future study can prove otherwise.

1.6.2 Explicitation is a recurrent feature of translation

Generally, explicitation of syntax will be observable in both of the translated corpora included in the present study.

Specifically, there will be higher frequencies of optional syntactic elements in each translated corpus than in each non-translated corpus.

In an attempt to clearly report what is written by someone else in another language, do translators tend to construct sentences with more explicitly elaborated syntax, perhaps intuitively parsing out deep structure components, as suggested by Folkart (1991: 132-134)?[6]

Our general hypothesis is that the written texts produced by a population of translators will contain more explicitated—more spelled-out—syntax, and that this will result in the observably higher use, in translated texts, of optional syntactic elements.

Our specific hypothesis is that we will find more instances, in our English corpus, of reporting verbs written with optional *that*, and that we will find more instances of optional *that* or *which* as the "Object of a Defining Clause" (henceforth **ODC***; note that defining clauses are also called "restrictive" clauses). We also predict that in our French corpus, we will find more occurrences of the "euphonic" consonant *l'* in *l'on*, and more occurrences of the "expletive" *ne* (i.e. of the *ne "explétif"*).

[6] Folkart (1991: 132) shows how a single complex sentence in a source text is broken down into eight simple sentences, each of which is then loosely re-combined into two translated sentences containing considerably more function words, and more explicit syntactic linking, than the original sentence. "Is syntactic explicitation perhaps a tangible artifact of a more or less intuitive syntactic processing which traces surface structures back to deep structures?" Folkart asks (personal communication).

1.6.3 *Simplification is a recurrent feature of translation*

Generally, simplification of vocabulary and syntax will be observable in each of our translated corpora.

Specifically, there will be:

(a) lower ***type/token/n ratios**** *in each translated corpus than in each non-translated corpus, showing each translated corpus to have a more restricted range of vocabulary.*

(b) lower proportions of ***content words**** *to* ***running words**** *in each translated corpus than in each non-translated corpus, showing each translated corpus to carry lower information loads.*

*(c) lower mean sentence lengths in each translated corpus than in each non-translated corpus, showing each translated corpus to have shorter clauses (**T-units***), and therefore simplified syntax.*

Do translators tend to write more simply than writers of non-translated ("original") texts? In an attempt to clearly report words written by others in another language, do translators tend to construct sentences using more basic vocabulary and with simpler syntax? Do translators perhaps intuitively use less complex sentence structures with more linking of elemental components, as suggested by Folkart (1991: 132-134)?

Our general hypothesis is that the written texts produced by a population of translators will have smaller vocabularies, fewer overall content words, and shorter sentences.

Our specific hypothesis is that we will find comparatively lower type/token ratios, lower proportions of content words to running words, and shorter sentences among translated texts.

1.6.4 Levelling-out is a recurrent feature of translation

> *Levelling-out, the probability that translation will exert an "equalizing effect" on the position of a text in a continuum, will generally be observable in each of our translated corpora.*
>
> *Specifically, within the pre-established statistical continuum provided by readability indices, translated texts will score as more "readable," and will have a narrower range of scores around their middle score value, than will non-translated texts.*

Is there less variety in translated texts compared to non-translated texts? Our general hypothesis is that, due to what Shlesinger (1989: 170-171) has called "the **equalizing effect***" of translation, the degree of similarity (the homogeneity) of groups of translated texts will be measurably greater than that seen among non-translated texts. Translated texts will tend to have observably less variation in traits qualified along a given continuum.

Our specific hypothesis is that when readability indices are applied to them, translated texts will generate readability scores with a narrower range of highs and lows (along the continuum ranging from "very easy" to "very difficult" to read) than will the readability scores of non-translated texts. We also hypothesize that translated texts will tend to score as more "readable."

1.7 Overview by Chapter

The present study is organized in the following manner. Chapter 2 presents the theoretical literature and previous research on "translation universals," or recurrent features of translation. In Chapter 3, the corpora gathered and the methods of corpus analysis used are described in detail. Chapter 4 describes the present study's empirical investigation of the three hypothesized recurrent features

(normalization, explicitation, and simplification, respectively) for which there are data taken from other corpora. Chapter 5 describes the empirical investigation of a fourth hypothesized recurrent feature of translation, levelling-out, for which, to the best of our knowledge, there is no data from previous CTS research. A description of how readability indices were used to investigate levelling-out is also included in Chapter 5. Finally, in Chapter 6, the results of the entire study are summarized and discussed, as are the ramifications of the present findings for future study of this topic.

2. Literature Review

What sets translation apart from other types of writing? There has been much discussion of this question over the years, although empirical research on the topic of "translation universals" dates mainly from the last decade. We include in this review some theoretical research that depends on a comparison of translated texts to source texts, where universal features of translation are assumed to fall under the heading of "shifts" in the translated or "target" text (henceforth **TT***) compared to the "original" or source text (henceforth **ST***). This view is summarized in Toury (2004: 20-25).

This more traditional perspective notwithstanding, there does seem to be some agreement in the literature that translation is a distinct type of speech act in itself, with features that set it apart from other types of speaking or writing in any given language. There appear to be a multitude of opinions on the exact nature of these distinguishing features. Mauranen and Kujamäki (2004: 2) have noted that theoretical discussion has covered many aspects of this question, including the "plausibility, kinds, and possible determinants of universal tendencies of translation."

Mona Baker has offered clearly stated hypotheses as to the kinds of translation universal that exist. According to Baker, there are at least four recurrent features of translation: normalization, explicitation, simplification, and levelling-out (1996: 176-177; 180-185). As noted in Section 1, normalization, explicitation, and simplification have been the object of much scholarly discussion. In DTS,

frequent mention is made of the strict adherence to linguistic norms that renders translated texts "conventionalized" or normalized.[7] Along with a number of other theorists reviewed below, Toury calls explicitation a translation universal (1980: 60).[8] Ria Vanderauwera posits simplification as a universal feature of translation (1985: 97-99) along with normalization and explicitation (1985: 76). The fourth recurrent feature of translation hypothesized by Baker, levelling-out, is a relatively new concept in Translation Studies. Baker appears to have been the first to fully formulate it as a hypothesis, predicting that large translation corpora will prove to have distinctly "regular" statistical profiles compared to large non-translated corpora (1996: 177; 184-185).

A general overview of the theory and corpus-based empirical study of recurrent features of translation is provided in Kenny (2001: 22-47; 1999b: 30-46) and in Laviosa (2002: 42-86). In what follows, we will review and update relevant aspects of the literature, providing a theoretical framework for the present study. We will also review the evidence available from previous empirical testing of these hypotheses.

2.1 Normalization

Translation is a type of reporting: to translate is to report something that has already been written in another language (Jakobson 1966; Mossop 1998 and 1983; Folkart 1991). Possibly as a result, translation is familiarly conceived as writing that is not original (witness the common description of the ST as "*the* original"). It

[7] Even-Zohar (1990: 46; 48-50); Toury (1977: 6-34); Toury (1978); Toury (1980: 51-70, 122-151); Toury (1981: 23-24) and Toury (1995: 53-69 and 267-274); Weissbrod (1992 and 1992b); Du-Nour (1995).

[8] "There is an almost general tendency—irrespective of the translator's identity, language, genre, period, and the like—to explicitate in the translation information that is only implicit in the original text." Simplification is also implied in Toury's listing of omission, shortening, "implicitation," and "condensation" among typical translation solutions (1980: 101-102).

follows that translation is commonly considered to be a type of writing that lacks creativity. Below, we will review the literature on the hypothesis that normalization, a theoretically exaggerated adherence to norms (to the detriment of any kind of creativity, such as the attempt to coin words), is a recurrent feature of translation.

2.1.1 Theoretical background

Laviosa-Braithwaite's listing of normalization among the possible universals of translation in the *Encyclopedia of Translation Studies* article on "Universals of Translation" (Baker, ed. 1998: 289-290) is an outcome of two decades of scholarly discussion centred on the effect of linguistic and cultural norms on the production of translated texts. The hypothesis of normalization as formulated by Baker (1996: 183) states that the translator's tendency to "conform to the typical patterns" of the TL can be seen in translated texts.

Theo Hermans notes that so much has been published on the subject of norms of translation that it is "safe to assume" most Translation Studies scholars will be familiar with the concept (1999: 50). The topic of translation norms has been reviewed extensively in the literature (see Baker 1998: 163-165; Blum-Kulka 1986; Chesterman 1993; Harris 1990; Hermans 1996 and 1999; Kenny 1997, 1998d, 1999b, 2000, 2000b, and 2001; Laviosa-Braithwaite 1998: 289-290; Nord 1991 and 1997; Schäffner, ed. 1999; Snel Trampus 2002; Toury 1980 and 1995; Millán-Varela 1999; Øverås 1998).[9]

[9] In the Bibliography, Scandinavian names are listed using Scandinavian alphabetical order: names containing letters with a dieresis (e.g. Jääskelainen, Øverås) are listed at the end of a series. The entry for Øverås is therefore listed after names starting with the letter "Z."

We can make three theoretical assumptions about normalization based on the above literature:

(a) That the norms for writing in a language generally apply to translating in that language.
(b) That translators are usually expected to conform more closely to the norms of a language than are other writers.
(c) That some element of creativity is nonetheless involved in the act of translating. Translators will innovate some of the time, but to a lesser extent than do writers.

In *Norms of Language*, Renate Bartsch (1987: 4) describes linguistic norms as notions of speech behaviour "correctness" that are a social reality in a speech community. If we agree that writing—whether it involves translating or not—is a type of speech behaviour, and that translators and other writers are part of the same speech community if they belong to the same culture and are working in the same language, then we can consider translation norms to be a subset of language norms.[10]

The Translation Studies literature on norms reiterates that where the rules of language use are concerned, a greater degree of conformity is seen in translators than in other writers. We can regard normalized translation as one of many types of conventional use of language if we accept that prevailing rules for writing apply to both translators and writers, and that the former behave as if they received more pressure to conform than the latter.[11] If normalization is a recurrent feature of translated texts, the social context that produces this feature is at least in part explained by Toury's "law of growing standardization" (1995: 267-268).

---

[10] Bossé-Andrieu (1994) provides an overview of the "social reality" of contemporary French language norms. There are many reviews of contemporary English language norms, but see especially Milroy and Milroy (1998).
[11] Kenny notes (1998d: 1) that "pressure to conform" applies to all writers, whether they are translating or not: the reading public is believed (by publishers) to have a rather low tolerance for idiosyncrasy in any written text, not just in translations.

Especially in societies that consider translation "peripheral," that is, of marginal importance, translators will tend to "accommodate" established norms to a greater degree (Toury 1995: 271). Translation norms theoretically "propagate" translation behaviour that favours convention ("stock equivalents" Toury 1980: 108) over invention ("living metaphor" Toury 1995: 274) in the translated text.

Kirsten Malmkjaer (1998) notes that "hyper-typicality" is a common feature of her corpus of translated texts. Her corpora of translated texts regularly reproduce what she calls the "Eliza Doolittle" phenomenon: "the translated texts, more than the texts in the control corpus, tend to contain those [TL] phrases, structures, and so on, which, from a comparative point of view, seem particularly characteristic of the [TL]" (1998: 2).

In DTS, a corollary of the assumption that translators are more conformist than (literary) authors has been the notion that translators will be less innovative in their writing. In this view, authors develop new norms and accept new influences from other languages before translators do.[12]

A stronger version of this view is that translators' writing will be old-fashioned, following behind the current norms of a literary polysystem (Even-Zohar 1990: 48-49). Translated literature, Even-Zohar says, is always "modelled according to norms already conventionally established" and is in fact a "major factor of conservatism" in the literary polysystem of a culture. Rachel Weissbrod (1990), looking at a corpus of literary translation into Hebrew, traces the influence of "English norms" on the Hebrew target polysystem as a whole, showing how they

[12] See also Apter (1987) on the apparent "compulsion" of Victorian English literary translation to emulate source language form. She also notes that Ezra Pound, who reacted strongly against this Victorian aesthetic, still considered translation to be "preparation for writing" (1987: 132) and therefore not creative or "Original" in itself.

first affect non-translated literature, and how they are only later adopted in translated literature.

Kenny (1998d) proposes "sanitisation" as a corollary of lexical normalization. At points where ST collocations evoke difficult-to-read, controversial, ironic, or unpleasant (semantic) associations, their corresponding TTs may tend to switch toward more palatable imagery and metaphor (1998d: 1, 6), a shift similar to bowdlerization, but lacking the deliberate, policy-driven prescriptivity of censorship.

Along with the acknowledged conservatism of translation, however, a number of scholars recognize the persistence of a certain degree of creativity. According to Toury, one of translation's universal characteristics is its variety, or "variability" (1995: 31). Minting unique forms instead of duplicating accepted ones can be considered a form of innovative translation.[13] It is reasonable, then, to expect to find a number of coinages in a corpus of translated texts. This type of creativity in translated texts could be considered the "flip side" of their normalization, the exception, as it were, that proves the rule.

Baker makes a direct link between creativity and non-normative translation choices, noting that the latter are not necessarily erroneous, and that they are not insignificant to Translation Studies:

> ... la tentation est forte de mettre l'accent sur la norme, sur ce qui est spécifique, aux dépens ... de l'utilisation de la langue la plus créative, au point, parfois, de décrire cette utilisation comme «fautive» parce qu'elle dévie de la norme. ... Les structures constituent la toile de fond qui donnera forme à la créativité : les normes autorisent [aussi] l'usage créatif de la langue .... (1998: 5)

[13] In Toury's terms, the use of unique "textemes" rather than the selection of set "repertoremes" (1991: 187).

Translators may, as Baker has hypothesized, make certain sets of choices consistently. It does not follow that these choices are uniformly and without exception conservative, or that they always have a stultifying effect on the literary and linguistic polysystem.[14] When permutations of forms—or even new forms—enter a language through translated texts, all that can be reasonably inferred is that unusual forms have been used in that language.

We may assume, then, that creativity and conformity can co-exist as tendencies in translated texts (just as they can co-exist in non-translated texts). It is possible to observe concrete evidence of this. In any population (such as a set of translated texts), conflicting statistical trends can co-occur and provide "counterexamples" to each other (Laviosa 2001: 77). In Kenny's (1999b; 2001) study of literary translations, a small set of exceptionally "creative" renderings stands out against the predominantly conventionalized wording of the overall translated corpus.[15]

In the literature on translation norms, we have noted a general assumption that translators are conformist, following language norms more closely than authors do. The hypothesis that there will be observable evidence of normalization in translated texts, whether literary or non-literary, emerges from this idea. What evidence of normalization can be found in a translation corpus?

---

[14] Ladmiral (1991: 25) dismisses the notion of "contamination" of a language through translation, pointing out that inter-linguistic contact is often a trigger for the actualization of latent structures in the language, for the realization of what Folkart (in a personal communication) has called a language's "virtual reserve." See also Folkart (1991: 155-156), on the competent translator's range of choices when faced with a "case vide référentielle" in the target language.

[15] Of the six literary translators whose group tendencies are compared in Kenny (1999b and 2001), three were found to be "very likely" to normalize source text forms such as "odd collocations" and idiosyncratic words, while one exceptional translator "avoided normalization as a translation strategy," and produced creative solutions in keeping with the "avant guarde" expectations of his publishing house (Kenny 2001: 142-188, 208-209).

Morphology and syntax would, Baker (1996: 183) predicted, prove in empirical study to be consistently more conventional in translated texts than in non-translated texts. Malmkjaer (1997) and Øverås (1998) maintain that in literary translations, the wording is more likely to be conventional, with fewer unusual word combinations. Even when unusual word combinations (strange strings) are used in translated texts, Mauranen (2000), one of the few to discuss non-literary ("academic" and "popular non-fiction") texts, believes that these unconventional phrases will be combined out of a smaller number of words—a more restricted vocabulary.[16] Munday (1998) believes that (literary) translated texts are more likely to have conventional word order.

But creativity in some form is also predicted by several Translation Studies scholars. Blum-Kulka and Levenston (1983) as well as Kenny (1999b; 2001) believe that word coinage, creative orthography, and word play are quite likely to be present in literary translation.[17] Ladmiral (1991: 25) notes that coinages which might be criticized, in a literary translation, as evidence of the undesirable influence of one language on another (e.g., as *anglicismes* in French), may in fact be formed out of the "reserve" of potential forms that are "latent" in a language.[18] As we will see below, most of the research on normalization in the literature has focussed on vocabulary, specifically comparing the relative proportion of unusual words in non-translated and translated texts that are literary.

---

[16] "Even if the combinations are unusual or infrequent in original texts, the individual items that constitute the combinations [in translated texts] may still be frequent." Mauranen (2000: 137).

[17] The examples of translation given in Blum-Kulka and Levenston (1983) are mainly literary and Biblical.

[18] Attempted coinages may be well formed by the target language's rules of word-formation, even when they appear similar to a word or phrase in the source language: see Ladmiral's (1991: 25) discussion of "le français possible." Attempted coinages may be considered more "acceptable" if they resemble words in a prestigious language: Toury notes that tolerance of such "interference" tends to increase when translation is carried out from a "prestigious language" (1995: 278). Furthermore, an individual speaker's introspection—his or her judgement of how well formed and acceptable a writer's attempted coinage is—may be highly subjective. Mauranen (2000: 138) observes that "a speaker's intuition as to what is typical in his or her language tends to be much more uncertain than a speaker's intuition about what is possible...."

2.1.2 Empirical research

A number of studies from the 1980s and early 1990s report finding evidence of normalization of translated texts, including the use of rare words in the , the shifting of phraseology toward the idiomatic and familiar, and the suppression or omission of aspects of the source language (henceforth **SL***) or ST that were considered unacceptable or too difficult to translate.

Toury (1980: 128-130) observes that in his corpus of modern Hebrew literary texts, the overall translation is noticeably "conventionalized." Particularly noted is Israeli translators' avoidance of neologism (the creation of new words) by locating extremely rare words (**hapax legomena***) in ancient canonical Hebrew texts, and by then using these rare words in lieu of coinages (Toury 1980: 148). Similarly, Gellerstam (1986) reports that in his corpus of translated Swedish novels, rare Swedish words are used repeatedly to translate words that are common in the language of the original (English), allowing the translator to avoid coining words (1986: 91-92).

Ria Vanderauwera (1985: 76-77; 93) sees a general tendency for her Dutch literary translated texts to shift toward "textual conventionality" in such a way as to make them more readable, idiomatic, and familiar than their originals. She notes, among the translators' more normalizing efforts, their "systematic attempts to suppress all kinds of irregularities, smoothen [sic] out unusual style and rhythm, and remove irrelevant fragments" (Vanderauwera 1985: 72-73). Kitty M. van Leuven-Zwart (1989 and 1990) found that a key feature of her corpus of (mainly Dutch) literary translations was a series of micro-shifts that cumulatively push

whole translated texts toward the "acceptablity" end of Toury's range of initial norms (1990: 86-92).[19] Shlesinger (1991: 150-151; 153) found that court interpreters working in Hebrew were very likely to omit inappropriate personal comments made in the courtroom by the people whose speech they were interpreting. Both pejoratives and compliments had a high probability of being dropped: "laudatory remarks" made about the interpreter were likely to go untranslated, as were insults directed at anyone (Shlesinger 1991: 153).[20] This is possibly an example of "sanitisation," defined by Kenny (1998d: 1, 4, 6) and noted above.[21]

Some word use regarded as innovative (rather than as normalized) was also found in earlier studies of translation. Blum-Kulka and Levenston (1983: 126-127; 136-137) tested the strategies involved when translators must deal with cross-language vocabularies that are in some way mismatched. They concluded that translators working with Hebrew and English would use a strategy of "word coinage" to respond to "semantic voids" in their TL (126; 132-134): translators did not rely on existing vocabulary, but instead invented new words.

During the 1990s, extensive statistically-based empirical studies became possible with the growing use of electronic corpora. Kenny (1999b, 2001) studied

---

[19] Van Leuven-Zwart coined the term "microshift" in this study. She predicted that her corpus of translated texts would show "differences or *shifts*" (1989: 153) that would occur "on the microstructural level, i.e. the level of sentences, clauses and phrases" (1989: 154). These would indicate the translation norms involved (1989: 151). Most of the microshifts she found were clearly instances of explicitation. However, their cumulative impact was one of normalization, in her opinion.

[20] Depending on the culture involved, interpreters may make a number of such omissions during the course of a court hearing. The extent of parenthetically informal talk in the North American courtroom, where very formal, stylized and often formulaic speech is the norm, is made clear in Gaines (2003). It may be inferred from Shlesinger's findings that "informal asides"—remarks that fail to adhere to the normatively formal register of the North American courtroom—stand a fairly high chance of being omitted by courtroom interpreters working there.

[21] In a study of the translation behaviours of Finnish, Swedish, and German translators writing subtitles for an American English film, Jaskanen (1999: 44-73) found that sexual innuendo in the script was not usually suppressed in translation, whereas culturally-specific references (e.g. to commercial products not available in Europe) were usually "naturalised." An attempt was usually made to translate even the earthiest of humour. This example may help to illustrate the difference between Kenny's "sanitisation" and bowdlerization: the former seeks to avoid making the TL reader uncomfortable where the SL reader presumably would not be; the latter seeks to censor without regard for the TL reader's presumed "range of good taste."

how British literary translators had handled "creative" words and phrases found in German literary STs. Kenny found that in her corpus of translated texts, the overall number of "creative" words that had been translated in a non-creative, normalizing way was greater than the overall number of instances of "creative" translation (1999b: 116). This did not, however, contradict Kenny's general assumption that there is sometimes an element of creativity in translation: there was evidence of both normalization and creativity in Kenny's corpus of English literary translations.

The continuum of initial norms, with its two "polar opposites" (Toury 1980: 54-55) of adequacy and acceptability, is observable in Kenny's findings: one translator out of the six studied produced consistently "creative" solutions, while another (out of the six) produced almost nothing but normalized solutions (2001: 181-183; 208-210).[22]

A number of recent small-scale studies have found evidence of hyper-conventionality in translated literary texts. Munday (1998) did a computer-assisted study of normalization in one Spanish-English translation of a single short story. He concluded that the translation contained "numerous" shifts in the use of possessive pronouns (1998: 7-8), and that in general, there was a trend toward normalizing shifts in the sample (1998: 14). Øverås (1998), whose manual study of a **parallel corpus*** (English-Norwegian/Norwegian-English) of 40 excerpts of literary texts was primarily an investigation of explicitation in translation, nevertheless arrived at a conclusion related to normalization. She found a general tendency in her corpus toward phraseological shifts from "collocational clash to

[22] The "creative" translator, Kenny notes, was working for an avant-garde small publisher, while the "normalizing" translator worked for mainstream multinational publishers (2001: 209-210). That the latter was also a recipient of several major translation prizes throws light on how the "norm of acceptability" seemingly prevails in English: translators who adhere to it are apparently rewarded. For more on this topic see Venuti (1998).

conventional combination" (Øverås 1998: 10-12 and 16).[23] Malmkjaer (1998) also provided evidence of normalization in translation when she compared ten different translators' renderings of one unusual ST phrase in a Danish children's story, and found that most of the translations used a conventional phrase in its place (1998: 4-5). Hansen and Teich (2001) found evidence of normalization (2001:7) in their very small (20,000 word) German corpus. They assumed that extensive use of German-language alternatives to the passive voice[24] is a normative feature of scientific texts in German, and found that these "passive alternatives" were used "significantly more frequently" in their German translated texts than in their German non-translated texts, showing, they claimed, that the translated texts were normalized (*ibid.*).

Apart from Kenny (1999b; 2001), to the best of our knowledge only one other large corpus-based study has looked for innovation in translated texts.[25] Mauranen (2000: 122) hypothesized innovative translation: she predicted that in her large-scale corpus of Finnish-language "academic prose" and "popular non-fiction," translations would contain more unusual word combinations than comparable non-translated texts. However, she had mixed findings. While there was some unusual "lexical patterning" in the translated texts (Mauranen 2000: 136), the individual words that constituted those patterns were more common than the words that constituted the "lexical patterns" in the non-translated texts (137). The vocabularies of the translated texts were in fact a "fair indicator of

---

[23] For example the unusual collocation "plump tunes" in the English ST was translated as "*trevlige melodier,*" a more conventional collocation like "pleasant tunes," in the TT.

[24] German grammatical constructions that are not passive voice, but that like the passive voice leave the agent unspecified or "underspecified" (4-5).

[25] Mauranen's (2000) corpus consisted of two million words of non-translated Finnish texts, and 1.6 million words of comparable translated Finnish texts.

conventionality," Mauranen concluded (*ibid.*). Her findings thus supported the hypothesis of normalization.

In sum, previous Translation Studies research has found evidence of normalization in translated texts. While the pre-1990s research examined translated texts in several different languages (Hebrew, Dutch, Swedish), all the corpora studied were very small, and the studies were not carried out as systematically as is possible with the corpus analysis tools available today. Since 1990, some major corpus-based research has been carried out, but on only one TL per study, and primarily (though not exclusively) on literary texts. Large-scale corpus-based study of normalization in at least two languages of non-literary translation still remains to be done.

2.2 Explicitation

The belief that translation tends to be more explicit than non-translated speech or writing appears to have been an enduring one in Translation Studies. Like normalization, this is a concept with which many translation scholars are likely to be familiar, and a number of versions of it have been reviewed elsewhere (Englund Dimitrova 2003; Delisle, *et al.* 1999 *passim*; Klaudy 1998; Laviosa-Braithwaite 1998; Shuttleworth and Cowie 1997). The present review focusses on the theoretical and empirical research pertaining to one particular version of this concept, what Klaudy (1998: 83) calls "translation-inherent" explicitation, which is attributable to "the nature of the translation process itself," as opposed to explicitation made necessary merely by the presence of some feature of the source

or target language. As we will see below, the discussion of explicitation generally describes it as both intrinsic to the process of translation and attenuated or intensified by differences between individual pairs of languages.

In what follows, we will review the theoretical literature relevant to a discussion of explicitation as a hypothetical recurrent feature of translation. We will then summarize the empirical studies that have tested this hypothesis.

2.2.1 Theoretical background

The inclusion of explicitation among the "Universals of Translation" in the *Encyclopedia of Translation Studies* (Baker, ed. 1998: 289) was an outcome of decades of theoretical research on the subject. As early as 1923, Walter Benjamin conjectured that there has, since Babel, been a hidden "reciprocal" relationship among all languages which translation renders explicit (in Schulte and Biguenet 1992: 74). Roman Jakobson, who ranks translation alongside other types of reported speech (1966: 233), says that explicitation results from the process of mentally formulating a translated "report," although the translator may consider greater or lesser degrees of explicitness appropriate, depending on the context (*ibid.*). Michael Rifaterre (1992: 205-208; 215) describes the search for equivalence in literary translation as a process of making explicit the implicit "presuppositions" of the writer and the translator.

Barbara Folkart (1990: 39; 1991: 209), who builds on Jakobson's (1966) and Mossop's (1983) definition of translation as reported speech, notes that the process of translating brings an enhanced functional "value" to the translated text. The process of translating may also activate latencies in the "*marge originaire*," the

range of connotations of the ST (Folkart 1991: 217-218; 446-447); this can be seen as explicitation of the overall enunciation. Mossop (1983: 259-264) describes the avoidance of "undertranslation" through "the addition of material" that is "appropriate to the pattern of meaning" without imitating the "compositional structure" of either the source or target language. In doing so, he appears to be proposing something similar to the hypothesis of explicitation.

Vladimir Nabokov (1992: 142-143) provides what is perhaps an extreme view of explicitation as a necessary part of the process of translation. Nabokov (1992: 141) advocates making absolutely explicit all that is implicit in the ST or that cannot be conveyed literally in the TT, by translating the body of the text very literally and then parsing out the entire remainder (whatever cannot be rendered literally) in footnotes.[26] These footnotes may take up more space on the page than the TT itself.[27]

Marianne Lederer (1994: 56-57) believes that the degree of explicitation will vary from language to language: "Jamais dans plusieurs langues l'explicite n'est le même pour renvoyer à la même unité de sens."

A complete hypothesis of explicitation was formulated by Shoshana Blum-Kulka (2000: 300), who posits that there will be an observable increase in explicitness in TL texts compared to SL texts, and that this explicitation is inherent in the process of translation. The increase in explicitness will occur beyond any that is traceable to differences between the two languages, she predicts (Blum-Kulka

[26] For a discussion of the theoretical "remainder" in translation, see Lecercle (1990) and Venuti (2002: 219-220), who notes that "translators [...] can never entirely avoid the loss that the translating process enforces on the foreign text, on its meanings and structures [...]. And translators cannot obviate the gain in their translating, [...] the creation of textual effects that go far beyond the establishment of a lexicographical equivalence to signify primarily in the terms of the translating language and culture. [...] I call these effects the 'remainder' in a translation."

[27] Nabokov declares, "I want translations with copious footnotes, footnotes reaching up like skyscrapers to the top of this or that page so as to leave only the gleam of one textual line between commentary and eternity" (1992: 143).

2000: 304). She recommends empirical studies to distinguish between explicitation that results from differences in languages versus explicitation that is the result of a process which is "inherent to" the effort of translating (Blum-Kulka 2000: 300; 312).[28]

Explicitation may be an intrinsic part of the process of translation, but according to a number of translation scholars, it can be consciously controlled and directed by the translator, to whom an arsenal of strategies is available.[29] One of the most complete catalogues of these strategies is offered by Vinay and Darbelnet (1995; c. 1958: 339; 342), who list "explicitation" among their numerous techniques of translation. Several other translation strategies that closely resemble one another, and that fit the hypothesis of explicitation, are also described by Vinay and Darbelnet.[30]

Eugene Nida (1964: 227-229) lists "amplification from implicit to explicit status" among his translator's techniques. Literary translator William Weaver (1989: 119-122) recommends the occasional "boosting" of sentences or "swelling" of a clause by "amplifying" it when the context makes this "permissible." Hervey and Higgens discuss the option of unpacking implicit meaning as part of a process of "particularizing translation" in their manual of translation techniques (1992: 95-99). Mildred Larson (1998: 41-46, 492-499) devotes considerable space to the translator's explicitating treatment of "implicit information which is referential,

---

[28] Blum-Kulka also maintains that more experienced translators will learn to hide the signs of this effort, and that there will be more observable explicitation in the work of self-trained or trainee "non-professional" translators (*ibid.*).
[29] Lörscher (1991:76) defines translation strategies as "potentially conscious" procedures solving problems which an individual faces "when translating a text segment from one language into another."
[30] Among these are "amplification" (1995: 339), a translation technique by which a greater number of words is used in the translated text to express "the same idea" as that found in the source text; "dilution," the spreading of "one meaning" (presumably expressed by one word in the ST) over "several lexical items" (1995: 341-342); and a "special case of amplification," called "supplementation" or "*étoffement*," the inclusion of a word or phrase to augment the idiomaticity of the translated text (1995: 350).

organizational, and/or situational" in her guide to cross-language equivalence. Jeremy Munday (2001: 101) notes that 'mismatches' supposedly discovered between ST and TT in Juliane House's model of register analysis (House 1977 and 1997) may in fact be something other than translation errors. They may simply be evidence of a conscious strategy of explicitation, Munday maintains (*ibid.*).

We have noted, in the literature, a prevailing assumption that there may be a greater tendency to "spell out" (either consciously or unconsciously) in translation what might in original writing be more often left implicit. If this is the case, then there should be observable evidence of explicitation in translated texts compared to non-translated texts in the same language. What concrete evidence of explicitation might we find in a corpus of translated texts? A number of suggestions are made in the literature. Vinay and Darbelnet (1995; c. 1958) suggest that it will be possible to identify "explicitation of pronouns" (116) and the avoidance of structural ambiguity through explicitation of the translated text (185).

Baker (1996: 180-181) predicts that both morphology and syntax will be subject to explicitation; she specifically recommends that studies compare the use, in translated texts, of "explanatory vocabulary and conjunctions" such as *cause, reason, due to, lead to*, *because, therefore, consequently*, and of semantically void optional words such as *that*.

Scott Burnett (1999), Olohan and Baker (2000), Silvia Hansen and Elke Teich (2001), and Vilma Pápai (2004: 155) all consider greater TT use of semantically void optional words and phrases (e.g. more instances where the optional conjunction *that* [English] *daß* [German] or *hogy* [Hungarian] is included

instead of the "zero form" being used) to be potential evidence of explicitation in translated texts.[31]

Øverås (1998: 45-7) cites such evidence of explicitation in parallel corpora as the insertion into the TT of conjunctions (e.g. "long, slender" rendered as "*langt og smalt*") and of pronouns ("*det gamle ansiktet til bestemor*" [literally "the old face of grandmother"] rendered as "the aged face of his grandmother"). Snel Trampus (2002: 52) suggests that fewer metaphorical phrases will be found in translated than in non-translated texts, and regards this as evidence of explicitation, assuming, contra Lakoff and Johnson, that it makes the TT easier for the reader to process and interpret.[32]

We can see in the above literature that some translation scholars are preoccupied with the individual ST and its explicitation in translation, while others point to how explicitation marks TTs regardless of the features of their STs, apparently as a result of the process of translation.[33] Below we will see that some of the most recent research has emphasized the latter approach, comparing corpora of translated and non-translated texts.

---

[31] It should be noted that "zero-*daß*" (e.g. *Ich glaube, es ist hier* versus *Ich glaube, daß es hier ist*) may be less frequent in German than is "zero-*that*" in English. With "zero-*daß*" the verb is re-positioned, making this a more complex syntactic construction than the use of "zero-*that*" in English.

[32] Lakoff and Johnson (1980; 1989; 1999) make the opposite argument, that metaphor is a cognitive organizer that allows abstract or complex concepts to be more easily processed and understood. They also argue that metaphor is fundamentally conceptual, not linguistic, and that metaphorical language is a surface trace of thinking organized by underlying metaphors. They are, of course, speaking of language in general, and have not applied their theory specifically to translation, although there appears to be no hindrance to doing so.

[33] Chesterman (2004: 39-40) suggests that this difference in focus can be used as a basis for categorizing the recurrent features of translation. Evidence observed in parallel corpora of the "way in which translators process the source text" should be called "S-universals," while evidence observed in comparable corpora of the "way translators use the target language" should be called "T-universals," according to Chesterman (2004: 39). We would argue that this distinction is somewhat artificial, and potentially confusing: to translate is in fact to "process the source language" in the target language. Unless one is purporting to speculate about the unknowable "black box" of the translator's mind, it will be most difficult to provide concrete evidence of a taxonomic difference between "S-" and "T-" universals.

2.2.2 *Empirical research*

As with normalization, the research on explicitation has emphasized the surface features of the translated text. Much of the research on explicitation has been carried out on vocabulary. Blum-Kulka and Levenston (1983: 126-127; 136-137) administered cloze tests to student translators in order to discern strategies involved when source and target vocabularies are in some way mismatched.[34] They concluded that novice translators would respond to gaps in TL vocabularies by increasing the number of words used while possibly reducing the information load of the overall text (*ibid.*). Ria Vanderauwera (1985: 97) found "minor explicitations" in the wording of English translations of Dutch literature.[35] This explicitation was both semantic (1985: 81, 93, 96-97, 106-107) and syntactic (1985: 48-49, 79, 98, 101, 105). Information deducible from the ST context was "often made explicit or repeated in the target text," Vanderauwera noted (1985: 73). Based on her findings, she predicted that explicitation would prove to be one of the most common practices of translators into any language (Vanderauwera 1985: 113).

Øverås (1998: 16-17) found evidence confirming the explicitation hypothesis, in her manual study of a parallel corpus of 40 literary excerpts in English and Norwegian. Øverås sorted evidence of explicitation into two types:

(1) "addition," in which conjunctions are added to the TT, hypotactic syntax is replaced by parataxis, or "pro-adverb" insertions are made (1998: 5-6; 8)

---

[34] The use of approximate equivalents, superordinate terms, synonyms, and converse terms (e.g. "she is not married" for "she is single") were among the strategies adopted by trainee translators (*ibid.*).

[35] "A whole series of minor 'explicitations' and 'corrections' streamline the target renditions," writes Vanderauwera (1985: 97). For instance, *de kennis* (literally "the acquaintance") is translated as "his non-Jewish friend," and "*alleen viel je op aan de Via Veneto*"(literally "alone felt I out on the Via Veneto") is spelled out as "a woman by herself looked conspicuous on the Via Veneto" (*ibid.*). In each case, the target text becomes longer as the meaning of the source text is reported.

(2) "specification," in which determiners (articles, demonstratives, or possessives) are added to the TT, or in which a noun is substituted for pronoun (1998: 7; 10).

Øverås also counted the number of instances of each type of grammatical feature said to represent an explicitating "shift" in her two translation corpora (of translated English texts and of translated Norwegian texts respectively). She found that there were 4.9% fewer instances of the grammatical features deemed to be instances of explicitation in the English translated corpus than in the Norwegian translated corpus. However, no comparable same-language texts were included in the study, so it is not possible to say whether these results reflect differences due to translation.

In her study comparing Finnish, Swedish, and German subtitled versions of an American film, Jaskanen (1999: 59, 62, 63, 64, 67, 68, 71) found numerous instances of explicitation in TTs. Jaskanen categorized these as part of a "global translation strategy" of "naturalisation" that appeared to have been followed by the translators.

Jaskanen's analysis was as follows. For audiovisual translation, she adapted Toury's continuum of initial norms, the two poles of which are adequacy and acceptability. Jaskanen divided this adapted continuum into three categories, "exoticisation" (corresponding to Toury's SL-oriented adequacy), "naturalisation" (corresponding to Toury's TL-oriented acceptability), and "neutralisation," the mid-point in Jaskanen's continuum (1999: 14; 43-44).

More extensive, statistically based studies became possible with the growing use of electronic corpora. In one of the first major corpus-based studies of

explicitation, Burnett (1999; cited in Olohan 2001 and in Olohan and Baker 2000) found significantly more use of optional, semantically void syntactic elements in translated texts, regardless of the degree of formality of the text style. Specifically, the use of optional *that* was found to be more common than the use of zero-*that* in the translated English texts as a whole, while zero-*that* was more common in the non-translated English texts as a whole (*ibid.*). For this study, Burnett (1999) used a corpus of fiction and biography translated into English from several languages; he reviewed the occurrence of optional *that* after selected forms of seven reporting verbs (*suggest, admit, claim, think, believe, hope*, and *know*).[36]

Similarly, Olohan and Baker (2000), who also counted and compared the use of optional *that* in translated and non-translated English texts (after only two reporting verbs, *say* and *tell*, but in a much larger version of the corpus originally used by Burnett), showed that optional *that* was significantly more frequent in the translated texts of their (mainly literary) corpus of narrative English.[37] In subsequent research, Maeve Olohan (2001: 425-427; 2002: 6) reached the same conclusions when she analyzed a still more enlarged version of the corpus used by Olohan and Baker (2000).[38] On the basis of this finding, she concluded that translated English texts were likely to favour the use of explicitated syntactic forms "even in contexts which do not warrant it, e.g. for purposes of disambiguation or for the signalling of more formal style" (2001: 423). Olohan (2002: 158-163)

---

[36] Burnett's study is reported in both Olohan (2001) and Olohan and Baker (2000). As we did not have access to his unpublished dissertation, we were not able to ascertain the exact size of his corpus at the time of his study. The corpus used by Burnett was a smaller version of a corpus later much enlarged, and subsequently studied by Olohan and Baker (2000) and by Olohan (2001 and 2002).

[37] The enlarged version of the corpus used by Olohan and Baker (2000: 151-152) contained approximately seven million words of fiction and biography.

[38] The version of the corpus used by Olohan (2001: 424; 2002: 6) totalled approximately 12.8 million words.

subsequently investigated five optional syntactic features, four more than in her 2001 study.[39] These features were:

1. Optional use of the *that* complementizer (used as a measure of explicitation in the 2001 study)
2. Optional use of four other relative pronouns (*who*, *which*, *whose*, *whom*)
3. Optional use of the complementizer *to* following *after having*
4. Optional use of *while* (in *while –ing*)
5. Optional use of *in order to.*

Olohan (2002) found all five optional syntactic features to be significantly more frequent in her translated corpus, lending further support to the hypothesis of explicitation as a recurrent feature of translation.[40]

In 2003, Birgitta Englund Dimitrova published partial findings of on-going cognitive research on explicitation, which showed that implicit contrasts in the ST tended to be explicitated in TT (2003: 26).[41] Recently, Tiina Puurtinen (2004) and Vilma Pápai (2004) have both found evidence in support of the hypothesis of explicitation in translated Finnish and Hungarian texts, adding valuable data on two non-Indo-European languages of translation to the growing list of studies.

---

[39] For this study, the corpus was divided evenly into translated fiction and biography texts and non-translated "imaginative writing" texts in English (Olohan 2002: 154-155).

[40] Olohan's method of investigating explicitation using semantically void optional words was duplicated in miniature by Silvia Hansen and Elke Teich (2001), who used a very small (20,000 word) electronic German comparable corpus to test the hypothesis of explicitation. They counted occurrences and omissions of the semantically void optional *daß* (that) in concordances containing frequently used German reporting verbs, and predicted that the translated corpus would show explicitation in its higher-frequency of use of the optional *daß*. Contrary to their expectations, however, they found fewer *daß*-clauses, and more zero-optional uses, in their translated German corpus. The extremely small size of their corpus makes it difficult to predict whether the same trends would be seen in a larger sample, however.

[41] Englund Dimitrova studied the Swedish translations of four professional translators, two student translators, and three language students with no translation training or experience. Her partial results report whether these subjects rendered explicit an implicit contrast between two Russian text segments with two sentences each, the second sentence of which "is in strong contrast to what the reader is expecting, having read and comprehended the first sentence" (2003: 24). We are extrapolating these figures from Englund Dimitrova's results. First, we have excluded Dimitrova's data on the translation behaviour of the language students, reasoning that neuropsychological research has demonstrated clearly that translation takes place in a separate part of the brain from speaking (see for instance Paradis 1984, 1985, 1995, 2000, 2001, and 2001b), and that bilingualism cannot be equated with the ability to translate. In Englund Dimitrova's study, the professional translators and student translators considered together chose to render the text segment contrast explicit twice as often as they left it implicit. The professional translators by themselves explicitated three times as often, suggesting that the tendency to explicitate in fact grows with experience.

Puurtinen found an overall tendency, in the translated texts of her million-word Finnish corpus, for "connectives" to appear more frequently; this finding supports the hypothesis of explicitation (2004: 169; 171-173). Pápai, who employed both manual and automated methods of perusal to find instances of "optional addition" of various words in her very small (45,000 words) Hungarian corpus, had the preliminary finding that in general, translated texts showed a "higher level of explicitness" than non-translated texts (2004: 156-157; 159).[42]

Most of the studies cited above found evidence indicating that explicitation is indeed a feature of translation. However, these studies investigated mainly literary translated texts, in a single language only. Where large corpora were involved, measurement of explicitation was limited to only one optionally-included word, phrase, or grammatical category. To the best of our knowledge there has as yet been no research that investigates explicitation of translated texts through a variety of measures, using exclusively non-literary corpora in more than one TL.

2.3 Simplification

Toury has claimed that translators consistently simplify the form of the translated text, and that this is one of the "most persistent, unbending norms in translation" in all languages (1991: 188). Baker calls this assumed tendency simplification, and includes it among her hypothesized universals of translation (1996: 181-183). She defines simplification as a hypothetical tendency to "simplify the language used in translation," and to make translated texts "easier" to read without making the text more explicit (*ibid.*). In what follows, we will review some

[42] Since Pápai used a number of measures that could only be automated if extensive annotations were put in place, such as "culture-specific items with added information" and "situational addition," she had to limit her corpus to a manually perusable size.

of the relevant theoretical statements that translation scholars have made concerning this feature. We will then review the available evidence from empirical testing of the hypothesis that simplification is a recurrent feature of translation.

2.3.1 *Theoretical background*

Simplification is listed among the "universals of translation" recognized by Laviosa-Braithwaite (and by Mona Baker, ed.) in the *Encyclopedia of Translation Studies* (1998: 288-289). A number of scholars (Baker, Berman, Bernardini and Zanettin, Blum-Kulka and Levenston, Chesterman, Hervey and Higgens, Laviosa[-Braithwaite], Mossop, Nida, Pápai, Toury, Vanderauwera, Vinay and Darbelnet, and Øverås) view simplification both as an intrinsic (and possibly unconscious) feature of the translation process, and as a conscious strategy for following Toury's initial **sociolinguistic norm*** of acceptability, by attempting to meet perceived expectations of the (imagined) readership.[43]

A number of scholars describe simplification as something that is inherent in the act of translating, regardless of the language pairs involved. Baker (1996: 176) says that, whatever the constraints imposed by differences between the two languages, "translators subconsciously simplify the language or message or both." Chesterman (2004: 40-45) lists simplification among the "T-universals," those recurrent features of translation that are among the "universal differences between translations and comparable non-translated texts, i.e. characteristics of the way translators use ... target language." Bernardini and Zanettin (2004: 59-60) view

[43] The readership is "imagined" in the same way that a national community is imagined. See Anderson (1991) and Ignatieff (1993). We are grateful to Dr. Timothy Stanley for pointing out this similarity.

simplification as having been proved to exist only if it is found to be both a distinguishing feature of the translated texts in a corpus that contains both translated and non-translated texts, and a feature traceable in a parallel corpus to strategies (whether conscious or unconscious) for dealing with language differences.

Some scholars describe simplification as a conscious translation strategy. According to Mossop (1983: 261), omission of "excess verbiage" is among the legitimate decisions made by the translator in the process of reporting what is said in the ST. If this omission simplifies the reader's task, it can be considered simplification. Baker (1992: 26-28) claims that translating a hyponym by a superordinate is "one of the commonest strategies" for dealing with "non-equivalence."[44] Such "strategies" of simplification can have the effect of reducing the information load that the reader must process, Laviosa maintains (2002: 43-44, 48, 60-61). Hervey and Higgens (1992: 250) seem to be describing a kind of simplification in their definition of "generalizing translation," in which "details that are explicitly present in the literal meaning of the ST" word are omitted, and a word that is "wider and less specific" in meaning tends to be used instead.

Along with Mossop (1983: 261), several scholars describe simplification as something akin to the obverse of explicitation: a kind of reduction or compacting of the translated text.[45] For instance, Eugene Nida (1964: 231-233) includes "subtraction" (e.g. reduction or deletion of repetitions and conjunctions) among his

[44] Baker posits "non-equivalence" as instances where "the target language has no direct equivalent for a word which occurs in the source text" (1992: 20). Blum-Kulka and Levenston (1983) refer to this as "lexical voids." Folkart (1991) refers to this as "cases vides."

[45] They do not seem to be suggesting, however, that the effect on the reader is opposite: while simplification makes a text easier for a reader to process, they do not claim that explicitation makes a text harder to read.

"techniques of adjustment." This change affects the structure, but not the meaning: surface features such as awkward constructions are simplified and therefore easier to read (*ibid.*).

Mildred Larson maintains that it may be necessary for the translator to reduce the amount of redundancy "for easier understanding," and that translators must make adjustments accordingly, since "redundancy patterns and functions" differ by language and text type (1998: 491). The information made implicit in the TT may thus be grammatical, semantic, or stylistic, Larson says (1998: 497-498), postulating simplification as a kind of reverse explicitation. Øverås (1998: 17) depicts the tendency to "explicitate and implicitate" in translation as two sides of the same behavioural coin, a shifting of translation toward the "acceptable" pole of Toury's continuum of initial norms of acceptability.[46] Pápai (2004: 160) states that the "lexical repetitions" involved in the process of explicitation lead to a "lower variety of vocabulary" and hence, she says, to lexical simplification, making it likely that these two recurrent features (simplification and explicitation) will appear together in a corpus of translated texts.[47]

It should also be acknowledged that some observable instances of simplification could be interpreted to be the result of translation error, in which systematic failure to fully understand a text's complete range of references (i.e. its morphology, syntax, and meaning) leads to systematic failure to restate it fully in the TL, or what Folkart (1991: 33-35) has called "opacification."

---

[46] Since explicitation is the only hypothesized recurrent feature included in her study, Øverås considers evidence of "implicitation" to be "norm-disconfirming" (1998: 5), rather than evidence of another, possibly co-existent feature of translation.

[47] Lower type/token ratios are brought about by such explicitations as the addition of conjunctions which "add up, indirectly, to the number of repeated items, i.e. [to a higher] number of tokens, and [which] therefore lessen the number of types," Pápai notes (2004: 160). As we will see below, Laviosa instead considers type/token ratio to be among the possible measures of simplification.

It is worth pausing here, before we conclude this review of the literature on theories of simplification, to note that simplification of text structure is not tantamount to increased text readability. Furthermore, although a number of scholars appear to equate implicitation with simplification (see above), the two terms cannot be considered synonymous. By condensing a text, "implicitation" may make reading harder, reducing redundancy and increasing the cognitive load that must be processed. Although replacing one "hard" word with an explanatory phrase also increases repetition and redundancy in the text, the phrase may nonetheless be easier for (an experienced) reader to understand. This makes it possible to consider the text to be simpler in terms of its readability. A text with a grammatically simple, compact structure may be very difficult to read, while a text with a more complex syntactic structure may be much easier to read. Consider the following examples:[48]

a. Inspect the forward strut rear angled needle roller bearing housing.
b. Inspect the housing of the rear-angled needle-roller bearing on the forward strut.

While example (a) has a shorter and simpler syntactic structure, example (b) is clearly easier for educated adults to read.[49] In addition, it must be stressed that simplification cannot be equated solely with reduction in the number of signifiers, any more than explicitation can be equated with a mere increase in the number of signifiers.[50]

---

[48] We are grateful to Dr. Lynne Bowker (personal communication) for providing these examples.
[49] Both sentences clearly require a high level of reading skill; neither would be suitable reading material for pupils in the lower grade levels.
[50] Berman, for instance, maintains that translation has twelve consistently "deforming tendencies" that represent a falling away from the ideal of perfect text pair equivalence (2000: 288). Among these tendencies, "quantitative impoverishment" (in which the translated text "contains fewer signifiers than the original") leads to "lexical loss" according to Berman (2000: 291-292). That is, there are fewer words in the translated text and the vocabulary is therefore impoverished. This might appear at first glance to fit the hypothesis of simplification, but the crucial element of comprehensibility is missing.

In the literature containing theories of simplification, we have noted a pervasive assumption that translators tend to use language more simply than writers do. The hypothesis that there will be observable evidence of simplification in most translated texts emerges from this idea. What concrete evidence of simplification may be found in a corpus of translated texts? A number of suggestions are made in the literature, all pertaining to vocabulary and grammar.

Blum-Kulka and Levenston (1983) maintain that translation strategies involving "overgeneralization" and the use of superordinate terms may have the effect of reducing the translated text's vocabulary range, which can then be considered lexically simplified (1983: 126-127). Toury (1980: 104) notes that translated Hebrew titles of novels seem to have a "limited range of syntactical patterns" that might, he believes, be part of a general "tendency toward simplicity." Vinay and Darbelnet (1995: 341-343) associate the comparative quantity of content versus grammatical-function words with processes and strategies that fit the hypothesis of simplification.

Baker (1996: 183) predicts that **lexical density ratios*** and type/token ratios will show translated texts to be comparatively simplified. Hansen and Teich (2001: 2, 3, 6) also hypothesize that overall lexical density ratios will be lower in translated texts. As we will see below, the empirical research on simplification has investigated aspects of both vocabulary and grammar in translated texts, focussing mainly on comparing word frequencies and sentence lengths.

2.3.2 *Empirical research*

For her doctoral research on English translations of Arabic novels, Najah Shamaa (1978: 168-171; cited in Baker 1995: 228) concluded that manually-counted "common" words such as *day* and *say* occurred with a higher frequency in the TTs than in the STs. Her findings suggested that the vocabulary of translations will be simplified compared to the vocabulary of non-translated texts, opening the way for future investigation of simplification using measures, such as lexical density ratio and type/token ratio, that are based on word frequency distribution. Vanderauwera (1985: 98-99) found that elaborate phrases in the ST were often replaced by shorter phrases in the TT. She also noted that long strings of paratactic clauses in the ST ("Jerky juxtaposition of syntactically identical and independent clauses" Vanderauwera 1985: 101) were often condensed into shorter (but syntactically more complex, hypotactic) sentences in the TT (Vanderauwera 1985: 101-102; 104-105; 110-111; 115). Shlesinger (1989: 150-151) found that court interpreters were likely to reduce, condense or omit "salient features of density in literate discourse" (such as pre- and post-posed modifiers, nominalizations, series, and sequences of prepositional phrases). Repetitions of phatic expressions (such as "if it please the court") were often deleted as a matter of course (1991: 151-153). Likewise, inappropriate or insulting comments (such as "the guy with the bald head") and direct references to the interpreter (such as compliments made about themselves in their capacity as interpreter) were often deliberately omitted (*ibid.*).[51]

[51] These may be taken as examples of what Kenny (1999b) calls "sanitization."

In one of the first major corpus-based studies to include empirical investigation of simplification as a recurrent feature of translation, Sara Laviosa-Braithwaite (1996) did doctoral research on a very large corpus of English texts that included translations into English from a variety of languages.[52]

Laviosa-Braithwaite (1996) identified four regular features of her translated corpus that she believed to be "core patterns" of simplification: (1) lower lexical density ratios; (2) a higher proportion of high-frequency words; (3) a larger percentage of the total vocabulary in the **list head***; and (4) fewer lemmas in the TT list heads (1996: 147). Laviosa-Braithwaite found, in her translated corpus, that there was likely to be a generally higher proportion of grammatical-function words versus content words in the text (i.e. something similar to a higher lexical density), along with a "relatively higher proportion of frequent versus less frequent words," a "relatively greater repetition of the most frequent words," plus "less variety of the words most frequently used" (1996: 157). This confirmed Laviosa-Braithwaite's specific hypothesis that translated texts would have both comparatively simplified vocabularies, and lower information loads (1996: 116-118).

Federico Zanettin (2000) measured the vocabulary range of a large (at 1.5 million words), literary, English-Italian parallel corpus, comparing each to a reference corpus.[53] Zanettin found that "above and below a certain threshold," there

---

[52]Note that the general objective of Laviosa-Braithwaite's (1996) doctoral research was to demonstrate the usefulness of corpus-based study of translation. Laviosa (2002: 58-63) summarizes the methodology and results of Laviosa-Braithwaite's unpublished doctoral research.The source-text languages included in Laviosa-Braithwaite's corpus were Germanic (Danish, Dutch, German, Norwegian, Swedish), Romance (French, Italian, Spanish, Portuguese), Slavic (Czech, Polish, Russian, Serbo-Croat), Modern Greek, Arabic, and Finnish (Laviosa-Braithwaite 1996: 88). Her translated English corpus (TEC) was over a million words (1,074,736 words), and her non-translated English corpus (Non-TEC) was nearly a million words (817,813 words). She designed each corpus so that it divided into similar proportions of comparable types (fiction and non-fiction): TEC 13.68% non-fiction and 86.32% fiction; Non-TEC 26.49% non-fiction and 73.51% fiction (Laviosa-Braithwaite 1996: 87). Note that the majority of texts in her corpora were literary, and that since the non-fiction texts were biography and reportage, all of the texts included in her corpora were narrative.

[53] Zanettin's parallel literary corpus consisted of five complete novels and one short story by Salman Rushdie and their published Italian translations (2000: 109).

was no significant variation in vocabulary ranges (as measured by standardized type/token ratios) between STs and their translations (2000: 113). Within that threshold, however, he found that translations consistently proved to be "lexically less varied" than non-translated texts (*ibid.*).

Diva Cardoso de Camargo (2003) recently published one of relatively few corpus-based studies devoted exclusively to non-literary texts.[54] She found some evidence of simplification in her corpus of English to Brazilian-Portuguese translations of technical and journalistic texts. In her corpus, the TTs had slightly lower type/token ratios, indicating that their range of vocabulary was slightly lower.

The above corpus-based studies found evidence in support of the simplification hypothesis. However, most of the corpora studied were literary, and most were limited to one language of translation per study. It remains to be seen whether simplification will recur in an exclusively non-literary corpus that includes more than one language of translation.

2.4 Levelling-out

The literature reviewed so far has investigated hypothetical universal characteristics of translation using measures that have depended on the particularities of the vocabulary and grammar of each language of translation studied. These measures have offered us some evidence that translated texts generally follow target-language norms more closely, and that they often have more explicit syntax, and simpler vocabulary and sentence structure, compared to

[54] Along with, for instance, Mauranen (2000).

non-translated texts in the same language. An additional recurrent feature of translation, considered to be a result of "the equalizing effect" of translation by Shlesinger (1989: 96-97; 170-171), and called "levelling-out" by Baker (1996: 184), is hypothesized in the present study. Although this feature has been noted in the literature (as part of the unexpected outcome of studies of other recurrent features), as far as we know, only one previous study (Shlesinger 1989) has included the specific testing of this hypothesis in any form, and only a few theorists have commented on it. Furthermore, to the best of our knowledge, extensive and systematic corpus-based study of a hypothesis of levelling-out as a recurrent feature of translation has not as yet been undertaken. In what follows, we will review the theoretical literature that is available, and will discuss the methodological options for future testing.

2.4.1 Theoretical background

Miriam Shlesinger (1989) was to the best of our knowledge the first in the literature to propose this recurrent feature of translation.[55] In her master's research on simultaneous interpreting, Shlesinger (1989: 96-97; partially cited in Baker 1996: 184-185) hypothesized that simultaneous interpretation "exerts an equalizing effect on the position of a text on the oral-literate continuum" and that as a corollary, "the range of the ... continuum is reduced." Shlesinger found the predicted "equalizing effect" was indeed present, and that the range of the

[55] Baker (1996: 177 *ff*) attributes both the concept and the term "levelling-out" to Shlesinger (1989). In fact, the term originally used by Shlesinger was "equalizing effect" (personal communication). In her Master's thesis, Shlesinger states that "simultaneous interpretation exerts an *equalizing* effect on the position of a text on the oral-literate continuum, i.e. it diminishes the orality of markedly oral texts and the literateness of markedly literate ones. ... The range of the oral-literate continuum is reduced in simultaneous interpretation (Shlesinger 1989: 96)." Shlesinger agrees that the concept of levelling-out applies to both written translation and oral interpretation (personal communication).

continuum was accordingly narrowed in the translated corpus (Shlesinger 1989: 170-171). As Shlesinger sees it, what distinguishes translation from non-translation is not the continuum, but the range within it. Translation will have different ranges within any continuum that is pre-established, generally recognized, and in general use.[56]

Baker (1996: 177, 184-185) proposes the specific hypothesis that "levelling-out" will be a recurrent feature of translation, and that translated texts in any language will "gravitate towards the centre" of a given continuum.[57] In Baker's study, the concept of the continuum is different from Shlesinger's, in one important respect. Baker's continuum can be internally-generated by the corpus studied; instead of being independently pre-established, it may be ad hoc. Translated texts may "steer a middle course between any two extremes, converging towards the centre, *with the notions of centre and periphery being defined from within the translation corpus itself*" (Baker 1996: 184; emphasis added). Baker hypothesizes (1996: 176-177) that the degree of similarity among individual translated texts will be measurably greater than that seen among individual non-translated texts, which will generate statistics that are less homogeneous and more dispersed. Baker cited evidence from Laviosa-Braithwaite's then-ongoing research that "the individual texts in an English translation corpus are *more like each other* in terms of features such as lexical density, type-token ratio...and mean sentence

[56] For instance, the oral-literate continuum, which Shlesinger (1989) uses in her study, is widely accepted: speech acts can be oral, written, or somewhere in between, with markers of orality in writing, or with prosodic style markers in speech. See Bikerts (1994), Crystal (1987: 177-181), Halliday (1994: 3-7; 349-352), Martin (1989), McLuhan (1962), Ong (1982), Tannen (1982) Tannen, ed. (1982) and Tannen, ed. (1984).

[57] Examples of an independently-established continuum would be the range from oral to literate used in Shlesinger (1989), the range of initial norms from "adequacy" to "acceptability," or the continuum of "readability," within which texts range from "very easy" to "very difficult" to read (see Section 2.4.2 and Chapter 5).

length than the individual texts in a comparable corpus of original English" (*ibid.*; emphasis added).

The evidence cited by Baker (1996; see above) was part of Laviosa-Braithwaite's doctoral research findings. Laviosa-Braithwaite (1996) found that measures she was using to test the simplification hypothesis on her corpora (i.e. lexical density ratio, type/token ratio, and mean sentence length) were generating statistics whose **dispersion*** (**standard deviation*** of individual values above and below the **mean***) was less than that of the same statistics for comparable non-translated corpora. These calculations showed that her translated corpora were more homogeneous (Laviosa-Braithwaite 1996: 126; 130; 134-135).

Baker (1996: 184) argued that Laviosa-Braithwaite's findings were evidence in support of the hypothesis of levelling-out. Laviosa-Braithwaite (1996: 135-136) expressed her agreement on the basic question of the existence of this recurrent feature of translation. However, she disagreed with Baker's name for it (*ibid.*), given the latter's concept of continuum. Laviosa-Braithwaite argued that the centre and periphery of the continuum cannot be internally generated and self-defined. Since each set of statistical highs, lows, and middle values is defined from within each corpus, scores can only be compared against an independently defined continuum (i.e., a pre-established set of references).

Shlesinger posits an external, pre-established set of criteria for a continuum, whereas Baker posits an internally generated continuum. Laviosa-Braithwaite gives different names to each: when the continuum has been independently defined and is

therefore pre-established, she calls the feature "levelling-out"; when the continuum is internally generated, she calls the feature "convergence" (1996: 134-135).[58]

Several other scholars subsequently entered into this theoretical discussion. Margherita Ulrych (1999: 41) used Baker's term, "levelling-out," but defined it somewhat differently. Ulrych posited levelling-out as "translational patterns" that occur with an *absolute frequency* that is significantly lower or higher in translated texts compared to non-translated texts. Ulrych thus did not define levelling-out in terms of relative ranges of highs and lows within a continuum. Hansen and Teich (2001), who listed levelling-out along with the three other recurrent features of translation discussed above, cited Baker's definition of the former. However, they investigated only normalization, simplification, and explicitation, and left levelling-out untested. Recently Mauranen (2004: 77-78) has noted that translations are more similar to one another than to comparable non-translated texts in the same language: they "show an affinity" for one another.[59]

Shlesinger's (1989) concept of the "equalizing effect of translation" is one upon which we can build in the present study, since it postulates the probability that translated texts will have a narrower range of extremes—of highs and lows—within a set of scores that are placed in an array along an independently-established continuum. Shlesinger's previous application of the (widely recognized) "oral-

---

[58] Laviosa-Braithwaite writes, "I will adopt 'convergence' to refer to the relatively higher level of homogeneity of translated texts with regard to their own scores on the measures of simplification selected in this study." Levelling-out, on the other hand, "will be used to refer to the shifts that take place along the oral-literate or any similarly pre-defined continuum in either interpreting or translating" (Laviosa-Braithwaite 1996: 136). Note that Laviosa-Braithwaite did not go on to use either of these two terms in the "Universals of Translation" entry she wrote for the 1998 *Encyclopedia of Translation Studies*. Instead, she coined a third, far more general term: "distinctive distribution of target-language items," a type of statistical levelling-out which is apparently limited to scores related to vocabulary (Baker, ed. 1998: 291).

[59] One example may be the small literary corpus in *TTR 12*:2. Folkart (forthcoming: 392) notes that these translations of Auden into French tend to have similar prosody and diction (e.g. "poetic' inversions"), and that this similiarity contradicts the assumption that there will be greater divergence among multiple renderings of a single poetic text than would be found in multiple translations of a single technical text.

literate continuum" to test the "equalizing effect of translation" (or "levelling-out," as Baker 1996 calls it) allows us to predict that within any pre-established continuum, a corpus of translated texts will score closer to the middle of the continuum and further away from the high and low extremes, compared to a corpus of non-translated texts.

Baker's hypothesis furthermore allows us to consider the internally-generated **median*** of these scores, and the range of scores that produces it, as another standard against which to test the hypothesis of levelling-out. Specifically, the dispersion (standard deviation) of readability scores can be used to gauge the comparative homogeneity of the present translated and non-translated corpora (see Chapter 5). This satisfies Shlesinger's (1989) original criteria, while simultaneously building upon Baker's (1996) hypothesis and updating and enlarging the scope of our hypothesized recurrent feature, which we will henceforth refer to as "levelling-out."

2.4.2 Methodological options

Although she did not set out to test it as a recurrent feature of translation, Laviosa-Braithwaite (1996: 126, 130, 132,134,135-136,139, 142, 144, 146) did contribute the measure that Baker (1996:184-185) cites in her proposal of levelling-out as a recurrent feature of translation. In her doctoral dissertation and related publications, Laviosa-Braithwaite (1996) calculated the variance for the scores obtained from her measures of simplification.[60] She found that translated texts in her smaller non-fiction corpus generally had scores with lower variance, while

[60] Lexical density, type/token/*100* ratio, and mean sentence length. Laviosa-Braithwaite also calculated the variance for the "proportion of high frequency words" in her corpora (1996: 143-146).

translated texts in her larger literary corpus had scores with slightly higher variance for three of the measures used: lexical density ratio, "proportion of high frequency words," and type/token/*100* ratios (Laviosa-Braithwaite 1996: 144). She also found that her literary corpus was "fairly homogeneous" (1996: 146), having low variance for all of the measures used except mean sentence length.[61]

We were able to find no other methodological precedent for testing the hypothesis of levelling-out as a recurrent feature of translation. Our study is, to the best of our knowledge, the first to explicitly design and carry out tests of levelling-out as a recurrent feature of translation.

If we are to test this hypothesis, we will need to do so using a pre-established statistical continuum that is based on generally-recognized, surface-observable linguistic features. Our continuum must, in other words, be one that has gained general acceptance and that is a valid and reliable method of testing.

We believe that readability indices are a good starting point for our investigation of levelling-out. Readability indices have been in use for decades, are validated for various languages, and operate along a recognized pre-established continuum. Although we will cover the practical literature on readability indices used in the present study in Section 3.3.2 below, we will present here a brief historical overview of the readability index.

Empirical methods of evaluating the reading difficulty of instructional materials were developed as early as the late nineteenth century (Venezky 1984: 24). Their main focus then, as now, was on vocabulary (word type and word

[61] The mean sentence lengths of the translated fiction texts had an anomalously high (91% greater) variance than those of non-translated English fiction texts (*ibid.*). Laviosa believed that mean sentence length was affected by the source language (129, 138, 146).

length) and sentence length (*ibid.*) as indicators of the level of reading difficulty of a text.[62] In general, a readability index uses either or both of these text features in a formula that provides a score which evaluates the general level of reading difficulty of a text or texts. Today, readability indices are in common use among technical writers (Cherry [1982]; Gunning [1968 and 1964]; Kincaid *et al.* [1975]; Macdonald *et al.* [1982]; Pigeon [2002]; Roberts *et al.* [1994]; Talburt [1986]; Tremblay [2000]).

A number of readability formulas were selected for use in the present study, based on two criteria: (a) their widespread acceptance among educators and technical writers; (b) their applicability to the present corpora. The latter criterion is particularly important. The readability indices selected must be so constructed as to allow their use on a large number of words of text.[63] Furthermore, they must produce a set of scores along a single statistically-based continuum of "readability," that is, of reading difficulty in relation to a large population of readers. In the present study, the purpose of using readability indices is not to measure readability *per se*, but rather to obtain a set of scores with which to measure the hypothesized recurrent feature called levelling-out. Below, in Section 3.3.2, we present the individual formulas selected on the basis of the above two criteria.

---

[62] The continued focus on vocabulary is seen in the fact that "robust predictors of word-recognition performance" in various populations are reckoned to be among the latest technological advances that make it possible for researchers in education, psycholinguistics, and cognitive psychology to predict text readability (Lété *et al.* 2004: 156-157). These "predictors" are based on word-frequency counts in very large corpora, but the preoccupation with vocabulary that motivated their development is probably centuries old. See Lorge (1944).

[63] For this reason, readability measurements based on "holistic" teacher estimates (e.g. Russell 1993), or that require reader participation (e.g. reader recall, reading speed, or eye movement), cannot be used for the purposes of the present study.

2.5 Summary

In the theoretical literature and empirical studies surveyed, there is evidence that recurrent features exist which consistently distinguish translated texts from non-translated texts, and which consistently recur, regardless of the influence of SLs. Most recurrent features of translation can only be discerned through specific surface characteristics of the TL. Levelling out, however, is a hypothetical recurrent feature that can theoretically be observed in translated texts by comparing any statistics they generate to any pre-established statistical continuum (Baker 1996: 184).

2.6 Conclusion

The above-cited studies provide some evidence that the wording and syntax of translated texts tend to be more explicit, simplified, and normalized than that of non-translated texts. They also suggest that translated texts will tend to score closer to their statistical median within a given continuum. However, all of the studies surveyed searched for only one recurrent feature at a time, in one language of translation, using corpora that were mainly literary. In the present study, we propose to test all four hypotheses on non-literary corpora in two TLs, French and English.

3. Corpora and Method of Analysis

Our methodology is based on the use of electronic corpora. A corpus can be any collection of spoken or written texts, but is more often conceived as a sample of authentic texts gathered in electronic format and used as a qualitatively representative reference for linguistic research (McEnery and Wilson 1996: 21-24; Bowker 2002: 43-46; Bowker and Pearson 2002: 9-10, 20).[64] An overview of the history and theory of corpus use in Translation Studies research is provided in Laviosa (2002: 5-31) and therefore will not be discussed further here.[65]

It should be noted that the present study is "corpus-based" rather than "corpus-driven," to use Tognini Bonelli's terms. Tognini Bonelli (2001; 2000: 207; 1996) distinguishes between a purely inductive, "corpus-driven" exploration and discovery of patterns in corpus data, versus the more deductive approach of formulating a hypothesis and then modifying it based on tests on a corpus. The latter is the "corpus-based" method of the present study.[66]

---

[64] Note that this is not the same thing as a strictly randomized, quantitatively representative sample.

[65] We do not mean to imply that corpus-based study of translation constitutes a methodological paradigm shift; it is in fact built on earlier work by Descriptive Translation Studies scholars using corpora that they were obliged to gather and analyze, painstakingly, by hand. The work of these forerunners needs to be acknowledged, as the present study would not be possible without them. See for instance the analyses of Lefevere and Jackson (1982), Hermans (1985), Vanderauwera (1985), Blum-Kulka (1986), Schlesinger (1989), and Even-Zohar (1990). For linguistics-oriented manual corpus analysis see for instance Du Nour (1995), and Weissbrod (1992, 1992b, 1990). For culturally-oriented research, see for instance Folkart (1990, 1991) and Brisset (1996).

[66] For more on the distinction between "corpus-based" and "corpus-driven" methodology, see also Hunston and Francis (2000: 18-19).

3.1 Design criteria

Corpus design criteria depend on the envisaged use of a corpus in a given study. Since our general hypothesis is that a translated text is distinct from other types of writing and that it therefore has distinct characteristics, we have chosen to use univariate **comparable corpora***, i.e. sets of translated and non-translated texts in the same language. We have opted to study translation into two languages instead of one, reasoning that where the identification and description of the specific characteristics of translation are concerned, it is important to study as many languages as possible (see for instance Mauranen and Kujamäki 2004: 3). In the present research, we therefore compare four electronic corpora of texts, in two in English (translated and non-translated) and two in French (translated and non-translated).

To designate these corpora, acronyms commonly used in Translation Studies for "Source Text" and "Target Text" (ST and TT) are combined with the first letter of the language as follows:

1. English corpora: **EST*** (non-translated); **ETT*** (translated)
2. French corpora: **FST*** (non-translated); **FTT*** (translated)

In other words, we are using four independent corpora designed to form two sets of comparable corpora, one in English and one in French, as a means of isolating one variable (i.e. a text's being a translation or not).

The size of a corpus can vary tremendously, depending on the purpose of the research. For instance, with lexicographic research, where the corpus is used to

study the meaning and to track the usage of a large number of words, a corpus may contain hundreds of millions of words.

However, where the goal is to explore text qualities, such as features that may distinguish translations from other forms of writing, researchers are often obliged to use corpora that are considerably smaller. Where the hypotheses are precisely defined and the tests intended, as the present one is, to be univariate, unavoidable practical restrictions often apply.

In the present study, since we have distinguished texts by the single variable of "translated vs. non-translated," we are obliged to depend on our Government of Canada sources to identify which of the text pairs on their official Web sites are translations.[67] We also had to gather our corpus from scratch, given the lack of a large "Canadian National Corpus" on the scale of the British National Corpus, from which the corpora used in the previous British studies of recurrent features of translation were drawn. Our Government of Canada sources were able to confirm only around 125,000 words of text on their Web sites to be translations. This unavoidably restricted the size of our corpus to 250,000 words (i.e. 125,000 words of translated text, and 125,000 words of non-translated texts, evenly divided up into each of the two languages of Canada). Presented in Table 1 is the number of words in each of our overall corpora, which contain both specialized and non-specialized texts.[68]

---

[67] An alternative procedure would have been to guess. While this method would have allowed us to gather millions of words, guessing has the fatal flaw of being entirely subjective. Since the whole point of the present study is to set aside prejudices about the nature of translation in favour of empirical observation, we opted for the more rigorous method of obtaining confirmation of each text's status. This restricted the size of our corpus, but allowed us to be certain that what we were studying were in fact translated texts.

[68] Note that the counts shown in Table 1 were obtained with WordSmith 3 (2002), using WordSmith's default settings for each language. In 2004, after all of the research for this study had been completed, WordSmith 4 was released. This version of the software gave different word counts for the same corpora, as follows: EST 64,331; ETT 62,919; FST 62772; FTT 59246. At 250 words per page, this altered the page counts as follows: EST 257; ETT 252; FST 251; FTT 237. All of the

Table 1
Overall corpora used in the study[69]

	Number of Words	Number of Texts	Number of Pages (@ 250 words/page)
EST	61,699	63	247
ETT	60,435	53	242
FST	60,725	47	243
FTT	60,798	33	243
Total English	122,134	116	489
Total French	121, 523	80	486
Total All	243, 657	196	975

As mentioned above, the chief objective of the present study is to compare translated written texts with non-translated written texts, looking for features that may be considered (qualitatively) to distinguish translated texts. It follows that for the purposes of the present study, all the texts selected are written texts (and not transcriptions of oral texts). Moreover, each text has been confirmed, by its source, to be either a translation or an "original." This is a key corpus design criterion. Having texts in at least two languages is also essential, as is the restriction of the corpora to non-literary texts (whether specialized or not).

Texts meeting all of these criteria were taken from a single very large source: Government of Canada Web sites. With such texts, identification of individual translators and authors is not possible. The authorship of documents posted to these sites is rarely identified and can arguably be designated as "institutional," since most of these documents are in all probability the work of

measures were subsequently double-checked using WordSmith 4. Except in the few cases noted, there were no substantial changes to the results reported in Chapters 4 and 5.

[69] In all Tables, percentages are rounded up from the second decimal.

more than one person.[70] In their anonymous authorship, the non-literary texts included in the present corpora can be assumed to have been produced in the manner of what is called "factory translation" in Milton (2000: 5-6, 11).[71] Among the characteristics of "factory translation" noted by Milton are the fact that it is produced anonymously, through teamwork (2000: 5). [72]

Within the large single source of the Government of Canada, a number of institutional "authors" (government departments and agencies) can be identified. A list of each department or agency that contributed a text or texts is given in Appendix I, along with the acronym used to designate the source for the purposes of identifying each text in a corpus. The total number of texts and the total number of words contributed by each department or agency are also given in Appendix I.

Since features distinguishing translation from non-translation can appear in any part of a text, we did not take extracts and included full texts only, following the recommendation of Bowker and Pearson (2002: 49) for corpus design tailored to the needs of studies such as the present one. However, this meant that we would be working with widely varying text lengths, as can be seen in Table 2.[73]

---

[70] This was confirmed in personal communications from most of the Government of Canada departments and agencies that contributed texts.

[71] Lambert (1994: *passim*) also discusses modern non-literary translation's "industrial element."

[72] Other characteristics of "factory translation" noted by Milton are that it is often tailored to fit particular markets or target readerships, that the texts's theme, style, and length will be subject to standardization or "fordism," and that the push to meet deadlines will supersede quality standards. This is as opposed to what Milton calls "aristocratic translation," the tradition of "crafted literary translation, often carried out within the university" (2000: 5).

[73] The word counts shown in Table 2 were obtained with WordSmith 4 (2004) using the set of individual texts comprising each corpus, and using the default text processing setting for each language.

Table 2
Shortest and longest texts in the overall corpora

	Shortest text	Longest text
EST	115 (CBEST1)	8482 (NREST4)
ETT	94 (CTETT8)	4455 (PWETT2)
FST	91 (DTFST15)	4920 (CTFST8)
FTT	134 (CBFTT1)	29,315 (NRFTT4)

Representatives from each contributing department or agency stated, in personal communications, that all of the documents about which we made inquiries had been posted to the Web sites sometime between 1995 and 2002, the year we began the process of gathering the corpora used in the present study. Although it is a product of circumstance rather than of choice, the seven-year period covered by the corpora seems a fitting length for our synchronous study.

To summarize in Laviosa's terms (1997: 291-295), for the purposes of the present study we have designed a full-text, synchronic, mixed-terminological (both general and specialized) written corpus with translations and non-translations in two languages. Following Laviosa's recommendations (1997: 291-295), the texts that comprise our "translation-dependent" comparable corpora have been selected from the point of view of the translator, that is, on the basis of the composition of the "translational" set, with the texts comprising the non-translated component being selected to match the translated component, and not the other way around. Each of the translated corpora is "mono-source" (i.e. each has a single SL). Finally, since all the texts have been published, providing some guarantee of quality in writing and translation, we assume that the texts included in our corpora

are the work of professionals.[74] Presented in Table 3 is a summary of the attributes of our corpora.

Table 3
Summary of overall corpus attributes

Status	Translated or Non-translated
Languages	English or French
Text type	Non-literary: general or specialized
Authorship	Institutional
Size	Approximately 250,000 words
Text coverage	Full texts (no excerpts)
Texts per corpus (mean #)	49
Medium	Writing
Subjects	All available
Publication dates	1995-2002

3.1.1 Design of the Non-specialized Sub-corpora

We used readability indices to measure levelling-out (see Chapter 5). Most readability indices were designed for manual application on single texts samples of around 100 words each, taken from the beginning, middle, and end of the text, and adding up to about 300 words per text tested for readability. We therefore considered it necessary to work with readability indices using smaller subsets of the above corpora.

Furthermore, a number of the readability indices selected for use in the present study require highly accurate syllable counts. These counts could not be made using the corpus processing software that was available for the purposes of the present research.[75] The syllable counts were therefore performed manually. We felt justified in using smaller sub-corpora, since it would take a single researcher a very long time to do a manual count of all the syllables in all 250,000 words of the

[74] Government of Canada policies and directives generally stipulate that translations published under the aegis of the government be the work of accredited professionals (see for instance Mossop 1990). Laviosa (1997: 296) makes the same assumption about her own corpus.

[75] Some of the indices (such as the Henry-de Landsheere: see Sections 2.4.2 and 5.1.5) require several manual counts to be performed on text samples that are kept to a strictly limited size. These text excerpts were taken from the non-specialized sub-corpora, to keep the application of these indices as similar as possible to the application of the other indices.

present corpus, and since unacceptably high rates of human error would have been likely if a manual count had been attempted on such a scale.

Wherever possible, we wanted to run several readability index formulas on the same set of texts. It was also imperative that the readability index scores be calculated on non-specialized texts, since specialized texts will automatically score among the hardest to read (e.g. Hunt 1977; Tremblay 2000), potentially skewing our results. Furthermore, it was desirable that we isolate the non-specialized texts, because such texts are least likely to be marked for **register*** (Quirk *et al.* 1985: 24-27). In the investigation of explicitation, this could be useful: the sub-set of non-specialized texts could serve as a kind of control sample, allowing us to gain a tentative idea of whether register had influenced the overall findings for explicitation (see Section 4.2.2).

We therefore created four sub-corpora of non-specialized texts, by removing all the specialized texts from each of the four overall corpora (of specialized and non-specialized texts) used in the present study.[76] The resulting sub-corpora consist of texts that we believe would probably be deemed, from the point of view of the majority of professionally-trained, experienced translators, to lack sufficient specialized terminology to warrant extra research time.

Presented in Table 4 are the word counts and number of texts for each of the non-specialized sub-corpora.

[76] We also removed a few non-specialized texts from each of the overall corpora, to even up the final word counts of the non-specialized sub-corpora (Table 4).

Table 4
Non-specialized sub-corpora

	# words	# texts
Non-specialized EST	15,954	25
Non-specialized ETT	15,939	31
Non-specialized FST	15,947	29
Non-specialized FTT	15,977	25
Total English	31,893	56
Total French	31,924	54
Total All	63,817	110

Since their word counts were nearly equal (the variation being being within 100 words), the non-specialized sub-corpora can be considered comparable in length for the purposes of the present study.[77]

3.2 Corpus compilation

We gathered all of the texts that were available for a corpus of this design. As indicated above, our corpus was gathered from Government of Canada Web sites. All texts on these sites are published in both English and French, in accordance with the Official Languages Act (Government of Canada 1985). In practice, this means that many (up to 50%) of the texts posted are translations. The texts are whole (un-excerpted), and they cover a wide range of non-literary topics. They are in the public domain, which means they may be copied electronically without contravening copyright law. This allows the researcher to amass texts relatively quickly. This makes Government of Canada Web sites a seemingly ideal source of material for the present study: large numbers of translated and non-translated texts can theoretically be gathered in both French and English.

Unfortunately, Government of Canada Web pages rarely indicate whether the text displayed is a translation. This leaves two options: attempting to discern the

[77] As can be seen in the final word counts (Table 4), there were approximately 40,400 more words of "specialized" text in each of the overall corpora. This is not surprising: according to Bowker (2003: 173-174), most non-literary translation today deals with specialized subject matter.

translation for oneself, or getting confirmation from a reliable source. To gather a corpus designed to rigorously differentiate between characteristics of translation, one cannot merely guess which of the texts on a Web site is the translation; a corpus gathered in such a way would reflect the gatherer's prejudices, rather than the actual characteristics of authentic material. To meet our corpus design criteria without skewing our results, it was clearly necessary to contact each government department and request verification of each text's status (as a translated or non-translated document). Many of our requests for this information received no response, because the Government of Canada departments were not sure which of their Web pages were in fact translations.[78] We were therefore obliged to work with those texts whose status was confirmed. We compiled all of these into an archive.[79] Since there were a limited number of confirmations, the necessity of following this procedure in order to ensure proper corpus design ultimately limited the size of our corpus.[80]

Since the texts included in the corpora were all downloaded from the World Wide Web, there were a number of technical concerns that needed to be addressed. We wished to save the texts minus their HyperText Markup Language (HTML) tags, which could potentially hinder some of the processing carried out by our corpus analysis tools. We therefore saved each text in plain text format (i.e. with the extension ".txt"), and proceeded to manually remove the unwanted line breaks

---

[78] The status of the texts posted on some Government of Canada Web sites had not been recorded and could not be retrieved; it was therefore impossible for many of the representatives of the departments or agencies to provide confirmation of the status of the texts on their Web sites.

[79] See Bowker (2003b: 170) on the distinction between an archive and a corpus. The Canadian government Web page texts that were confirmed to be translations were all copied and stored along with their source texts. Note that the comparable corpora used in the present study were selected from this archive of parallel texts.

[80] However, following this procedure ensured that the work of a large number of anonymous writers, translators, and revisers was qualitatively sampled; as such it followed the recommendations of Bowker and Pearson (2002: 51, 54). Furthermore, there was, as recommended, "wide linguistic coverage."

inserted by the word processor, as recommended by Bowker and Pearson (2002: 65-66).

While this process ensured that we saved all text appearing on the Web page, it also meant that certain other types of information were not included in the corpora gathered: among the data not retained were sound files, video files, and in particular the graphics (which in some cases included text, although we made a concerted effort to manually transcribe as much of the text included in graphics as possible). We believe, however, that such audiovisual information should not be considered essential to the purposes of the present study, which focusses on the written word. Furthermore, we believe that video and sound files are in most cases probably not an integral part of the texts. Audiovisual files accompanying a given Web site may be considered unlikely to carry much in the way of useful linguistic information about its texts.

Representatives of the Government of Canada departments and agencies that contributed texts informed us in personal communications (by email and telephone) that almost all of the documents posted to their Government of Canada Web sites had not at that time (i.e. in 2002) been written specifically for the Web, but had in most cases been commissioned for earlier print publication, and had later been selected for Web posting.[81]

The attributes of each document gathered were recorded by a header adding 11 annotations to its text file, in keeping with Bowker and Pearson (2002: 81-84) and Bowker (2003: 170-171). To prevent the corpus analysis software from including the words in the annotations, the recorded attributes were sandwiched

[81] A small number of Canada Post texts created for commercial purposes were the sole exception.

between opening and closing SGML tags (<*/attribute*\>).[82] The annotations were as follows: text designation, number of words, URL, text status, type, date, source (department or agency), translation-related policies and directives of the source (if divulged), the type of translation service used (if divulged), the text's general topic or genre, and the text's title. The attributes included in the annotations are further explained below.

1. Text designation, e.g. </ACEST1\>

This designation consists of the source department's acronym (here "AC," which stands for "Agriculture Canada"), the corpus designation (in this case, the English non-translated corpus, or EST), and a number representing the order in which the text was gathered.

2. Number of words in the text, e.g. </Words 6993\>

This designation shows how many words are contained in the text proper (not including the words in the annotation).

3. URL

The address of the Web site from which the text was copied, e.g. </http://www.rural.gc.ca/conference/documents/mapping_e.phtml\>

4. Status, e.g. </Confirmed English Source Text\>

All of the texts copied were either translations or source texts for translations. A text gathered into a non-translated corpus was therefore designated as a "source

[82] Note that it is possible to include annotations in a search, by changing the settings of the corpus processing software. This allows the researcher to extract all texts with a certain designated attribute. For instance, from the corpora used in the present study, texts posted during a given year can be quickly assembled by searching for dates within the SGML headers enclosing recorded attributes.

text" (ST), while a text gathered into a translated corpus was designated as a "target text" (TT).

5. Text type (specialized or non-specialized), e.g. </non-specialized\>

Texts were designated as "non-specialized" from the point of view of the translator: texts were considered to be non-specialized when they did not appear to contain enough specialized terminology to warrant significant terminological research by a professional translator.

6. Date posted or modified, e.g. </Date Modified: 2002 11 06\>

The date that a Web page is posted or modified usually appears at the bottom of that Web page. Wherever possible, we noted the date given. The Government of Canada department or agency that had posted the Web page, and that had subsequently provided us with the necessary information as to its status (#4 above), often had no records as to when the text had been written or translated prior to posting. This information was in most cases the closest we could come to ascertaining that the text indeed belonged to the period of time (1995-2002) that the present corpora were intended to represent.

7. Source, e.g. </Provenance: Agriculture Canada\>

Here we noted the official name of the source department or agency.

8. Type of translation service used (if known), e.g. </ In-House and Consultant\>

The various Government of Canada departments and agencies that answered our queries reported that they had access to one of three types of translation service: "Translation Bureau, In-House, or Consultant."

Services designated "Translation Bureau" covered two alternatives, each amounting to a budget subsidy for the department or agency receiving the service.

In some cases, a "dedicated translator" was sent by the Translation Bureau to work on site at a department, while in other cases involving lower volumes of text, the government department sent texts to be translated at the Translation Bureau or by its contractors.

Some departments and agencies employed translators directly, paying their salaries through their own budgets. The service these translators provided was designated as "In-House."

Finally, some departments regularly hired commercial firms or individual freelance translators as consultants on contract. This service was designated as "Consultant" in the text's header. In practice, this third service usually supplemented the others, whenever salaried translators providing the other two types of service took leave, went on holiday, or were fully booked.

Many of the departments and agencies reported that they used both. The texts submitted by these sources were given a combined designation, that is, as "Translation Bureau and Consultant" or as "In-House and Consultant" (as in the example above).

9. Translation-related policies (Translation Bureau, In-House, or Other), e.g. </Policies and directives: TB\>

We asked all sources to describe their translation-related policies, that is, their prescribed procedures for word counts and translation quality assessment, and their procedures for identifying and employing qualified translators. These were identified by the type of service, in the two cases where that service was offered by salaried professionals following government directives.

Although the third type of service, the "consultants," were under government contract, they were in most cases anonymous, and could not be contacted to ascertain whether any of their policies differed from those contained in the government directives.

10. General Topic, e.g. </How-To Manual for the Public\>

Although it was not intended for use in the present study, we made note of the information contained in this annotation for possible use in future research that might include genre and style among its hypotheses. The general topic or genre of each text was assigned based on common sense, following the precedent of researchers such as those who gathered the Brown corpus, the LOB (Lancaster-Oslo/Bergen) corpus, and the London-Lund corpus. Their categorization of texts was, according to Biber (1990: 261), based on "folk 'genres'," which were "readily distinguished by mature speakers" and defined "primarily on the basis of format, purpose, and situational context." Laviosa (1997: 291; 298; 305; 310) followed suit in conceiving a corpus typology that includes 14 different common sense classifications (which she called "text genres" or "institutional text categories") to the texts in her comparable corpus.[83]

11. Text title, e.g. </Title: Canadian Rural Partnership: Asset Mapping: A Handbook\>

The text title was used to determine the general topic.

[83] Laviosa's text categories included biographies, business guides, coursebooks, customs and folklore, fiction (general and short stories), culinary, "general knowledge," tourism guides, promotional pamphlets, in-flight magazines, newspaper articles, "official reports (published and/or public domain)," transcripts of speeches, and general travel writing (1997: 298). The assignment of categories was "based on consensus" (1997: 305; 310).

The annotations listed above allowed quick, automatic retrieval of various types of information from our corpora. For instance, texts identified by type as non-specialized could be retrieved from each of the four corpora and gathered as sub-corpora for use with readability indices (see Chapter 5) by performing a search for annotation #5 (see Bowker and Pearson 2002: 82). As with annotation #10, some of them may also leave open the possibility of performing other types of analyses (of a sort not envisaged for the purposes of the present study) in future research.

We did not assign part of speech tagging to the corpora used in the present study, despite the advantages this would have brought to the research (outlined in Bowker and Pearson 2002: 83-84), chiefly because budget restrictions precluded the purchase of a tagging program license, and because demo versions of these programs were (and still are) limited to a maximum sample of around 10,000 words, with repeated use forbidden. Nor were our corpora lemmatized, for the same reasons.

3.3 Tools used to analyze the corpora

Once the corpora had been compiled in the manner described in this section, we were ready to begin our qualitative analysis. In this section we will describe three different types of tools used to observe the features of the corpora gathered for the purposes of the present study. The first is a corpus processing software suite called WordSmith Tools, affordable software well-suited to studies such as ours.[84] The second is a selection of readability indices, which were developed for

[84] WordSmith Tools is recommended by Bowker and Pearson (2002: 108) for student researchers who are interested in using electronic corpora, and who require inexpensive, user-friendly, generally robust software.

professional use by educators and technical writers, and which were readily adapted to the purposes of the present study. The third is a set of measures of corpus features, each of which was adapted from existing corpus linguistics methods used in previous corpus-based study of recurrent features of translation (see Sections 2.1.2, 2.2.2, 2.3.2, and 2.4.2).

3.3.1 WordSmith Tools, Versions 3 and 4

WordSmith Tools, a suite of corpus processing software that includes programs which make it relatively easy to search through and analyze an electronic corpus, was the main implement used to observe the characteristics of the corpora included in the present study. When the present study began in 2002, only Version 3 of WordSmith Tools was available for purchase, since Version 4 was still in development and was available only as a beta. WordSmith Tools 4, the current version, has an improved user interface and a number of added features, some of which were used in later stages of the present study.

The three main tools of theWordSmith Tools suite are **WordList***, **Concordancer***, and **Keywords***.[85] Two of these (WordList and Concordancer) were used for the present research. In what follows, we will describe those of their functions that were used in the present study. The tool used most often was WordList, which displays a set of three separate sub-lists. These lists are:

a. **Frequency List***, which displays every word in the processed corpus, in descending order of frequency, alongside the frequency of each word (i.e. the

[85] All of the features described below are fully illustrated, and their functioning explained, in Scott (2004) and Bowker and Pearson (2002: 109-133). We therefore do not include illustrations or instructions for use of the software described.

number of times it occurs in the corpus), that frequency as a percentage of running words in the text(s) from which the list was prepared, the lemmas (if any have been manually prepared)[86] and, in Version 4 only, the number of texts in which the word appears, followed by the percentage of the total texts which that number represents.

b. **Alphabetical List***, which displays every word in alphabetical order, followed by the frequency of that word (the number of times it occurs), that frequency as a percentage of running words, any lemmas, and (in Version 4 only), the number of texts in which the word appears, followed by the percentage of the total texts which that number represents.

c. **Statistics List***, which displays statistics for the overall corpus processed, as well as for each text contained in it. The Statistics List displays information about the corpus in the following order:

(1) Text files: the title of each text file included in the corpus. In the first column, the file title "overall corpus" appears.[87]

(2) Bytes (Version 3); File size (Version 4). A rough indicator of the total number of characters, according to Scott (2004).

(3) Running words in the text (**tokens***).

(4) Tokens used to produce the WordList feature output (Version 4 only).

(5) Number of different words (types).

(6) Type/token ratio.

(7) Standardized type/token ratio.

(8) Standardized type/token ratio standard deviation (Version 4 only).

---

[86] In WordSmith Tools 3, the output of the WordList feature may be lemmatized after an initial word list is obtained, by following a time-consuming but accurate manual process of hand-joining entries, or by adding to this procedure a "hit and miss" auto-joining process which must subsequently be manually corrected. In Version 4, it is possible for the researcher to develop a .dll (data language library) file and run it under the "advanced settings/advanced" feature of WordSmith Tools 4. The author of WordSmith Tools, Mike Scott, notes in the Version 4 handbook that this custom lemmatization is "not for those without a tame programmer to help" ("Lemmatization: Custom Processing"); in other words, those who attempt the process should have advanced knowledge of programming languages, such as C++, Java, or Pascal. Given our lack of knowledge of these programming languages and the lack of funding available for the purposes of hiring a software programmer, the custom lemmatization function of WordSmith Tools 4 was not attempted during the present study.

[87] The titles are numbered across the top of the display, so that it is possible to scroll horizontally to the right and see how many text files are involved in the word-list. This is a display that is meant for heavy use in Version 3, since this information is not easily obtained elsewhere. In Version 4, WordList displays a fourth list, called the Filenames List, in which the name of each text file included in the corpus is handily displayed.

(9) Standardized type/token basis (Version 4 only).

(10) "Average word length" (Version 3); "Mean word length in characters" (Version 4).

(11) Word length, standard deviation (Version 4 only).

(12) Number of sentences in the text.

(13) Mean sentence length (in words).

(14) Standard deviation of sentence length (in words).

(15) Number of paragraphs in the text.

(16) Mean paragraph length (in words).

(17) Standard deviation of paragraph length (in words).

(18) Number of headings in the text.

(19) Mean heading length (in words).

(20) Standard deviation of heading length, in words (Version 4 only).

(21) Number of sections in the text (Version 4 only).

(22) Mean section length in words (Version 4 only).

(23) Standard deviation of heading length in words (Version 4 only).

(24) Numbers removed (Version 4 only).

(25) **Stoplist*** tokens removed (Version 4 only).

(26) Stoplist types removed (Version 4 only).

(27) The number of 1-letter words.

(28) The number of 2-letter words.

(29) (The number of *n*-letter words, up to 50).

The Statistics List was essential, in the present study, for the calculation of many of the measures. However, not all of the statistics on the Statistic List were of interest: for the purposes of the present study, only those that pertain to words and sentences were used.

Another program used in the present study and included in the WordSmith Tools suite is the Concordancer, which displays all the occurrences of a search pattern (i.e. of a given word or phrase) in **KWIC*** (key word in context) format.

This program was put to use in the present study whenever it was necessary to check the context of individual occurrences of words, and particularly as a means of retrieving the sentences in which words that might be transient coinages were to be found (see Section 4.1).

We also used the **Dispersion Plot*** feature of the Concordancer, which provides a display that shows whether a given search term is unduly concentrated in a single area of the corpus. In both Version 3 and Version 4 of the Concordancer, the Dispersion Plot marks each occurrence of the search term as a physical distribution within the texts that make up a corpus.

Finally, it should be noted that there proved to be certain limitations to the above tools. First, certain statistics, such as the number of words counted, varied slightly when certain settings, such as type/token ratio and collocate horizons, were changed. Moreover, since our corpora were not tagged or lemmatized, it was necessary to manually discard thousands of irrelevant instances from the data retrieved. Nonetheless, it would have been extremely difficult to undertake the present study if WordSmith Tools had not been available.

3.3.2 Readability Indices

Readability indices are instruments which measure the degree of complexity of a written text's vocabulary and syntax, and which correlate the results with a given reading population's skill level. Readability indices provide a widely recognized, independently-established conceptual continuum (ranging from "easy to read" to "hard to read") which is useful to our investigation of levelling-out.

Although readability indices date back only to the early twentieth century, a concern for readability has possibly been present since the invention of writing. Irving Lorge (1944) notes that the first recorded attempts to examine what has come to be called readability were made in the eighth century by Talmudist religious teachers, who did word counts as a way of distinguishing unusual meanings by frequency of occurrence.[88]

Foundation vocabularies based on the "scientific principle" of word frequency counts were produced by Victorian-era researchers, who attempted to relate vocabulary to the level of a text's reading difficulty (Klare, ed. 1969: 30). In 1921, the first *Teacher's Word Book*, a list of 10,000 words forming a (theoretically) basic vocabulary, was published in English by E. L. Thorndike, based on his estimate of the frequency of occurrence of words in a sample of English texts. Further editions of the *Teacher's Word Book* were published in 1932 and 1944, each adding 10,000 more words to the vocabulary, and ranking the words according to a more reliably calculated frequency and range of occurrence, using more extensive counts. Klare, ed. (1969: 30, 38) notes that in 1923, Lively and Pressey developed the first readability formula based on Thorndike's frequency tabulations.

Vocabulary was not the only means by which early researchers attempted to analyze readability. In 1893, L. A. Sherman attempted to link differences in the readability of various literary styles by means of sentence length (mean number of

[88] Also mentioned in Klare, ed. (1969: 29-30).

words per sentence), and by counting simple vs "predicated" (complex) sentence structures.[89]

Readability formulas measuring both vocabulary and syntax have been in common use since at least the 1940s, by teachers wishing to estimate the reading difficulty level of educational material. Technical writers have since adopted them to estimate whether a given professionally-written text is too difficult or too easy for its target readership.[90] The formulas are validated, that is, tested and shown to be valid, for the reading population of a given language within a given educational system.

The readability indices which we have selected for use will be discussed in more detail below.[91] There are many different validated formulas for assessing readability in English and French, although there are fewer for French than for English.[92] Most are based on two variables: words and sentences. Words are measured by their length, in characters or in number of syllables. Sentences are measured by their length in words. Some formulas include a third variable: the number or percentage of linguistic units in a text that fall into a given category,

---

[89] As Klare, ed. (1969: 195) notes, Sherman found that both sentence length and predication were decreasing, while the use of simple sentences was increasing, at the end of the nineteenth century.

[90] Simple, speedy, inexpensive application is considered important for these formulas, which must work on a large number of texts on a daily basis in a professional setting.

[91] A number of alternative measures of text readability do exist, but for the most part, they involve reader participation. These include testing for individuals' reading speeds, having readers rate texts for comprehensibility, giving memory tests in which subjects are asked to reproduce a text after reading it, giving cloze tests, having readers follow think-aloud protocols, and recording a subject's eye movements as he or she reads. These alternative measures do not seem to be widely used, perhaps because they require heavy investment of time and money, and because their results can be difficult to interpret. Implementing them would have been beyond the scope of this corpus-based study, which also has limited resources. The simpler "surface"-measuring formulas we selected all have excellent track records. We considered it feasible to adapt readability indices to our purposes, since they were originally validated for the educational milieu and then later appropriated for professional use by adults.

[92] We were able to find only two validated formulas that had been developed for use with French texts in Canada. One of them, SATO, is a widely marketed, comprehensive software package of text-analysis tools that include readability levels calculated using the Gunning formula, (defined and explained below). SATO is meant to be used by government project managers and analysts, the legal and commercial sectors, and various branches of academia including Education, Sociology, Psychology, and Linguistics (Daoust 2003). The other is a manual formula adapted from the English-validated Gunning-Fog Index by the Institut Canadien des Actuaires (Tremblay 2000). There is also a formula that is strongly validated for European French: the three-variable Henry-de Landsheere, the manual version of which is rather labour-intensive. The Henry-de Landsheere formula is apparently available in a European software package (Vandendooren 2000).

such as the percentage of sentences written in the passive voice, or the percentage of "hard" vs. "easy" words.[93]

A series of readability indices was selected for use in this study. For the English corpus, four different formulas were used: a) the Microsoft Office Word 2000 Readability Scores (Flesch Reading Ease and Flesch-Kincaid Grade Level); b) the Fry Readability Graph; c) the Gunning-Fog Index; and d) the Lix formula for measuring readability. For the French corpus, only three different formulas were deemed to be useable for the purposes of this study: a) Henry-de Landsheere; b) ICA (adapted from de Landsheere and Gunning-Fog); and c) Lix.[94] Each of these indices will be discussed in greater detail in the chapter on levelling-out (Chapter 5), where their applications can be clearly shown.

3.3.3 *Type/token ratio, lexical density ratio, mean sentence length*

It is generally recognized that breadth of vocabulary can be measured in a corpus using type/token ratio, which is the ratio of word forms (types) to running words (tokens).[95] This ratio must usually be standardized.[96] In Laviosa-Braithwaite's 1996 study, type/token ratio was approached as a complex statistical workup on individual sets of texts. Following Laviosa-Braithwaite (1996), we calculated the mean of the type/token ratios (in percentages) of all the individual

---

[93] "Hard" words being defined as all words not appearing on a list of "easy" words included with the formula and equated with a basic vocabulary.

[94] For more on French readability indices, see for instance Björnsson (1968), Cherry (1982), Gougenheim (1969), Gougenheim, *et al.* (1967), Gunning (1968), Nilsson (2002), Tremblay (2000), and Vandendooren (2000).

[95] See for instance Zanettin (2000: 109-110): "The more word forms are used with respect to the total number of words..., the wider the range of vocabulary used in the text ... ."

[96] Unstandardized type/token ratios simply drop as the number of words in a corpus rises. This is because the number of tokens (running words) begins to exceed the number of existing types. For instance, a 1,000 word text might have a type/token ratio of 40%, whereas a corpus of 4 million words would have a much lower ratio, e.g. 2%. The unstandardized type/token ratio therefore tells us little. It is when the type/token ratio is standardized, that is, calculated every *n* words, that it shows vocabulary range. WordSmith Tools' concordancer standardizes type/token ratios every 1,000 words by default. However, it can be set to standardize at shorter or longer intervals.

texts in each of our corpora, standardizing at fifty words (***n* = *50***), in order to take into account the shortest texts in our corpus.[97] Type/token/*n* ratios were obtained in the present study from automatically-generated WordSmith Tools output.

The lexical density ratio is the proportion of content words to running words in a single text or whole corpus. This ratio is usually expressed as a percentage. In an electronic corpus, the number of content words can be obtained automatically, by subtracting the number of **function words*** from the number of running words (Stubbs 1986: 33). We obtained a Frequency List for each corpus, with the appropriate (English or French) function-word stoplist in place.[98] Using the frequency counts so obtained, we then calculated lexical density ratios, using the following formula to express the ratio (i.e. the proportion of content words to running words) as a percentage:

> Lexical density ratio = 100 x L/N
> Where
> L = the total number of content words or "lexical" words ("L" is for "lexical")
> N = the total number of running words ("N" is for "number")

Mean sentence length is a measure used both in the field of readability instruments, where it is often designated by the acronym "**ASL***" (for "average sentence length"), and in corpus-based study (e.g. Scott 1999b).The hypothesis that simplification is a recurrent feature of translation predicts that, in a comparable corpus, the translated texts will have shorter mean sentence lengths than the non-

---

[97] In the 1996 study, Laviosa-Braithwaite had standardized every 100 words, the shortest length of a text in her corpus (personal communication). Among the texts in our corpus, the shortest was 91 words (in FST). The two lowest intervals at which our concordancer (*WordSmith* 3.0) could be set to standardize were 50 and 100. We therefore standardized at $n = 50$. Note that for calculating the standard deviation of the type/token/*n* ratios of our texts, we standardized at the value closest to the **mode*** value for each corpus. Thus, for the English texts, we standardized the type/token ratios at 350 words, while for the French texts, we standardized at 250 words.

[98] Two stoplists were used, one for English (published by WordNet) and one for French (published by Jean Véronis). These may be viewed at Lee (2001).

translated texts. To test this hypothesis on the present corpora, we obtained mean sentence lengths automatically, using WordSmith Tools.[99]

3.4 Corpus analysis

Detailed explanations of the many uses of corpus linguistics methods for Translation Studies research are found in Olohan (2004), Bowker (2002), and Bowker and Pearson (2002). Therefore, in what follows, we will only briefly discuss the advantages and disadvantages of the methodology for the present study. We will then explain why we have taken a mixed approach to corpus analysis, one that is both (descriptively) quantitative and qualitative.

3.4.1 Advantages of the methodology for the present research

There are three main advantages to using corpus-based methodology for the present research. First, since it allows large numbers of authentic translated and non-translated texts to be grouped together and investigated, corpus-based methodology makes it possible to observe the object of the present study both directly and on a large scale. Second, corpus processing tools such as wordlisters are particularly useful for finding—and verifying— the observable presence of recurrent patterns such as those that are investigated in the present study. Third, computer-assisted analysis is likely to be much more accurate than a manual analysis would be, since more time can be spent studying relevant material (Bowker 2003: 175). In what follows, we will discuss each of these advantages in detail.

[99] WordSmith Tools provides the arithmetic mean of all sentence lengths in the selected texts. For this calculation, we used WordSmith Tools 4.

The qualities of corpus research make it an excellent tool for descriptive studies of translation such as ours. Tymoczko (1998: 656) stresses that corpus research is flexible enough to offer a variety of new directions for the discipline of Translation Studies as a whole.[100] In general, one must agree with Andrew Chesterman (2004: 46) that "corpus-based research into translation universals has been one of the most important methodological advances in Translation Studies during the past decade...in that it has encouraged researchers to adopt standard scientific methods of hypothesis generation and testing."

Based as it is on principles of observational study, analysis of electronic corpora provides authentic, accurate, verifiable data, which can then function as a standardized reference for further study. Compared to the method of using introspection, which relies heavily on subjective judgement or intuition, observation of texts assembled into electronic corpora can provide much more reliable data for research on the frequency or rarity of a given word or grammatical form. Corpus-based research is complementary to introspection: naturally occurring text data can serve as a control for findings made with anecdotal evidence, elicited performance, or computer simulation, as well as with with artificial "thought up" data produced through pure introspection.[101]

---

[100] Corpus-based research presents an opportunity for the discipline to "turn away from the invidious competition and isolation" between the linguistic and cultural approaches, which can now proceed to "explore their reciprocal relationship," she adds (Tymoczko 1998: 657).

[101] Furthermore, with qualitative analysis, "very fine distinctions" can be drawn, while with quantitative analysis, findings can be "generalised to a larger population" and "direct comparisons may be made between different corpora, at least so long as valid sampling and significance techniques have been employed" (McEnery & Wilson 1996: 62).

Corpus-based research provides much more than what Chomsky (1965) dismisses as uninteresting "performance data."[102] Corpora show how multiple individual linguistic "performances" (spoken or written) aggregate, forming patterns that individual users cannot have intended to produce.[103]

These patterns are rendered observable when data is sorted and displayed using corpus analysis tools such as wordlisters and concordancers. Large-scale patterns of language usage become relatively easy to discern, making it possible to hypothesize which patterns will be repeated in other corpora (Kennedy 1992: 335).[104] This ready observation of large-scale patterns is ideal for Translation Studies, a fact which has been noted by a number of scholars in this field. The patterns made visible can be both quantitative and qualitative—numerical and non-numerical. For instance, it is possible, according to Venuti (2000: 336), to discover "regularities" in the history of translation by studying relevant diachronic corpora. Attributes of many types (on cultural and other non-linguistic aspects of a corpus) can easily be retrieved and studied by the researcher if this information is appended, using annotations, when a text is inserted into a corpus (McEnery and Wilson 1996: 23-24; Laviosa-Braithwaite 1996: 69-74, 80, in her Appendices 9 and 18; Kenny 2001: 117-120, 215-216; Bowker 2002: 68-70; Bowker and Pearson

---

[102] Chomsky objects to performance data on the grounds that it is not a good mirror of competence, because it is affected by "memory limitations, distractions, shifts of attention and interest, and errors" (Chomsky 1965: 3). But overall language use in a population must be distinguished from such individual "performance."

[103] An example of such a verifiable pattern would be the percentages of certain classes of words in a corpus (Stubbs 1996: 234).

[104] Stubbs (2001: 221) notes that corpus methods can "organize huge masses of data" into "coherent, easily visible patterns," with discernable proportions in relation to other patterns. Kennedy (1992: 339, 341) points out that much traditional linguistic description contains no such statements of proportions: "It is as if chemists knew about the different structure of iron and gold, but had no idea that iron is pretty common and gold is very rare." Frequent occurrence of lexical or grammatical patterns is "good evidence of what is typical and routine in language use," and the existence of these patterns is easily verified, Stubbs notes, since corpus linguistics is based on "two principles of empirical observational study: 1) The observer must not influence what is observed... 2) Repeated events are significant" (2001: 220-221). See also Teubert 1999 and 1999b, *passim*.

2002: 75-91).[105] The use of corpora is particularly well suited to the search for evidence of patterns of linguistic behaviour, such as recurrent features of translation.

3.4.2 Limitations of the methodology for the present research

Potential concerns about corpus-based study of translation include the degree of representativeness that a modern language corpus can be assumed to have, the necessity for subjective judgement about what constitutes adequate corpus size, the practical difficulties involved with gathering a corpus, the scope of contextualization that a corpus-based study can provide, the method's emphasis on surface features and statistical norms, the difficulty in detecting features that are lacking in a corpus, and the inaccessibility of abstract concepts to first-order observation such as that provided by the methodology. Below we list and discuss each of these limitations in the context of our study.

1. The representativity of the sample and the findings.

Strictly speaking, it is unlikely that an electronic corpus of modern language texts can constitute a strictly randomized, statistically representative sample. All living languages are in a constant state of flux, and new spoken and written texts are added to them every day. The population of texts is practically infinite (i.e. at the time we are writing, unknowable), and with very few exceptions, a rigorous randomized statistical sampling of all existing utterances and/or texts in a single

[105] However, as with many techniques, simplicity is desirable, and excessive annotation should be avoided. Leech (1991: 24) comments that data whose elaborate annotation is governed by a set theory may produce biased results. Sinclair (1987: 107) notes that if the objective is to make generalisations based on one's observations, "a means of recording structure must be devised which depends as little as possible on a theory. The more superficial the better." See also Hunston and Francis 2000: 18-20.

corpus, no matter how large, is likely to be a very difficult undertaking.[106] As Partington (1998: 148) puts it, "any corpus of data is only truly representative of itself and not of the entire universe of study."

Since the notion of "language as a whole" is purely theoretical (Biber 1988, Biber 1993, Biber *et al.* 1994), findings of corpus-based empirical study can rarely be generalized to an entire language. As with most empirical work, we do not claim that our findings will allow any definitive statistical inferences about the total population sampled.

There is no expectation, in fact, that research of this kind will eventually yield universal "laws" of translation construed to be applicable to all instances of translation, at all points in time and space. Rather, in keeping with the principles of inductive empirical research, we merely harbour the humbler hope of arriving at findings that can be put to use in future study.[107]

2. The adequacy of the corpus size.

Adequate sample size is rather difficult to determine. Calculating sample size is a matter for the researcher's judgement, according to Bowker and Pearson (2002: 45-46), who list factors such as the particular needs of the study, the availability of text data, and the amount of time available to the researcher for gathering the corpus. In general, the larger the corpus, the more reliable the results

[106] However, as Folkart (1989) points out, there is a type of text with a "mathematical" dependency on formulaic language that makes translation highly predictable. Such specialized texts as weather bulletins or course outlines constitute limited homogeneous populations from which randomly sampled, statistically representative corpora might more easily gathered.

[107] Our findings must, in other words, be considered with at least one caveat: Partington (1998: 146) cautions that results of a corpus-based study can be considered reliable "only for the portion of language contained in that corpus"; only if other empirical studies have the same results can they be interpreted as generalizable to larger populations.

are believed to be, although Stubbs notes that even "relatively modest" corpora do provide adequate evidence for many of the features of language (2001: 224).

Biber (1990: 269) demonstrates that a corpus of "relatively small" size can be adequately representative of the linguistic variation of the genre sampled. Bowker and Pearson (2002: 45-46) recommend against assuming that "bigger is always better," since a small but well-designed corpus may provide more information pertinent to the goals of a study than a larger corpus that is not customized to meet research needs. Our corpus, although limited in size, comprises rigorously authenticated translated texts, and clearly distinguishes them from non-translated texts. This makes it possible to meet our objectives.

3. The logistics of gathering a corpus.

Unless a suitable pre-existing corpus is available, a great deal of time must be spent designing, compiling, annotating and testing an electronic corpus, before any investigation of its features may begin.

Preparing and collecting the texts can be an especially time-consuming task, even when, as in the case of the present study, the texts are available on Government of Canada Web sites, i.e. in electronic format, and merely require file conversion, rather than (the considerably more onerous) procedure of scanning and optical character recognition (Bowker 2002: 23-30; 37-39).

Moreover, in our case, we had to identify texts as originals or translations, and this information was not readily available from the government departments and agencies concerned. However, we did manage to overcome these logistical problems and compile an adequate corpus.

4. The contextualization.

It might be argued that KWIC concordances, along with other output of corpus processing software, lack the complete context of communication. For example, in a parallel corpus, it could be difficult to find instances of compensation in the TT that are distant from their location in the ST.[108] An automated search not executed carefully enough could fail to find the point of compensation in the corresponding translation, since the translated "nuance" may have been conveyed well away from its location in the original text. However, such cases occur infrequently and are unlikely to affect the overall findings of research carried out on large corpora.

Furthermore, most corpus processing software, including that used in the present study, allows the user to view output in incrementally larger contexts: the display may be expanded to whole sentences, whole paragraphs, or even whole texts, and one individual text is interpreted alongside many others.[109] Moreover, absolute contextualization was not considered necessary for the purposes of the present study, which is restricted to observation of words and sentences.

5. Surface structure emphasis.

It might be argued that surface features are generally overemphasized in any corpus study, since concordancing software does not allow first-order observation of either "deep" linguistic structures or sociocultural factors. To this it must be countered that defined focus is a necessity for any research. "Every observational

---

[108] It should be recalled that compensation is a translation strategy by which "a nuance that cannot be put in the same place as in the original is put at another point" of the phrase or text (Vinay and Darbelnet 1995: 341).

[109] Indeed, the letter "C" in the acronym KWIC stands for "Context": see Glossary.

tool emphasizes something, and we can't study everything at once," Stubbs notes (2001: 222; see also Partington 1998: 144).

While our study does focus on surface structure features such as words and sentences, it constitutes a beginning that could lead to the observation of other, more subtle features in future study.

6. Exclusion of rare features.

Baker (1996: 179) warns that researchers may be tempted to ignore "one-offs" among the large-scale patterns found in a corpus. While dwelling on the statistical norms of a corpus can hinder recognition of exceptional cases, the latter should not be too readily discarded, as they can provide clues to the creative use of language in translation, according to Baker (*ibid.*).[110]

Care must be taken, then, to build observation of rare occurrences into corpus-based studies, when this is appropriate. Biber (1990) notes that the study of smaller specialized corpora can reveal variation in language that might be averaged out across very large corpora. However, with one exception (the use of possible coinages to measure normalization), the present study is a deliberate search for trends, not rarities.

7. The impossibility of verifying lack of existence.

---

[110] "Once we have very large amounts of text on the computer, and the ability to generate all kinds of statistics and frequencies at the press of a button, there is a strong temptation to emphasise the norm, what is typical, at the expense of the one-off, the more creative use of language... . One of the main reasons we want to study the patterning of any kind of language or text production, including translation, is that patterns are the backdrop against which creativity can take shape: norms enable the creative use of language and identifying them allows us not only to capture universal features of translation, and hence understand translation as a phenomenon in its own right, but it also allows us to make sense of the individual example" (Baker 1996: 179).

Rare is the methodology that can prove something to be non-existent: it is much easier to show the opposite.[111] Partington (1998: 146) notes that the absence of a structure or phrase from a corpus is no proof that it does not occur elsewhere. Concordancers can provide only positive data, and cannot show what is impossible, what typically does not occur, or what has not yet occurred anywhere, any more than they can display what has been omitted. Pym (1998: 69-70) discusses the difficulty of searching for such items, and of constructing a "wholly negative" corpus consisting of, for instance, translations that were never published.[112]

This epistemological problem makes it impossible to conduct truly "universal" studies, as Tymoczko (1998) notes. This is true even of studies on translation universals such as ours. However, just as our research partially replicates that of other scholars, such as Laviosa-Braithwaite (1996), so replication of our study using other corpora may in future eventually lead to confirmation of certain recurrent features of translation.

8. The impossibility of first-order observation of abstractions.

Direct observation of abstract concepts is no more possible with this methodology than with any other. We will find corpus-based study of little use in a search for such things as Lecercle's (1990) translation "remainder." Halliday (1992: 64) notes that "I cannot ... ask the system to retrieve for me all clauses of mental process or marked circumstantial theme or high obligation modality."

---

[111] On this and other logical problems of inductive method, see Popper (1972, 1972b) and Russell, Bertrand (1959). On poorly constructed arguments in linguistics, see Harpaz (1998).

[112] Pym goes on to note that "projects for such negative histories have nevertheless been devised," and cites studies by Coste (1988) and Venuti (1995: 253-260) among them.

However, the present hypotheses concern the product, not the process, of translation. They are arguably well suited to this type of research question.

9. The lure of engaging in statistics for statistics' sake.

Tymoczko (1998: 658) warns agains making statistical methods ends in themselves in Translation Studies. Just as dwelling on the statistical norms of a corpus can hinder recognition of exceptional cases, so can wielding advanced inferential statistics for their own sake hinder understanding of the significance of findings. Too-literal interpretation of tests for significance can cause the researcher to wrongly discard important findings. With the study of language especially, results that do not test out as highly significant do not always mean that the null hypothesis is correct, as Woods *et al.* note (1986: 127).[113] Preoccupation with statistical significance may in fact "prevent us from detecting a substantial failure in the null hypothesis" (*ibid.*). Nor can the linguistics researcher usually know "how large is the probability of making a Type 2 error," that is, how likely it is that we will incorrectly consider a null hypothesis confirmed and our own hypothesis therefore denied (*ibid.*).

Furthermore, overly complex presentation of results can obfuscate and confuse. For instance, Stubbs (1995) maintains that since there is no such thing as a "random corpus" in linguistics, attributing levels of statistical significance to values obtained from a corpus through the t-test (which compares the means of random variables, Sanders 1990: 178, 272-274; Norman and Steiner 2003: 35-40) is arguably not a valid procedure. Stubbs advises linguists to "keep an eye on the

[113] It should be recalled that when the null hypothesis is supported, the research hypothesis is considered to be unsupported.

original raw frequencies," using any clear way of summarizing them, and notes that hiding the original values can make them more difficult to interpret (1995: section 4.7).

In the light of these objectives and recommendations, we have aimed in the present study for clear presentation of descriptive statistics, and have avoided using inferential statistics. Moreover, where performing complex statistical analyses would have detracted from our goal of obtaining a clear comparison of translated and non-translated texts, we have opted for the simplest and most basic presentation of the data statistics—whatever was closest to the "raw" numbers.

We would argue that basic descriptive statistics (such as raw frequencies, percentages, ratios, means or **averages***, and standard deviation) are a suitable way of presenting many of the findings of corpus-based studies of translation, in particular those that aim at developing and exploring preliminary hypotheses.[114] Replication of corpus-based studies is key to their contribution to knowledge in our discipline. There is a potential problem with the replicability of those studies that rely heavily on advanced inferential statistical methods, some of which are probably less than accessible to most researchers trained in Translation Studies graduate programs.[115] It may not be too much of an exaggeration to say that many of the Translation Studies researchers who propose to do corpus-based work are

[114] Many of the recent corpus-based studies of translation are working on newly-formed, tentative hypotheses. And it does appear that sticking with the basic statistical measures is the choice of the majority of these researchers.

[115] We are by no means saying that advanced inferential statistics have no place in CTS research. Findings in support of hypotheses that have been the subject of repeated study, and that are thoroughly refined, can be verified using statistics that confirm significance, as well as reliability and validity. See for instance Biber 1990, in which certain issues of text length and text variety in corpus design are satisfactorily resolved using reliability coefficients, factor analyses, and other inferential statistics.

likely to be academically-trained translators, interpreters, or terminologists,[116] with little or no training in statistics.[117] Thus, while we do use descriptive statistics in the present study, we approach them qualitatively.

3.4.3 *Qualitative and quantitative analysis*

Corpus-based study of translation is compatible with both (descriptive) quantitative and qualitative methodology. Babbie (2004), Casebeer and Verhoef (1997), Creswell (2003) and Denzin and Lincoln, eds. (2000) all compare qualitative and quantitative methods. Quantitative methodology, they note, is usually associated with reasoning (i.e. logic) that is deductive and objective. Quantitative research questions are product-oriented and closed, the analysis performed is numerical, and the inferences drawn are statistical (*ibid.*). Conversely, qualitative methodology is associated with reasoning that is inductive and hermeneutic or subjective. Qualitative research questions are exploratory, process-oriented, and open-ended, that is, they may be modified during the research, and the analysis is both comparative and based on narrative description (*ibid.*).

The two approaches are not mutually exclusive, as can be seen in a number of corpus linguistic studies, in which it is standard to reason on an inductive "research-then-theory" basis, and in which corpora are perused for patterns of frequencies of occurrence. Patterns regarded as significant are those that subsequently prove to be repeated in different corpora, helping to establish what

---

[116] Indeed, Translation Studies research carried out by individuals with no training in the discipline is likely to be received with scepticism by those actively working in it.

[117] See for instance the curricula for the graduate programs and certificates at the following schools: Central Institute of Indian Languages, Mysore, India; Collège St-Boniface, Manitoba, Canada; Handelshøjskolen i København, Danmark; Kent State University (Ohio) Institute for Applied Linguistics; Universitat Autònoma de Barcelona Facultat de Traducció i d'Interpretació; Glendon College at York University (Canada); Stockholm universitetsTolk- och översättarutbildning; the École Supérieure d'Interprètes et de Traducteurs, Université de la Sorbonne Nouvelle - Paris III, and the Imperial College of London. Not one course in statistical methods is found among these programs.

Partington calls "inductively achieved hypotheses" (1998: 149). If a pattern consistently fails to turn up in further corpora, then the hypothesis is eventually discarded (*ibid.*). If, contrariwise, the pattern is repeated, then "it is necessary to collect still more data and analyse it for similar patterns" (*ibid.*). Since this is the overall line of reasoning of the present study, our research may be considered to be mainly qualitative, our reasoning inductive, our conclusions hermeneutic. This is a qualitative study that is partly supported by a quantitative method.

In a purely qualitative approach, little or no attempt would be made to assign frequencies to linguistic features identified in the corpus (McEnery and Wilson 1996: 62). In a purely quantitative approach, the features of the corpus would be assigned rigid categories and counted. The findings would be generalizable to a larger population with some certainty, but at the cost of inflexible classification of features which might be better described in such a way as to allow for what McEnery and Wilson call the "ambiguity which is inherent in human language" (1996: 62-63). This is why many corpus-based studies in linguistics and translation prefer to use a mixed quantitative and qualitative approach, with an emphasis on interpretation that is sometimes (but by no means always) supported by tests of statistical significance.[118]

This pragmatically mixed approach is seen in studies such as those carried out by Aijmer (1998), Altenberg (2002), Baker (2000), Camargo (2003, 2001), Doherty (1998), Dyvik (2004, 1998), Eskola (2004), Gellerstam (1986), Hasselgård (1998), Jääskeläinen (2004), Johansson (1997), Johansson and Hofland (2000),

[118] Comparatively few corpus-based studies of translation have employed strict quantitative methods alone (see however Harri Jantunen (2004).

Laviosa (2002b), Laviosa[-Braithwaite] (1996; 1997; 1998), Kenny (2001, 2000b, 1999b, 1998, 1998d, 1997), Kujamäki (2004), Mauranen (2004, 2000), Munday (1998), Nilsson (2004), Olohan (2003, 2002, 2001), Olohan and Baker (2000), Paulussen (1999), Pápai (2004), Puurtinen (2004, 1998), Salkie (2000, 1997), Santos (1998, 1996), M. N. Scott (1998), Tirkkonen-Condit (2004, 2002, 2000), and Øverås (1998). In these studies, basic descriptive statistics (such as raw frequencies, numerical means, percentages, and ratios) are used, and the results are qualified and interpreted.[119]

Like the above scholars, we too combine basic descriptive quantification with qualification. We reason inductively but pose product-oriented questions, and analyze numerically but interpret the results comparatively. However, our research questions are to a large extent pre-specified and outcome-oriented, and our analysis is always in part numerical and based on descriptive statistics, traits usually associated with quantitative research (Babbie 2004; Casebeer and Verhoef 1997; Creswell 2003; Denzin and Lincoln, eds. 2000).

Furthermore, in the present study we use tests of significance to determine whether our results are likely to be due to chance or not. Statistical tests that show significance add weight to our interpretation, because statistically significant findings indicate both that the sample size is probably adequate for the chosen measure, and that the measure is probably valid, i.e. that it is, in fact, discerning a

[119] The potential use of statistics for the purposes of hermeneutic, cultural studies-oriented translation research has also been acknowledged, despite the traditional scepticism towards empiricist work often encountered among cultural studies enthusiasts. Pym (1998: 71; 74; 77-79; 113) notes that although "the ideological use and abuse of statistics is notorious" and "no statistical distribution of translations...is entirely neutral," statistics can be presented in "cunning pictures and prose," and can provide the translation historiographer with a useful means of refining hypotheses and discerning trends that he might otherwise have missed. "Although quantitative methods cannot in themselves write good history, they can certainly help us head in the right direction," Pym declares. Statistics, he says, can provide clues to cultural-studies oriented questioning of issues involving power and ideology: "quantitative fluctuations ... can indicate sites of disturbance in the past" (*ibid.*).

distinction between translated and non-translated texts. Following consultation with Dr. Gilles Lamothe, we used the Z-test and the F-test (see Lamothe 2005). Each of these tests gave us a p-value, or level of statistical significance (i.e. the chance that the findings are random). On the advice of professional statisticians, who pointed out that ours is a social sciences study with no previously-established baseline, we accept the present research hypotheses at a confidence level (i.e. the chance that the findings are not random) of 90 to 100%.[120]

We used the Z-test to compare proportions. Like all tests of significance, the Z-test to compare proportions begins with a null hypothesis that assumes there is no difference between two population proportions, which we can designate as $\rho 1$ and $\rho 2$, respectively. In the context of statistical hypothesis testing, the research hypothesis is said to be the alternative hypothesis (H1). The null hypothesis in our case will be H0: $\rho 1 = \rho 2$. If the evidence from the sample is strongly in favour of the research hypothesis, then we may reject the null hypothesis and accept the alternative hypothesis (Lamothe 2005). In the present study, our research hypothesis concerning proportions is generally that $\rho 1 > \rho 2$. In other words, we hypothesize that $\rho 2$ (the translated texts in a given language), will have lower proportions of coinages, lower type/token ratios, lower lexical densities, lower proportions of syntactic explicitators, and lower mean sentence lengths compared to $\rho 1$ (these same ratios in non-translated texts in the same language). This alternative hypothesis can be expressed as H1: $\rho 1 - \rho 2 > 0$. The Z-test procedure for comparing proportions is shown in Appendix X (and see Lamothe 2005).

[120] Although the conventional confidence level is 95%, it was pointed out to us that many exploratory studies in the social sciences accept confidence levels of 90%. We are grateful to Dr. Ravi S. Pendakur, Department of Social Development, Government of Canada, to Dr. Gilles Lamothe, University of Ottawa, to Dr. Ian Bruce, Bell and Curve Statistics, Boston, and to Mr. Hasan Alam, Analyst, Human Resources and Skills Development Canada, for their advice in this matter.

We used the F-test to compare variances. In comparing our two population variations, which we can designate as σ^2_x and σ^2_y or the variance of the first and second populations respectively, the null hypothesis will assume that there is no difference: H0 : $\sigma^2_x = \sigma^2_y$.[121] In the present study, our research hypothesis concerning the standard deviation of the readability scores in each population is that the variance of the first population is greater than that of the second: $\sigma^2_x > \sigma^2_y$. In other words, we hypothesize that the translated texts in a given language (σ^2_y), will have scores with lower standard deviations compared to those of the non-translated texts in the same language (σ^2_x). This alternative hypothesis can be expressed as H1: $\sigma^2_x - \sigma^2_y > 1$. The F-test procedure for comparing variances is shown in Appendix X (and see Lamothe 2005).

Each of the above-described tests produces a p -value, which is "the probability of observing a more extreme value than the current observed value of a test statistic that is in favour of the alternative hypothesis" (Lamothe 2005). The p-value indicates the level of significance at which we can accept or reject the null hypothesis.

In what follows, we investigate the four hypothesized recurrent features of translation using various measures, all of which are dependent on objective counts. Some of these measures (such as the number of coinages, type/token ratios, lexical densities, number of optional syntactic elements) generate comparative proportions, while others generate comparative means (i.e. mean sentence lengths) and comparative variances (i.e. standard deviations of score sets). For each measure of a recurrent feature, we summarize the raw data and present the results of the

[121] Note that the Greek letter sigma(σ) is often used as a mathematical symbol to designate standard deviation.

appropriate statistical query. Complete tables for the statistical tests can be found in Appendix X.

4. Investigating Normalization, Explicitation, and Simplification

As indicated earlier, three of the four recurrent features of translation proposed in Baker (1996) have been investigated previously in corpus-based study of translation. These three hypothesized "universals" are normalization, simplification, and explicitation. In keeping with the methodological approach and goals of the present research, we have extended the investigation of these features to the corpora gathered for the purposes of the present study. The measures used in the previous studies have been replicated as closely as possible, in order to allow our research results to be compared with those of the previous studies. In what follows, our research and results are described, and our interpretation of the results reported, for each of these three hypothesized recurrent features of translation.[122]

4.1 Normalization

As discussed in the Literature Review, a previous study (Kenny 1999b) had investigated normalization through a number of coinages found in literary texts translated from German into English. We investigated normalization in the translated texts of our French and English corpora, which, because they are non-

[122] Puurtinen (2003: 148-150) argues that a consistent difference in the frequency of use of one grammatical form (non-finite constructions) in one language of translation (Finnish) is evidence against the hypotheses of normalization, explicitation, and simplification as "translation universals." However, the corpora cited by Puurtinen (2003) are restricted to a specialized sector of the literary market, that of the translation of canonical Victorian English children's literature (i.e. *The Wizard of* Oz) into contemporary Finnish.The corpus used in Puurtinen (2003) does not appear to us to have been designed for the specific purposes of investigating "translation universals" per se, and we have therefore not cited it above. We would add that there appears to be ample opportunity for the future gathering of comparable non-literary corpora in both Finnish and Swedish, given the official bilinguialism of Finland. This would allow comparison of translated and non-translated texts in more than one language, providing what might perhaps be a more suitable corpus design for the investigation of hypotheses of recurrent features of translation.

literary, and because they include two languages of translation instead of one, can add complementary information to the knowledge of normalization versus creativity gleaned in Kenny (1999b). To conduct this part of our study, we hypothesized (as noted in the introductory chapter) that the vocabulary of the present translated texts was normalized in that it contained fewer unattested (possibly "creative") words than the non-translated texts, in both corpora (English and French).

4.1.1 Normalization: Measures

Our general approach to analysis was inspired by that of Kenny (1999b), who had previously formulated a specific hypothesis of "lexical" normalization and tested it in pioneering research on a literary German-English parallel corpus gathered in Britain (Kenny 1999b: 2, 153-155, 191). She identified some of the "creative" words and phrases found in the original German literary works and then compared them to their English literary translations (Kenny 1999b: 140-143; 152-203). Many of the ST items that Kenny selected for comparison were coinages—words with unattested orthography or atypical morphology (*ibid.*). Kenny found that in most cases, the coinages identified in the ST had been translated into English using words or phrases that were both attested and common (1999b: 156-177). Since the creative vocabulary of the original literary works had most often been translated using conventional wording, Kenny inferred that these translations had generally been subjected to "lexical" normalization (1999b: 192-202).

Bearing in mind the theory that there are "translation universals" or recurrent features of translation, we hypothesized that our translated corpora are also

"lexically" normalized. We predicted that in the present study, there would be fewer unattested words (coinages) in our translated texts than in our non-translated texts. Like Kenny, our basis of comparison is the coined word. However, to test our version of the hypothesis of "lexical" normalization in non-literary texts, we adapted Kenny's measure instead of replicating it. For the purposes of the present study, we searched for all unattested words that could be found in each of the four (non-literary) translated and non-translated corpora, whereas in Kenny (1999b), a small sample (out of the proliferation of creative coinages available) was selected from the (highly literary) STs for comparison with the corresponding literary TTs.

Our broader scope of searching was necessitated by the fact that the texts included in the present study are non-literary, being taken from institutional rather than artistic sources. The texts included in our corpora were in all probability not written with high literary aims (i.e. with the intention of being "creative"), and it is thus not surprising that they contain comparatively few unattested words. Additionally, the languages included in our corpora probably had an effect on the number of coinages available for investigation in the present study, as can be seen in the results for normalization (Section 4.1.2). The present study does not include German, a language in which longer compound nouns may more commonly be formed out of other nouns than is probably the case in English and French.

Presented below is the method followed to test our hypothesis of normalization. We prepared a list of possible coinages from among the lowest-frequency words in each corpus.[123] Microsoft Office Word spellchecker systematically pointed out forms that were not in its database of recognized words,

[123] That is, from among the words that had occurred only once, twice or three times in a sub-corpus.

as recommended in Bowker and Pearson (2002: 213-214). Deleted from this list of possible coinages were all proper names, as well as any words found to be attested in Canadian reference works,[124] in TEXTUM,[125] or on the World Wide Web.[126] The following types of words were deleted from the list of possible coinages:

Abbreviations (e.g. "cont'd for "continued").

Commonly known and recognizable words (eating, halage).

Compounds—hyphenated or not—that were attested (e.g. baby-boom, biodiversité).

Errors, whether grammatical or typographical (e.g. "the rock *showned* no hardening"; "à court *erme*").

Foreign words whose status was typographically indicated (e.g. that were italicized, or that appeared between quotation marks).

Institutionalized acronyms (e.g. NATO, TCRPAP).

Letters designating items in an ordered list (e.g. "A-N" designating "items A through N").

Numerals of any kind.

Ordinals spelled out in letters (third, troisième).

Parts of email addresses (e.g. "climatechange," in <climatechange@gc.ca>).

Prefixes followed by hyphens (e.g. *self-*) that had been erroneously counted as single words by the software due to the presence of a space after the hyphen.

Second-level domain names (e.g. "worldweb" in www.worldweb.com and "SEESTANLEYPARK" in <www.SEESTANLEYPARK.com>).

---

[124] English: *Canadian Oxford Dictionary* (1998), *Gage Canadian Dictionary* (1983), *Canadian Style* (1997), *Canadian Writer's Reference* (2001), *Editing Canadian English* (2000), and the English references included in *Termium-Plus* (Government of Canada). French: *Multidictionnaire de la langue française* (1997), *Dictionnaire du français Plus* (1988), *Robert : Dictionnaire québécois d'aujourd'hui (1992), Dictionnaire universel francophone en ligne* (accessed 2003), *Grammaire BEPP* (Lemire 2001), *Nouveau Grand dictionnaire terminologique* (Gouvernement du Québec 2002b), *Banque de dépannage linguistique* (Gouvernement du Québec 2002c), and the French references included in *Termium-Plus* (Government of Canada).

[125] The Corpus TEXTUM was gathered between 1988 and 1996, as part of the *Bilingual Canadian Dictionary* project headed by Dr. Roda P. Roberts. Totalling 300.3 million words (English: 210 million; French: 100.3 million), the on-line corpus contains mainly Canadian-sourced texts, although smaller non-Canadian sub-corpora (English: 41.8 million, French 22 million) are included for comparison. The larger sub-corpora are mainly sourced from prominent local Canadian newspapers (English: the *Ottawa Citizen*, the *Vancouver Sun*, the *Montreal Gazette*, the *Calgary Herald*; French: *Le Droit, Le Devoir, La Presse, Le Soleil* and *L'Actualité*), while the smaller sub-corpora are taken from technical, scientific, and literary sources.

[126] As a last resort before placing a word on our list of (possibly transient) coinages, we checked among Web pages with Canadian domain names, conducting an advanced *Google* Web search specifying the country code ".ca."

Technical terms (identified by searching in *Termium*).

Once this process of elimination was complete, there remained a number of words that were retained on the assumption that they might be "fresh" or "transient" coinages.[127] Following the approach taken in Adams (1973), we then analyzed these (presumably ephemeral) coinages to determine which word-formation strategies had produced them.

To make sure that the coinages used to test the hypothesis of normalization in the translated corpora had not all occurred in one text, we also used the Dispersion Plot feature of the WordSmith Tools Concordancer.

A number of unattested words were found (e.g. "out-source," "upswell," "rockmass," "pre-prototype," "tarif-lettres, "facteur-charge," "collier ras-de-cou"). These are listed, and their contexts given, in Table 6.

4.1.2 Normalization: Results

There were 77 unattested coinages in total (i.e. in all four corpora). Most of these were found in the English corpus.[128] Each translated corpus (ETT and FTT) contained considerably fewer coinages than each corresponding non-translated corpus (EST and FST), as seen in Table 5 below. Both translated corpora contained nearly thirty percent fewer coinages. While the number of coinages differed by

---

[127] The term "fresh coinages" is Fowler's (1965: 253). Note that we are making no claim that the words retained are permanent neologisms in the French and English languages. Unless evidence turns up to the contrary in future study, we will assume these attempted coinages to have been "transient," as Valerie Adams describes the presumably ephemeral examples from her own corpus (1973: *viii*).

[128] There were considerably more English-language attempted coinages. Out of the 77 total attempted coinages identified and retained in the four corpora, 55 were found in the non-translated and translated English corpora, compared to 22 in the French corpora. As mentioned previously, we make no claim as to the permanence of these "coinages," which we will consider to be transient unless evidence to the contrary appears in future.

language, the proportional difference (between translated and non-translated texts) was exactly the same in English and French.

Table 5
Quantity and distribution of transient coinages

(English)	(55 Total)
EST	(35) 63.6%
ETT	(20) 36.4%
TT percentage difference	**-27.3%**
(French)	(22 Total)
FST	(14) 63.6%
FTT	(8) 36.4%
TT percentage difference	**-27.3%**

Many of the unattested words had been produced following familiar word-formation strategies (e.g. adjective compounds formed out of a *noun + -verb +ed*, such as *knowledge-related* or *scenario-based*). Both English and French had a variety of unattested derivations and compounds, most of which could be categorized as either noun compounds or adjective compounds.[129] In English there were also a small number of unattested verb compounds and adverb compounds, as well as unattested blends, clippings, "zero derivations,"[130] and a loan word.[131] In translated English only, there were two untraceable coinages, whose seemingly

[129] Valerie Adams (1973: 147) notes that "groups of words associated in form and meaning may grow up in a language," such as the "group-forming" pattern with the suffix –nik: (beatnik, folknik, straightnik, etc.). We found one group-forming pattern of compounds with the clipping *e-** (e-bidding, e-Catalogue, e-revolution, e-standards, e-tail).
[130] We use the term "zero derivation" in the present study, whereas "zero-suffix derivation" is the term used by Valerie Adams (1973: 16). It is also called "conversion" in Crystal (1997: 92).
[131] We are referring to "loan word" as defined in Crystal (1997: 227): "where both form and meaning are borrowed."

idiosyncratic pattern of word-formation made them difficult to classify.[132] Below we will discuss the distribution of the transient coinages found in each language.

Transient Coinages Found in the English Corpus

Thirty-five transient coinages are found in the English non-translated corpus. All are well-formed and transparent, that is, they are formed following familiar patterns. None appears to have been formed following any unusually "creative" strategy. Out of the total coinages in ETT, eighty-five percent (30) are compounds; the remaining fifteen percent (5) are derivatives. One of the compounds, "out-source," can be categorized as a "zero-derivation" (a type of compound) because "source," usually a noun, becomes a verb in combination with the prepositional particle "out." It is interesting to note that this coinage, which was not attested at the time this corpus was gathered, has since come into common use.

The 30 compound coinages found in the English non-translated corpus are listed below.

EST Compound Coinages

1. commercial-based
2. cross-government
3. e-bidding
4. e-Catalogue
5. e-revolution
6. e-standards
7. e-tail
8. endowment-building
9. in-press

[132] By "unidentifiable" coinage, we mean unattested words that do not seem to follow a pattern of word formation previously described in the literature, and that have apparently been coined "from scratch" (Heift 2003: 36). These are called "pure" coinages by Heift (*ibid.*).

10. IT-procurement
11. knowledge-related
12. mailflows
13. mediacard
14. media-rich
15. ocean-to-ocean
16. outmigration
17. out-source
18. over-capitalized
19. over-power
20. partner-operate
21. payment-related
22. revenue run-rate
23. process-management tool
24. scenario-based
25. store-level
26. sulphur-yellow
27. techno-minimalism
28. upswell
29. wave-washed
30. weight-based

EST Derivative Coinages

1. ex-urban
2. interarts
3. non-monetarily
4. pre-dialogue
5. predoctoral

In the translated English corpus, a total of 20 transient coinages were found. As in the non-translated English corpus, most of these had been formed following familiar and recognizable strategies of word-formation. Eighty percent (16) of these coinages were compounds, while ten percent (2) were derivatives. Two coinages were retained that had been formed following a strategy which appeared untraceable (and that might be interpreted to be "poorly formed" in the sense that they do not follow a recognizable pattern and that their meaning is not clear).

The sixteen compound coinages found in the English translated corpus are listed below.

ETT Compound Coinages

1. above-specified
2. apartment-studio
3. crewleader
4. council-wide
5. culturally-diverse
6. downstructure
7. drillhole
8. geocamera
9. geotomographic
10. poetry-on-the-bus
11. producer-artist
12. quartz-tourmaline
13. rockmass
14. script-clerk
15. sculptor-molder
16. youth-related

The two recognizable derivative coinages found in the English translated corpus are listed below.

ETT Derivative Coinages

1. accountably
2. pre-prototype

The two English translated coinages whose word-formation strategy was considered untraceable are listed below.

ETT "Unidentifiable" Coinages

1. alterned
2. deviatoric

We may speculate as to the strategy followed in forming the latter two ETT transient coinages: "alterned" is possibly a blend, or some sort of merging, of "alternated" and "altered"; "deviatoric" may be a "calque" (Crystal 1997: 51) of the ST word *déviatorique*, which is itself an unidentifiable coinage.[133] See Table 6 for the larger context in which these coinages appear.

A summary of the distribution of the English transient coinages according to corpus and word-formation strategy can be made as follows. There are 35 total coinages (30 compounds and five derivatives) in the non-translated corpus, while in

[133] The presence of these words among the selected transient coinages suggests an interesting topic of future study, in which the quality of an attempted coinage might be ranked along a scale ranging from "normative" to "non-normative" (rather than "creative"). See Section 6.5.2. It is tempting to interpret these particular attempted coinages as calques, as in Crystal (1997: 51): "a type of borrowing where the morphemic constituents of the borrowed word... are translated item by item into equivalent morphemes;...[a]...loan translation."

the translated corpus, there are 20 total coinages (16 compounds, two derivatives, and two "unidentifiable" word-formations).

There are thus nearly thirty percent fewer transient coinages in the translated corpus. Despite the fact that the translated and non-translated corpora have a similar total number of running words, they do not have a similar total number of coinages, as might be expected if the features of translated texts were indistinguishable from those of written texts (our "null hypothesis"). Furthermore, in ETT, the meaning of two of the transient coinages is not clear, while no such cases are found in EST.

Coinages Found in the French Corpus

Fourteen transient coinages were found in the French non-translated corpus (FST). Most of these follow familiar word-formation strategies. Almost all (thirteen of them) are compounds. The (typical) loan compound noun s*teady-cam* names a person (a specialized camera operator) in French, whereas in English, it names an object (a type of camera). We found the hyphenated version of this word to be unattested in French, and therefore have counted it among the transient coinages. One of the FST compounds, *vidéowall*, may be loan word, that is, a gallicization (using the *accent aigu*) of videowall, the English compound. There is one "unidentifiable" coinage, *déviatorique,* the meaning of which is not clear, and word-formation strategy of which is apparently untraceable.[134] With the exception of the last one listed (*déviatorique*), all are well-formed compounds (that is, they are based on familiar patterns of word formation). Given the apparent similarity of

[134] As mentioned above, its anglicized version, "deviatoric," appears to have been borrowed or "calqued" (Crystal 1997) into ETT.

the topic covered in FST coinages 1-6, we checked their distribution in the corpus, and found that while many were found in two different texts by one institutional "author" (CAPPRT-TCRPAP), there were also occurrences in an Environment Canada text. The 14 transient coinages found in the French non-translated corpus are listed below.

FST Coinages

1. assistant-coiffeur (compound)
2. assistant-menuisier (compound)
3. assistant-perruquier (compound)
4. assistant-prothésiste (compound)
5. assistant-régisseur (compound)
6. assistant-scripte (compound)
7. facteur-charge (compound)
8. géocaméra (compound)
9. négociation-cinéma (compound)
10. quasi-horizontal (compound)
11. sculpteur-mouleur (compound)
12. steady-cam (compound)
13. vidéowall (compound)
14. déviatorique (unidentifiable)

In the French translated corpus (FTT), eight transient coinages were found. The meaning of all of them was clear, and all followed recognizable word formation strategies. Of these, seven (almost all of them) were compounds, and one was a derivative. The eight transient coinages found in the French non-translated corpus are listed below.

FTT Compound and Derivative Coinages

1. artistes-participants (compound)
2. collier ras-de-cou (compound)
3. écrivain-philosophe (compound)
4. hydroclimatologique (compound)
5. intervenants-ministères (compound)
6. sculpture-fontaine (compound)
7. tarif-lettres (compound)
8. méga-initiatives (derivative)

The distribution of the French transient coinages according to corpus and word-formation strategy can thus be summarized as follows. In the French non-translated corpus (FST), there are 14 total coinages, of which 13 are classic compounds, while one is an "unidentifiable" word-formation. In the French translated corpus (FTT), there are a total of eight coinages (seven compounds and one derivative), all of which have transparent meaning and recognizable word-formation. There are thus nearly 30% fewer transient coinages in the French translated corpus. Despite the fact that the translated and non-translated French corpora have a similar total number of running words, they do not have a similar total number of transient coinages, as might be expected if the features of translated texts were indistinguishable from those of written texts (our "null hypothesis"). Furthermore, there is one unidentifiable, "opaque" coinage in FST, while no such cases are found in FTT.

4.1.3 Normalization: Test of Significance and Interpretation of Results

We used the Z-test to determine the statistical significance of our findings (see Appendix X). For both corpora of translated texts, the proportional difference in the number of coinages was highly significant in English at a 97% confidence level (p-value 0.025858), and was acceptably significant in French at a confidence level of approximately 90% (p-value 0.113897). We were thus able to accept the alternative (research) hypothesis. We interpret these findings to indicate that the results noted below were unlikely to be due to chance.

Compared to our non-translated corpora, both of our translated corpora have proportions of transient coinages to running words that are distinctly lower, providing evidence in support of our research hypothesis. This tendency toward conservatism was not absolute. As predicted by Toury (1991: 87 and 1995: 31; see Section 2.1.1), unique "textemes"—words coined for the purpose of translating a single text—are indeed present in both of our translated corpora.

However, it is striking that in each language of translation there are exactly 27.3% fewer total coinages, despite the fact that the translated and non-translated corpora all have a similar total number of words. The proportional difference between translated and non-translated texts is thus the same, regardless of the language of the text. This appears to offer support for our general hypothesis that translated texts will differ from non-translated texts in similar ways, no matter what the language. Furthermore, it supports our specific hypothesis that there will be fewer unattested words (transient coinages) in translated texts, regardless of the

language. The fact that the translated texts have significantly fewer total coinages in each language is therefore tentatively interpreted as possible evidence that greater "conservatism" is a norm in translation, and that translated texts may be normalized regardless of language.[135]

The results also point to possible differences in linguistic norms. Significantly more transient coinages were found in English than in French. This suggests that the overall acceptability of coining words differs in the two languages studied. Table 6 lists the transient coinages and quotes the context in which they were embedded.

Table 6
Coinages in context

English non-translated corpus (EST)	
Coinage	Context
commercial-based	CPIL will assist you to develop and implement a **commercial-based** wholesaler to retailer process, customized for your special requirements
cross-government	The CRP is being designed and implemented by an Interdepartmental Working Group (IWG) composed of representatives from 26 federal departments and agencies. Minister Vanclief, in his capacity as Minister Coordinating Rural Affairs, and the Rural Secretariat, provide leadership and coordination for this **cross-government** approach.
e-bidding	The Electronic Supply Chain Program will provide clients with an on-line marketplace, **e-bidding** services and electronic payments and settlements.
e-Catalogue	Future Directions Canadian General Standards Board will implement: an **e-Catalogue** of standards; the sale of and receipt of payment for e-standards documentation; an on-line list of certified products and services; ISO 9000 and ISO 14001 application information and forms; and a standards development work program.
e-revolution	Globalization, commercialization, technological change, the Internet and the **e-revolution**: around the world, postal organizations are facing unprecedented challenges and opportunities.
e-standards	Future Directions Canadian General Standards Board will implement: an e-Catalogue of standards; the sale of and receipt of payment for **e-standards** documentation; an on-line list of certified products and services; ISO 9000 and ISO 14001 application information and forms; and a standards development work program.
e-tail	How the RSC Works For Your Post

[135] Baker (1996: 183) originally proposed "conservatism" as an alternate term for normalization.

	The RSC "order to cash" total system supports order taking and fulfillment of all inventory for your retail channels. [...] Management of retail, wholesale and **e-tail** sales, and product distribution to corporate offices
endowment-building	The study, to be completed by the fall, will examine financing, planning and development, **endowment-building** and stabilization over a 20-year period.
ex-urban	Recognition of differences is also critical for all population groups that reside in rural Canada -- Aboriginal peoples, the elderly, farm families and **ex-urban** residents.
in-press	Leanne Small and Tony Fuller have an article, currently **in-press**, addressing the use of asset mapping as a tool in the examination of sustainable livelihoods in rural communities.
interarts	Three years ago, the council established an **interarts** office to work with the increasing number of artists whose work does not fit into one traditional medium.
IT-procurement	Management Services' mandate is to provide leadership, direction and high-quality support services to the CIO and other IMTB divisions, with particular reference to [...] contract and **IT-procurement** support; branch administrative, committee and technical support[...]
knowledge-related	The Branch currently has five divisions: Client Services and Partner Relations Division (CSPR)[.] CSPR helps IMTB meet client and partner needs for information- and **knowledge-related** solutions.
mailflows	"**Mailflows**" link the system to the physical flow of mail through the plant.
mediacard	Accomplishments include: [...] Direct Marketing: [...] Initiated and implemented the "**Mediacard**" product.
media-rich	Future Directions: Respond to increased demand for electronic service delivery and **media-rich** services and applications.
non-monetarily	Governments should recognize that in some instances the community cannot commit dollars and cents because of increased demands on its resources; however, it can contribute **non-monetarily** by building relationships.
ocean-to-ocean	Northern Services has established linked deliveries across Canada's North to create a network of **ocean-to-ocean** mail service, one that increasingly follows the shortest-distance rule of other networks.
outmigration	**Outmigration** A number of groups discussed the fact that many young people in rural areas are leaving in order to find jobs or go to school.
out-source	Director of Mail Management Services for Canada Post, managing **out-sourced** mail operations, distribution and transportation for major corporations in Canada; [...]
over-capitalized	Participants suggested that many resource based industries have been **over-capitalized** and, as a result, many rural communities are now in debt.
over-power	Within some partnerships, the interests of some groups **over-power** the interests of others; people do not always want to share responsibilities, power and control.
partner-operate	For example; almost a decade ago, we pioneered partnerships between Canada Post and Aboriginal bands, forging what have since become 158 **partner-operated** post offices.
payment-related	Pensioners will be able to access their own personal and **payment-related** information and complete discretionary on-line transactions.
pre-dialogue	**Pre-Dialogue** Focus Groups In the Spring of 1998, the rural issues and the Rural Dialogue process were tested and validated by focus groups held in five rural communities across the country.

predoctoral	She received her Bachelor of Music degree from the University of Toronto in 1999 and her Masters degree in concert repertoire and oratorio from the Hochschule für Musik Augsberg in Germany, where she is currently pursuing **predoctoral** studies under internationally-known Canadian soprano Edith Wiens.
run-rate	CGI's annualized revenue **run-rate** totals US$1.3 billion (CDN$2.1 billion).
process-management tool	IBS is responsible for the management of major corporate IM/IT projects. This includes ongoing development of the SAP **process-management tool**, and its integration with Agency needs and business practices.
scenario-based	Employees and pensioners will have the capability to initiate and amend voluntary deductions on-line and will have access to **scenario-based** pay and pension calculators.
store-level	The Key Benefits of the ROSS Retail Point of Sale System [...]Improves speed, accuracy and image of retail counter transactions [...]Improves **store-level** inventory management for fewer stock-outs and "smart" allocation of products and supplies
sulphur-yellow	**Sulphur-yellow** in colour, and slightly feathered with red, the 'Monte Carlo' is an early-flowering, short-stemmed and double-flowered tulip.
techno-minimalism	Many of its musicians are heavily into technology (electro-acoustic music is now 20 years old) and younger artists are experimenting with **techno-minimalism**, transforming works through digital layering into ambient pieces that last 60-70 minutes.
upswell	From across Canada, from political and religious leaders and a mass of organizations ranging from the YMCA to the United Farm Women of Alberta, there came an **upswell** of demand for a national broadcasting policy.
wave-washed	Photographer Terry James provided an image that features the **wave-washed** granite boulders and famous lighthouse, which also houses a post office.
weight-based	Inventory of mail volume waiting to be processed is calculated using **weight-based** conversion Factors.

English translated corpus (ETT)

Coinage	Context
above-specified	If the objection had been that the applicant had not met one or more of the **above-specified** requirements, or that its documents were forged or fraudulent, the objection would have been relevant.
accountably	Our governments have shown wisdom in creating strong cultural agencies that operate independently but **accountably**.
alterned	The holes were drilled along the vein at an angle of 55 degrees towards the top. The holes were **alterned** on two lines spaced 1.2 m (4 ft) apart. One line was drilled near the contact with the north wall, the other at the centre of the drift.
apartment-studio	The recipient has the use of an **apartment-studio** located on the Piazza Sant'Apollonia in the Trastevere district of the city.
crewleader	...set design, in particular the following positions: art co-ordinator, assistant art director, set designer, assistant set designer, set co-ordinator, set technician, set decorator, propsman specialist, props designer, propsman **crewleader**, studio propsman, location propsman, ...
council-wide	As Joanne Morrow, Director of the Arts Division, explains in the feature interview in this issue, the Board and the staff of the Council have established a **Council-wide** policy on peer assessment that outlines with clarity and rigor the rules for peer assessment practices.

culturally-diverse	Additional funding has been assigned to the Inter-Arts Office (for interdisciplinary and performance art), the Outreach program (for market development), the Aboriginal Arts Secretariat and Equity Office (for Aboriginal and **culturally-diverse** artists), the Canadian Commission for UNESCO, as well as to the administrative costs of jurying and awarding grants.
deviatoric	The magnitude of the mean and the **deviatoric** stresses at each location were also calculated.
downstructure	[...], the Board established four key priorities for Council operations: [...] managing resources rigorously and transparently and keeping administrative costs to the level achieved through **downstructuring**; and defining and applying the concept of Board governance as appropriate [...]
drillhole	The location of **drillholes** is shown on Figure 5.
geocamera	The boreholes were inspected with a **geocamera** to determine their general conditions.
geotomographic	The **geotomographic** survey
poetry-on-the-bus	Earlier this summer, thanks to the initiative of the Festival international de poésie de Trois-Rivières, the Canada Council-supported **poetry-on-the-bus** project was extended into Quebec.
pre-prototype	This project evaluated the technical concept of the **pre-prototype** in underground tests at the Experimental Mine.
producer-artist	The ... collective bargaining is conducted in conformity with the rules of public policy or private rules established by contract that govern the **producer-artist** relationship; the two must not be confused.
quartz-tourmaline	The zone contains some **quartz-tourmaline** veins.
rockmass	Rockbursts are grouped two ways: those associated to slippage ..., and those associated with high strain energies inside less fractured **rockmasses.**
script-clerk	editing and continuity, in particular the following positions: production co-ordinator, floor director - excluding dubbing directors - production assistant, [...], **script-clerk**, script assistant,
sculptor-molder	set design, in particular the following positions: [...], assistant painter, **sculptor-molder**, draughtsman, head carpenter, [...]
youth-related	In brief, the new funds have been directed primarily, through arts organizations, into **youth-related** initiatives and community outreach opportunities.

French non-translated corpus (FST)

Coinage	Context
assistant-coiffeur	Le regroupement a demandé à être accrédité pour représenter au Québec un secteur qui comprend :[...] 2) conception de costumes, coiffures et maquillages, notamment dans les fonctions suivantes : ... coiffeur, **assistant-coiffeur**, perruquier, assistant-perruquier, ...
assistant-menuisier	[2] Le regroupement a demandé à être accrédité pour représenter au Québec un secteur qui comprend :[...] (3) scénographie, notamment dans les fonctions suivantes : [...]chef menuisier, menuisier, **assistant-menuisier** [...]
assistant-perruquier	[2] Le regroupement a demandé à être accrédité pour représenter au Québec un secteur qui comprend :[...] 2) conception de costumes, coiffures et maquillages, notamment dans les fonctions suivantes : [...] perruquier, **assistant-perruquier**, préposé aux rallonges capillaires [...]
assistant-prothésiste	[2] Le regroupement a demandé à être accrédité pour représenter au Québec un secteur qui comprend :[...] 2) conception de costumes, coiffures et maquillages, notamment dans les fonctions suivantes : [...] prothésiste, **assistant-prothésiste**, [...]

assistant-régisseur	Les activités visées incluent : (4) montage et enchaînement, notamment dans les fonctions suivantes : … régisseur logistique, **assistant-régisseur** logistique, scripte, assistant-scripte, […]
assistant-scripte	Les activités visées incluent : (4) montage et enchaînement, notamment dans les fonctions suivantes : […] scripte, **assistant-scripte**, […]
déviatorique	On a aussi calculé la grandeur de la contrainte moyenne et **déviatorique** en chaque point de mesure.
facteur-charge	Le **facteur-charge** moyen s'établit à 0.2 kg/m3 ou 0.15 lb/t.
géocaméra	L'inspection à la **géocaméra** a révélé la présence de nombreux bris à la paroi des forages …
négociation-cinéma	Selon la requérante, lorsque cette **négociation-cinéma** sera complétée, les seuls producteurs d'importance qui n'auront pas d'entente avec elle seront …
quasi-horizontal	La position des émetteurs et des récepteurs définissait ainsi un plan de mesure **quasi-horizontal** situé à une distance d'environ …
sculpteur-mouleur	(3) scénographie, notamment dans les fonctions suivantes : coordonnateur artistique, […]peintre scénique, assistant peintre, **sculpteur-mouleur**, dessinateur,
steady-cam	(1) conception de l'image, de l'éclairage et du son, notamment dans les fonctions suivantes : assistant-réalisateur, premier assistant à la réalisation, […], cadreur, caméraman (incluant **steady-cam**,[…]
vidéowall	(4) montage et enchaînement, notamment dans les fonctions suivantes : […] opérateur de télésouffleur, vidéographe en régie, projectionniste vidéo (y compris écran géant, **vidéowall**), chef machiniste vidéo …
French translated corpus (FTT)	
Coinage	Context
artistes-participants	La firme a choisi au hasard les **artistes-participants** dans la base de données du Conseil sur les bénéficiaires de subventions, ….
collier ras-de-cou	Les campeurs ont appris, entre autres, à jouer du tambour, à chanter, … à confectionner des capteurs de rêves, **des colliers ras-de-cou** et à faire de la broderie perlée ….
écrivain-philosophe	L'auteur de science-fiction, Sir Arthur Clarke, …, le directeur artistique du Cirque du Soleil, Franco Dragone, et **l'écrivain-philosophe** Mark Kingwell faisaient partie des invités d'honneur.
hydroclimatologique	Dans le cadre de ce projet, l'Université du Manitoba élaborera un modèle d'évaluation **hydroclimatologique** dynamique ….
intervenants-ministères	Comment mettriez-vous en oeuvre les suggestions dont on a discuté-et qui doit participer à cette mise en oeuvre? (Quels **intervenants-ministères**/organismes gouvernementaux, …, secteurs précis)?
méga-initiatives	Les projets ponctuels de création d'emplois et les « **méga-initiatives** » de courte durée ne sont pas considérés comme la meilleure approche de la diversification économique ….
sculpture-fontaine	Il s'agit de la **sculpture-fontaine** intitulée Embâcle, l'oeuvre qui figure sur le timbre.
tarif-lettres	L'une des entreprises postales les plus modernes, …, la Société canadienne des postes : […] Possède un **tarif-lettres** de base qui se classe au deuxième rang …

4.2 Explicitation

As discussed in the Literature Review, explicitation (principally measured as the frequency of inclusion of optional *that* with a small number of reporting verbs) was previously investigated in English literary texts by Olohan and Baker (2000), and subsequently by Olohan (2001, 2002). To add complementary information to the knowledge of explicitation gleaned in the above previous studies, we also investigated explicitation in an additional language(French as well as English) and using non-literary texts. In each of the two languages (English and French) included in the present study, we attempted to observe whether the translated corpora contained higher frequencies of optional syntactic elements, rendering the syntax of the translated texts more explicit than that of the non-translated texts, as hypothesized in the introductory chapter.

4.2.1 Explicitation: Measures

Our approach to measuring explicitation was inspired by that of the above previous studies. To test the hypothesis of explicitation as a recurrent feature of translation, we compared the frequencies of redundant syntactic elements in the English and French translated and non-translated texts. This was done by sorting and analyzing the patterns of occurrence of *that* and *which* in the English corpus, and of *le* and *ne* in the French corpus. Each of these measures is described below. One of the English measures is an expanded version of a measure used in Olohan and Baker (2000), and in Olohan (2001; 2002); the others are new measures developed for the specific purposes of our study.

English Measures: Literature Review

In the English corpora, we investigated explicitation using two measures:

1) the comparative frequencies of optional *that* versus "zero" (i.e. no) *that* after reporting verbs
2) the comparative frequencies of optional *that* or *which* as the object of a defining (restrictive) clause.

In what follows, we will briefly review the pertinent literature on these optional syntactic elements in English, after which we will describe the measures themselves.

In the literature, certain "optional" syntactic elements are noted to exist among the markers of explicit reference in English. According to Biber, the primary syntactic markers of explicit reference are relative pronouns (especially *who*, *whose*, *whom*, and *which*) and "pied piping constructions" (preposition + *whom/which*), both of which can be considered "devices for the explicit, elaborated identification of referents in a text" (Biber 1988: 110).[136] Such elements have been described by a number of other scholars. Generally, a syntactic element is considered to be "optional" when it appears to be possible to omit it without changing the basic syntactic structure of the sentence.

Leech and Svartvik (1975) categorize the "zero" or asyndetic connective as a variant of the relative pronoun.[137] The asyndeton (that is, the "zero" element)

---

[136] The relative pronoun is furthermore a marker of "on-line informational elaboration" according to Biber (1988: 114; 154-160).

[137] It may be misleading to consider *that* to have been "dropped" in cases of asyndetic linking. There is some question as to which is the "original" object clause link. It could be either one, as both are apparently present in the Old English texts, and

is common in restrictive (also called "determining") clauses: "All the relative pronouns can be used in restrictive relative clauses: *who* (*whom*, *whose*), *which* and, particularly, *that* and the *zero* relative" (Leech and Svartvik 1975: 287).

Brinton (2000: 216; 220) states that either asyndetic ("zero") or syndetic (optional) linking may be used where the connective is the clause object, e.g. "Do you know the boy (*Ø/that/who/whom*) we met?", or the complement of a preposition, e.g. "This is the house (*Ø/that/which/*) we wrote to you about." According to Brinton (*ibid.*), it is generally acceptable to "delete the subordinating conjunction *that"*(i.e., to use asyndetic or "zero" linking) when the clause it introduces functions in one of three ways:

1) as a direct object (*I know Ø coffee grows in Brazil*),
2) as a direct object after an indirect object (*He told his mother Ø coffee grows in Brazil*), or
3) as a "complement of A" (*He is certain Ø coffee grows in Brazil*).[138]

Hunston and Francis (2000: 61) note that where either *that* or *which* comes before the subject of the verb, and where it is the object or complement of that verb in the relative clause (e.g. Most of the people that I met were academics; He tapped the file which Fox had brought in), "the relative pronoun is often not used" (e.g. The people Ø I met appeared capable).

In a diachronic study of British newspaper editorials, Westin (2002) compared the degrees of "referential explicitness" seen in various types of

---

the asyndetic link may be "at least as old as [the syndetic] *that*" (Rissanen 1991: 275). However, according to Rissanen, asyndetic ("zero-*that*") linking is common starting only with extant English texts from the 1600s.

[138] Three further conditions are listed, although Brinton notes (2000: 220) that the acceptibility of these must be left up to the judgement of the speaker: when the clause functions: 1) as a subject complement (*My understanding is Ø coffee grows in Brazil*), 2) as a "complement of N" (*His claim Ø coffee grows in Brazil is correct*), 3) as an extraposed subject (*It is well known to all Ø coffee grows in Brazil*).

relative clause structure, including relative clauses introduced by "*wh*-pronouns" (a term Westin borrows from Quirk *et al.* 1985: 77; 366) and by *that*.[139] Elsness (1984: 521), who studied object clause connectives in a very small corpus of American English, found that objects of defining clauses were "always" asyndetically (i.e. "zero") linked, in a small, manually-perused corpus containing a variety of types of non-translated texts.[140]

Two questions tend to arise in the literature concerning asyndetic ("zero") linking with relative clauses in English:

1) is "zero" linking stylistically marked as less formal?

2) are there conditions that effectively bar its use, and if so, are "optional" and "zero" linking equivalent?

Below we discuss each of these questions in turn.

In a number of reference manuals, it is suggested that the use of the asyndeton ("zero" linking) somehow imparts a difference in register: "more informal discourse will tend to have a preference for zero," and the conjunction *that* is "frequently omitted" in "less formal indirect speech," claim Quirk *et al.* (1985: 1252; 1026; 999). The "zero" connective is believed to make the text both less formal and less explicit, although there appears to be no rule describing conditions under which the writer is obliged to use it (e.g. Quirk *et al* 1985: 23-24, 845-846, 900; Leech and Svartvik 1975: 249; Collins COBUILD *English Usage* 1992: 708-709; Swan 1995: 588-589). The difference in register is generally attributed to a closer stylistic association of the "zero" asyndeton with

[139] Westin excluded the asyndetic "zero" variant construction from her study only because of the manual searching it would have required (2002: 132).

[140] "OBCLs [objects of clauses] in which the subject is raised into the matrix clause ('... the man [who] the President thought was responsible' are without exception ZEROs."

speaking, and of syndetic linking (i.e., inclusion of the optional element) with writing.

An assumption is therefore made that speech tends to be both less formal in register and less explicit in its syntax, and that writing is both more formal and more syntactically explicit. This distinction does not always stand up to scrutiny, as Bartsch (1987: 10; 12) points out: while there are likely to be differences in the degree of syntactic explicitness between spoken and written language, these differences are due not to the medium itself (speech or writing), but to the degree of contextual ambiguity:

> Spoken language is syntactically less restricted in general ... Written language *per se* is largely independent of the situations of writing and reading. It has to make explicit, by description, information which in daily speech can be available in the situation. Besides the use of more lexical items, this requires a large amount of socially controllable syntactic construction. [However,] there are also situations in which spoken language requires strong syntactic restrictions. Formal speech is an example, as in lectures about involved matters. Strict syntactic form is also necessary in stories and songs that report history, not only for facilitating recollection, but for keeping the facts straight about events that happened long ago and are not recoverable independently. The same precision is required in formulating predictions or plans about the future" where ambiguity must be avoided.
>
> (Bartsch 1987: 10; 12)

Storms (1966: 262-269) attributes inclusion of optional (syndetic) *that* to a desire on the part of the speaker to be "less personal, less familiar" and more "objective, factual, formal, official." He divides the "lexical items (the verbs, nouns, and adjectives) that take the *that* clause" into two main groups:

1) "those that have a modal function" and which reflect the speaker's "attitude as regards the 'truth'"; and

2) "those that have an affective or expressive function" (266).

Elsness maintains that asyndetic ("zero") linking is more common in texts of a "less formal" type, and that this is "amply confirmed" in his study (1984: 531).[141] Elsness (1984: 525-526) also found that "the choice of connective [was] influenced by the person of both the matrix and the OBCL subjects: 1st and 2nd person subjects make for higher proportions of ZEROs than 3rd person subjects." In other words, according to Elsness, asyndetic ("zero") linking is more likely to appear in texts that include the personal form of address than in texts with the (presumably less personal and therefore potentially more formal) third person singular and plural.

According to Bernstein *et al.* (1999: 217), the use of the conjunction *that* in a reporting clause is obligatory, rather than optional, under three conditions:

a) when a "time element" intervenes between the reporting verb and the clause (e.g. "She said yesterday that production was up");

b) when a long passage "delays" the reporting verb (e.g. "He said without pausing to furrow his beetle brow that he was sorry");

c) when two or more clauses "attach to" the reporting verb (e.g. Tiger griped that his game was way off, that his stroke had gone wonky, and that he'd rather play croquet from now on).[142]

In these examples, the use of the conjunction *that* before the reporting clauses does not appear, intuitively, to be optional, as it would be in sentences such as "He said (that) he was sorry." The (invented) examples seem to require disambiguation, much like sentences in which an element can be assigned one or

---

[141] This assertion is made on the basis of Elsness' findings that texts categorized as popular press and fiction in his corpus have asyndetic linking ("ZEROS") which are "slightly more frequent than THATs," and that "ZEROS are particularly rare" in texts categorized as scientific writings in his corpus (Elsness 1984: 531).

[142] Bernstein *et al.* do not offer corpus-based data in support of their claim, which appears to be based on logical introspection. The examples given are our own.

more grammatical categories, leading the reader "down the garden path."[143] Décsy (1987) would probably disagree that "zero" and "non-zero" linking are semantically equivalent variants. Every change in wording is a semantic change, she claims (Décsy 1987: 107).[144] This introspection remains to be tested on authentic texts, however.

If, *contra* Bernstein *et al.* (1999), asyndetic (zero) and syndetic (optional) linking are in fact a matter of choice on the part of the speaker or writer, then a number of factors may still influence which of the two choices (asyndetic versus syndetic linking) is selected for the construction of a sentence.

Elsness (1984: 531-533) believes that style, potential ambiguity, structural complexity, weight-distributional relations, and the closeness of the clause juncture all influence the choice between *that* and the "zero" or asyndetic variant. Where an adverbial intervenes between the matrix verb and the object clause, Elsness (1984: 523; 531) found that asyndetic linking is "rare," and concludes that "the proportion of THATs [sic] increases markedly" where an intervening adverbial makes the sentence potentially ambiguous. Elsness also feels that there is a "tendency for the proportion of THATs [sic] to increase with the length [in number of words] of the OBCL subject," again suggesting that "a motive for selecting *that* connective can be a desire to contribute to greater syntactic clarity in cases of structural complexity just after the matrix/OBCL boundary" (Elsness

[143] "A *garden path* sentence contains an element (a lexical item, usually) that can be assigned more than one category (N, V, etc.) or representation (meaning, pronunciation, stress, intonation, etc.)" as in: "If it works, run it, if it doesn't, can it" (Dougherty *et al.*, 2001: 2). In this example, the reader can assign either of two grammatical categories to the verb *can*, interpreting it as a modal or as a transitive imperative (colloquial) verb. A disambiguating element—a word or punctuation mark, for instance—must be added to clarify the syntactic structure. Without this explicitation, the sentence will still be grammatically correct. However, for its syntax to be clear and its meaning understandable, disambiguating explicitation is required, not optional. See Gibson (1991) for a catalogue of ambiguities in English leading to difficulty in sentence processing.

[144] "There is no such thing as free word order of the parts of the sentence. Every word order change in the sequence of the parts of the sentence changes the meaning of the statement (at least slightly)."

1984: 527). The lexical "load" of the subject influenced the type of connective selected, Elsness found: "constructions which are predominantly THAT [sic] tend to have OBCL subjects introduced by lexical or other comparatively heavy words: nouns, adjectives, numerals, and indefinite determiners and pronouns" (1984: 529).

Rissanen notes that while the word order differences that arise from the use of the (asyndetic/syndetic) variants may bring about slight semantic shifts, they have nonetheless been shown to have such a "high degree of synonymy" that they can for all intents and purposes be considered equivalent (1991: 273-274). Rissanen (1991: 276-277) also notes that although there are a number of "zero favouring structural environments," the question of whether there are syntactic conditions that exclude the use of asyndetic linking nonetheless remains. Storms (1966: 265) likewise says that "*that* has no inherent semantic meaning but only a functional, grammatical meaning, e.g., it constitutes the grammatical means of joining two clauses. Consequently the omission of *that* does not influence the semantic relation."

The hypothesis of explicitation as a recurrent feature of translated English texts, as seen in their greater frequencies of syndetic linking (i.e. linking with an optional syntactic element instead of with "zero"), was tested empirically in Burnett (1999), in Olohan and Baker (2000), and in Olohan (2001). Each of these studies compared the frequency of use, in the **TEC*** and **BNC***, of "optional *that*" with a small number of reporting verbs. In each study, *that* was considered to be a redundant syntactic element, since asyndetic linking ("zero-*that*") was an alternative option.

The findings for each study were that "optional *that*" (the syndetic linking) occurred significantly more frequently in the translated texts than in the non-translated texts after the reporting verbs selected for study (Burnett 1999, cited in Olohan and Baker 2000: 151; Olohan and Baker 2000: 154; Olohan 2001: 425).[145] In all three studies, these findings were considered to support the hypothesis of explicitation as a recurrent feature of translation.

English Measures of Explicitation in the Present Study

As mentioned at the beginning of this section, we investigated explicitation as a possible recurrent feature of translation by comparing the frequencies, in our translated and non-translated English corpora, of two cases of optional syndetic linking:

(1) reporting *that*

(2) *that* or *which* (object of a defining clause).

Our hypothesis (see Section 1.6.2) was that there would be higher frequencies of these optional syntactic elements in the English corpus of translated texts. In what follows, we will discuss each of these measures in detail.

As discussed above, English grammar permits the speaker or writer to either omit or include the "reporting *that*" conjunction, without significantly

[145] Olohan (2001: 427) also compared the frequency of use of the relative pronoun *which*, and found it to be "considerably higher" in TEC. However, she did not sort out, from among all the instances of *which*, those cases that were the actual syndetic linking she was searching for. While she discarded "irrelevant instances" of *which* that were "sentence-initial" and "sentence- or clause-final," she did not go on to select out those cases where *which* could have been omitted but was not, that is, where "the coreferential NP is not in subject function in the relative clause" (*ibid.*). What she counted was the remaining raw frequencies of *which*, without distinguishing further among the various uses of the word in the instances found. We felt prompted to build on this part of Olohan (2001): in the present study, as will be seen below, we have distinguished cases of *which*, object of a defining clause, from among all the occurrences of *which* in our English corpora. Olohan (2001: 427-428) obtained inconclusive results with the other optional elements she studied.

altering the denotation of the clause.[146] Here is an example from our English corpus: "These countries believe *that* it is high time to start looking at both institutions" (ETT; emphasis added). This sentence is neither grammatically nor semantically altered when *that* is deleted and asyndetic linking used instead (e.g. These countries believe it is high time to start looking etc.). However, inclusion of *that* may be considered a means of disambiguating the syntax.

To compare how often "zero-*that*" and "optional *that*" were used in our English corpora, we obtained concordances of all identifiable reporting verbs and sorted every occurrence of a reporting clause into either of the two categories (i.e. "zero" or "optional *that*"). Our list of reporting verbs was compiled from those of Quirk *et al.* (1973) and Collins COBUILD (2000). There are 167 reporting verbs identified on this list (see Appendix III), and all of them were searched in the present study.[147]

We also noted all cases where a time element, long passage, or multiple-clause construction (i.e., two or more clauses) intervened between the reporting verb and its reporting clause. This made it possible to observe whether the sentence structure appeared to meet any of Berenstein, *et al.*'s conditions for obligatory syndetic linking.

Sorting all the reporting clauses into cases of asyndetic "zero" linking and syndetic "optional" linking demonstrated how often reporting clauses were

---

[146] The conjunction is also called "zero-*that*" (Quirk *et al.* 1973: 734, 832-146; Elsness 1984 *passim*; Collins COBUILD 2000: 321), "expletive *that*" e.g. Bernstein *et al.* (1999: 217), or "*that*-object" Crystal (1999: 387). Kane (1983: 735) refers to cases of asyndetic linking where *that* is omitted as "contact clauses."

[147] It should be noted that "be rumoured" (in the past tense only) is listed as a reporting verb in Collins COBUILD (2000: 315, 321, 331), with the impersonal subject "it." We have therefore included "rumoured (be)" in our list of reporting verbs.

"explicitated" (according to the present study's hypothesis) in the translated English corpus versus the non-translated English corpus.

As discussed above, English grammar permits the speaker or writer to either omit or include *that* or *which* when these words are the objects of a defining or "restrictive" clause (Collins COBUILD 2000: 364; Crystal 1997: 329). Consider, for example, "*Shells that/which* my sister collected" versus "Shells Ø my sister collected" (an example of "zero" ODC). This sentence would be neither grammatically nor semantically altered if *that* or *which* were deleted and asyndetic linking used instead; however, inclusion of *that* or *which* ODC may be considered a way of disambiguating the syntax.

In order to compare how often *that* or *which* (ODC) had been optionally included in the English corpus, we obtained a concordance of all occurrences of *that* and *which*, and retained all cases of *that* or *which* (ODC) in the English translated and non-translated texts. This demonstrated how many defining clauses were "explicitated" (according to the present study's hypothesis of explicitation) in the translated corpus versus the non-translated corpus. However, since it would have been impossible to find every single subject of a defining clause (barring an extremely lengthy manual search, which would at any rate have been vulnerable to extensive human error), we were unable to compare all instances of "zero" ODC with this explicitating element.[148]

[148] For example, to find all instances with a structure similar to Shells Ø my sister collected, a search through all nouns (e.g. *shell*) and all pronouns (e.g. *my*) would have been necessary. This would have been possible with a well-tagged corpus, but extensive further manual searching would then have been imperative, in order to find out whether there were other "*zero*" ODCs in the corpus that were constructed with other categories of words. Given the fact that there is a time limit for writing doctoral dissertations, such an extensive undertaking was considered beyond the scope of the present research.

Once all possible investigation of the above optional syntactic elements (i.e. "optional reporting *that*" and *that* or *which* ODC) had been carried out, we checked the position of each occurrence found in the Disperson Plot feature of the WordSmith Tools suite Concordancer (described in Section 3.4). This showed that the elements retained were adequately dispersed throughout that corpus, and that they had not all been used in only one text.[149]

Finally, we ran the same tests (by obtaining concordances of the same optional elements as above, i.e., "reporting *that*" and "*that/which* ODC") on the smaller non-specialized sub-corpora (see Section 3.1.1) which were used with the Readability Indices in the present study (see Chapter 5). These non-specialized sub-corpora contain little in the way of vocabulary that would be considered specialized from the point of view of the translator. We reasoned that there was a good chance that the non-specialized sub-corpora might contain fewer markers of formality than the "main" corpora (EST, ETT, FST, FTT) used in the present study, simply because specialized "fields of discourse" tend to be more grammatically complex and more generally formal (Quirk *et al.* 1985: 23-24).

These "control tests" could not give us conclusive results, since the non-specialized sub-corpora are very small (about 16,000 words each). Nonetheless, they did give us some indication of whether or not text register (i.e. the degree of formality of the various texts in our corpora) had affected our results.

---

[149] Considering the "institutional" nature of the authorship of the Government of Canada texts included in the present study, it is unlikely that any of the texts represent the writing of solely one person. As mentioned in Section 3.2, most, if not all, government texts are probably the work of more than one person, including at least one writer or translator and at least one revisor or gatekeeper.

French Measures

There are at least two cases where redundant, semantically-void elements may optionally be used in French:

(1) the "*consonne euphonique" L'* (*le*) before *on*
(2) *ne* inserted into a sentence as a semantically empty particle

These two cases are presented below, after which we discuss our method of studying them.

First, in French, *L'* (the apostrophied form of the definite article *le*), may be added to the indefinite pronoun *on*. Here are two examples from a single text in our French non-translated corpus, the first containing the optional *L'* and the second without it:

(a) Qui peut dire où *l'on* va? (CCFST10).
(b) Ainsi, *on* a créé le Secrétariat général de la Francophonie et *on* a transformé l'Agence de coopération culturelle et technique en Agence de la coopération... (CCFST10)

In example (b) above, it is clear that it would be possible to add a prothetic *L'*, altering the syntax (and sound) without changing the meaning, especially with the second "*on*" as follows: "...et *l'*on a transformé l'Agence...."

Grevisse notes that *on* was in earlier centuries considered a noun to which the apostrophied article *L'* might be attached (1990: 277.4). Diderot explains that the noun *on* was a synonym of *homme: "On* est un abrégé de *homme*; on dit *l'on* comme on dit *l'homme. On m'a dit*, c'est à dire un homme, quelqu'un m'a dit" (1751, *Tome II*: 17). The inclusion of the prothetic *L'* before *on* reinforces the syntax without changing the denotation (de Villers 1988: 736; Grevisse 1990: 132; Dubuc 1996: 102; Frontier 1997: 368-370; Wilmet 1997: 275).

Second, an "expletive" or "dummy" particle *ne* may be added to a French sentence after a number of comparative adjectives and adverbs (such as *autre, meilleur, mieux, pire*), conjunctions (such as *à moins que, avant que, de crainte que, de peur que, plutôt que*), and verbs (such as *avoir peur, craindre, douter, empêcher, éviter, nier*). An example would be: "On craint qu'il [ne] s'agisse d'un problème de plus en plus important." Although it is a homograph of the negative particle *ne*, the optional *ne "explétif"* conveys no negative meaning and is therefore semantically empty (Gougenheim 1969: 264-265; Grevisse 1990: 222-224; Hanse 1994: 577-578; Dubuc 1996: 479-481; Wilmet 1997: 519-521; Riegel *et al.* 1998: 419; Salkoff 1999: 109, 163).

There is some question whether the optional inclusion of the above two elements is in some way connected with the register of the text. A number of authorities on the subject (e.g. Grevisse 1990, Wagner and Pinchon 1977, Hanse 1999, Gouvernement du Québec 2002c) imply that using *l'on* (as opposed to *on*) imparts a more formal register to the text. However, no rule is given, in any of the references consulted, for inclusion of the optional element. This suggests that the association of register with the use of the prothetic *L'* may be subjective. For example, Grevisse (1990: Section 277.4) notes that the "consonne euphonique" may be used when the writer's sense of aesthetics demands it. Since subjectivities vary widely, it cannot be argued that this "aesthetic" distinction is the same thing as an absolute distinction in register. The literature also associates the *ne "explétif"* with formal register, but inclusion of this optional element is again left entirely to the discretion of the speaker or writer (Hanse 1999; Grevisse *ibid.*).

To count how often the optional prothetic *L'* had been included in the French corpus, we obtained concordances of all instances of *on* (including *l'on*) in the French translated and non-translated texts. We then counted how many occurrences of *l'on* there were and compared them to the number of occurrences of "zero + *on*." Only those cases where *on* stood alone as an unapostrophied element were retained and counted as "zero": for the purposes of comparison, we did not count *qu'on* or *lorsqu'on* among the cases of "zero + *on*."

To count how often the optional *ne "explétif"* had been included in the French corpus, we obtained concordances of all instances of *ne* in the French translated and non-translated texts. We eliminated all instances of "non-expletive" *ne* (used with *pas*, *jamais*, and so on), which turned out to be most of the instances in the automatically-obtained list. From the remaining concordance lines, we counted the instances of *ne "explétif."* We then compared these to the total occurrences of *ne* in each French corpus.[150]

As with the English corpus, we also verified that the French semantically-void syntactic elements retained were adequately dispersed throughout the corpora, by observing their position in the Plot feature of the WordSmith Tools suite Concordancer.

Finally, in an attempt to get an indication of whether register had affected our findings, we ran the same tests (by obtaining concordances of the same "optional" elements as above, i.e., prothetic *L'* and *ne explétif*) on the smaller

[150] It should be noted that this is not the same thing as finding all the instances of "non-*ne*" *explétif*: With this particular measure, we cannot guarantee that we are comparing all "zero" to all "optional" instances, because the "zero" instances are so numerous in such a variety of structure as to be nearly impossible to retrieve systematically.

sub-corpora of non-specialized texts that were also used with the English explicitation measures, and that were used with the Readability Indices

4.2.2 Explicitation: Results

The results for many of the tests supported the present study's hypothesis of explicitation as a recurrent feature of translation. The translated corpora tended to contain proportionally more optional elements than the non-translated corpora. However, the results for the French optional *L'on* failed to support the hypothesis. These findings are summarized in Tables 7-10, and are discussed in detail below. All occurrences counted among the results were found to be adequately dispersed throughout the overall corpora (EST, ETT, FST, FTT) of specialized and non-specialized English and French texts, as could be seen using the Dispersion Plot function of the WordSmith Tools suite Concordancer.

In each English corpus (EST and ETT), we compared the number of sentences in which optional "reporting *that*" conjunction was included after a reporting verb. Our findings are shown in Table 7.

The English translated corpus (ETT) had about 12% more inclusions of optional "reporting *that*" than did the English non-translated corpus (EST). Appendix IV shows the comparative breakdown of occurrences of "*zero*" versus optional *that* in the English corpora (EST and ETT).

We had also expected to find no difference between translated and non-translated texts in the above-listed three cases where (according to Bernstein *et al.*) *that* and *which* may be necessary (rather than optional) disambiguating elements. Our findings were not entirely as expected. Cases that might be considered to meet

the above three conditions for obligatory inclusion of reporting *that* were considerably more frequent in the translated corpus.[151] There were eight occurrences in ETT of long passages intervening between a reporting verb and its clause (appearing to satisfy "condition c" listed by Bernstein *et al.*; see Section 4.2.1). All of these eight occurrences included *that* (hypothetically obligatory here, according to Bernstein *et al.*), and all were sentences constructed with a series of reporting clauses attached to a single reporting verb.[152]

Nonetheless, the inclusion of *that* in these eight cases could not be considered to satisfy the criteria for obligatory inclusion listed in Bernstein *et al.*, since in each case it was clearly possible to omit *that*, attaching the multiple reporting clauses to the reporting verb without altering the meaning or the fundamental syntactic structure in any way.[153] In the English non-translated corpus, only one instance was found to fit the above description of constructions which, according to Bernstein *et al.*, require the inclusion of reporting *that*. This was a sentence with two clauses attached to the reporting verb *find*:

> "The study also found that: As a result of new funding to the Council beginning in 1997, total funding to individual artists increased by about 18 percent and the number of funded artists rose from approximately 1,300 each year to about 1,900" (EST, sic).

Note how, in the above example from the English non-translated corpus, the use of (clearly optional) reporting *that* has been kept to a minimum: it introduces the first reporting clause, but not the second. Whether the use of reporting *that* was

---

[151] To find these cases, we searched for instances of the reporting verbs listed in Appendix III.

[152] In the following example (from ETT) we have underlined the reporting verb and the second *that*, which introduces the second in a series of reporting clauses attached to it: "In reply, the Guilde argued that it had already been granted recognition ... to bargain with 85 percent of producers under provincial jurisdiction and that it was perfectly natural ... ."

[153] E.g..: "argued *that*: 1) it had already been granted recognition ...; 2) it was perfectly natural...."

optional or not in these cases, it was included significantly more frequently in the translated corpus, having been used in every case found there. In the the only instance found in the non-translated corpus, however, its use was minimized.

In each corpus (EST and ETT), we compared the proportional number of inclusions of optional *that* or *which* as the object of a defining or "restrictive" clause (ODC) to the total number of occurrences of *that* and *which*. Our findings are shown in Table 7. *That* ODC was included nearly one percent more often (+.8%) in the English translated corpus (ETT).[154] ETT had 10.3% more inclusions of optional *which* as the object of a defining or "restrictive" clause (ODC) than did the English non-translated corpus (EST). Appendix V compares the overall instances of *that* and *which* to the total specific instances of *that/which* ODC in each English corpus (EST and ETT).

As a last step in the investigation of explicitation in the English corpus, we replicated the above tests on the smaller non-specialized sub-corpora, to get an idea of whether formal register had played a part in our findings. Our findings with these "control" tests are summarized in Table 8, and are discussed below.

Proportionally, there were 19.2% more inclusions of reporting *that* in the translated sub-corpus. Furthermore, *that* ODC (object of a defining clause) was included 0.5% more in the translated non-specialized sub-corpus.[155] There were no instances (0%) of ODC among the 23 total occurrences of *which* in the non-translated sub-corpus of non-specialized texts, while there were four (14.8%) instances of *which* ODC out of the 27 total occurrences in the translated sub-corpus

[154] This is a number which does not seem significant in itself, but which also does not contradict the findings for the measure "*which* ODC." We note also that TT contains nearly 33% more total occurrences of *that* (see Appendix V), although this finding cannot be interpreted without further study.

[155] Since the non-specialized sub-corpora are smaller, one would expect these percentages to be proportionally smaller.

of non-specialized texts. These "control tests" suggest that register was a factor in the findings obtained with the explicitation measures used in the "main" corpora (EST and ETT).

In sum, the three explicitating elements studied in the English corpora and sub-corpora were included more often in the translated texts. These findings appear to support the research hypothesis of explicitation.

Table 7
Optional syntactic elements in English (overall corpora)

Optional element	EST	ETT	Proportional occurrence in translation
That after reporting verb	26% of total reporting verbs	38% of total reporting verbs	TT +12%
That ODC	2% of total *that*	3% of total *that*	TT +1%
Which ODC	1% of total *which*	11.3% of total *which*	TT + 10.4%
			(Mean difference TT +7.7%)

Table 8
Optional syntactic elements in English (non-specialized sub-corpora)

Optional element	EST	ETT	Proportional occurrence in translation
That after reporting verb	16.7% of total	36% of total	+19%
That ODC	3.6% of total *that*	4.1% of total *that*	+0.5%
Which ODC	0	14.8% of total *which*	+15%
			(Mean difference TT +11.5%)

The French translated corpus (FTT) had a higher proportional number of inclusions of the optional syntactic element *ne explétif,* but had a consistently lower proportional number of inclusions of *L'* with *on* (despite the fact that the raw frequency of the explicitating element *L'* before *on* was twice as high in FTT). These results are shown in Table 9.

The control tests using the smaller non-specialized sub-corpora (where markers of formal register were least likely to appear) had similar results, which are shown in Table 10. These results suggest that register had probably not affected the findings obtained from the main French corpora (FST and FTT).

We note that "*on* + zero" is six times more frequent in FTT than in FST, but that the occurrences of "*on* + zero" are nearly equal in the two non-specialized sub-corpora, suggesting that the high occurrence of "*on* + zero" found in FTT may point to a feature of specialized translated French texts only.

Table 9
Optional syntactic elements in French (overall corpora)

	Ne (total occurrences)	*Ne "explétif"*	Proportional difference, Translated corpus
FST	269	2 (0.7% of raw score)	
FTT	244	3 (1.2% of raw score)	+0.5% (*ne "explétif"* proportional to total *ne* in corpus)
	On + "zero"[156]	*On + L'*	
FST	41	7 (17% of raw score)	
FTT	256	14 (5.5% of raw score)	-11.6% (*L'on* proportional to total *on* in corpus) (TT 67% of 21 occurrences in French corpora)

Table 10
Optional syntactic elements in French (non-specialized sub-corpora)

	Ne (all)	*Ne "explétif"*	Proportional difference, Translated sub-corpus
Non-specialized FST	9	1 (11.1%)	
Non-specialized FTT	14	2 (14.3%)	+3.2% (*ne "explétif"* proportional to total *ne* in corpus)
	On (all)[157]	*On + L'*	
Non-specialized FST	13	3 (23%)	
Non-specialized FTT	14	1 (7.1%)	-16% (*L'on* proportional to total *on* in corpus)

[156] The forms *qu'on* and *lorsqu'on* were not included in this count.
[157] As with the counts for the overall corpora (FST and FTT), *qu'on* and *lorsqu'on* were not included in this count.

4.2.3 *Explicitation: Test of Significance and Interpretation of Results*

We used the Z-test to determine the statistical significance of our findings. The test statistics drew the same picture for both the overall corpora (of specialized and non-specialized texts together) and for the non-specialized sub-corpora alone. With most of the measures used, the proportional difference between translated and non-translated texts in both languages was acceptably significant, and we were able to reject the null hypothesis in favour of the research hypothesis in 50% of the cases (see Appendix X).

The significance levels for the overall corpus (of specialized and non-specialized texts together) were as follows. For the measure "reporting verb + *that*" the p-value was less than 0.0001, a highly significant finding in that the chance of error is 1 in a million. For the measure "*that* ODC" the p-value is about 0.1, an acceptable level of significance in that the risk of error is 1 in 10. For the measure "*which* ODC," the p-value was less than 0.0001, a highly significant finding in that the chance of error is 1 in a million. For the measure "*ne* explétif," the p-value is about 0.29, an approximate 1 in 3 chance of error. In this case, the difference in proportions is negative as hypothesized, but the significance level is not strong enough to reject the null hypothesis. For *L'on* in the overall French translated corpus, the evidence does not support the alternative hypothesis. The p-value is almost 1, and the difference btween the two proportions is positive and not negative as hypothesized. This finding is pronounced and we most definitely cannot reject the null hypothesis.

The significance levels for the non-specialized sub-corpora were as follows. For the measure "reporting verb + *that*" the p-value is less than 0.002. This finding is very significant in that the chance of error is only 2 in 1,000. For the measure "*that* ODC" the difference in proportions is negative as predicted, so the evidence supports the research hypothesis. However, this evidence is not strong enough to reject the null hypothesis at an acceptable level of significance, since the p-value is 0.44, and the risk of error is close to 1 in 2. We therefore retain the null hypothesis. For the measure "*which* ODC" the difference in proportions is negative as predicted, which supports the research hypothesis. Furthermore, the p-value is less than 0.03, which is strong enough to reject the null hypothesis. For the measure "*ne explétif,*" the proportional difference is negative as predicted, supporting the research hypothesis. However, we cannot reject the null hypothesis because the risk of error is nearly 1 in 2 with a p-value of just under 0.46. The measure "*L'on*" provided no more support for the alternative hypothesis in the non-specialized French translated corpus than it had in the overall corpus. The evidence was not at all favorable, since the difference was positive, counter to predictions, and since the chance of error approaches 1 (p-value 0.992151). We definitely fail to reject the null hypothesis with this measure. These findings indicate that most of the results for the English measures were unlikely to be due to chance, while the results for the French measures have an unacceptable risk of error.

Reading the measures qualitatively, it was apparent that there were more optional syntactic elements (with the exception of *that* ODC in English and *l'on* in French) used in both languages of translation. We interpret these results as support for the research hypothesis of explicitation. We gather from the above findings that

where optional explicitating elements of syntax exist in the corpora studied, a number of these elements may tend to be used more often in translated than in non-translated texts, perhaps in order to shift translations toward greater syntactic disambiguation.

We make two specific inferences from the above results, one about the supposed "formal load" of the optional syntactic elements, and the other about how optional they in fact are.

First, there did not appear to be much evidence associating the above explicitating syntactic elements with the formal register in English or French. We cannot say with any certainty on the basis of the above findings that there is likely to be any decrease (or increase) in explicitation due to lack of formal register. Had text register governed the occurrence of the syntactic elements studied, the frequencies of these elements would probably have been similar in each corpus. The only difference actually found in their use was, on the contrary, between translated and non-translated texts. Additionally, the non-specialized sub-corpora appeared to be more explicitated overall. Since the non-specialized sub-corpora consist, at least in theory, of those specialized and non-specialized texts in the overall corpora (EST, ETT, FST, and FTT) that are least likely to contain markers of formality, the findings can be interpreted as indicating that translated texts may tend to contain more such optional, explicitating elements, regardless of register.

Second, there did not appear to be much evidence that the elements studied were not optional. Had Bernstein *et al.* been correct, then in the English corpus, the conjunction *that* would have been included, without exception, in most or all of the reporting clauses that met the three conditions for "obligatory" inclusion cited

above. Although we found very few sentences that in fact met those three conditions, we did (as noted in the previous Section) find one exception suggesting, *au contraire*, that syndetic linking may be "optional" under the conditions listed by Bernstein *et al.*

Finally, we note that one element (*on* +/- "zero," see Table 9) occurred considerably more often in the French non-translated corpus (FST) than in the French translated corpus (FTT), despite the fact that the overall word counts of these two corpora are comparable. The control test with the non-specialized French sub-corpora provided the beginnings of an explanation for this finding. Given that the non-specialized sub-corpora contain comparable numbers of occurrences of *on* (FST non-specialized sub-corpus 11; FTT non-specialized sub-corpus 12), and that specialized texts were excluded from them, it appears that those of the French specialized texts that were translated contain a comparatively high number of occurrences of *on*, both with and without other elements such as the explicitating *L'*. The other side of this coin is that the non-translated French specialized texts contain far fewer occurrences of *on*. The present study is not designed to allow us to discern whether these relative numbers of occurrence of *on* reflect a translation-related tendency. This could be the subject of future study (see Section 6.5).

4.3 Simplification

As discussed in the Literature Review, simplification was previously investigated by Laviosa-Braithwaite (1996) in a corpus of mainly literary English translated texts. To add complementary information to the knowledge of

simplification gleaned in the above previous study, we investigated simplification in non-literary translated texts in both French and English. In each of the two languages included in the present study, we attempted to observe whether, as hypothesized in the introductory chapter, the translated texts had vocabularies and sentence structures that were simplified in comparison with the non-translated texts.

4.3.1 Simplification: Measures

To test the hypothesis of simplification as a recurrent feature of translation, we used text analysis software to obtain three descriptive statistics: standardized type/token ratio, lexical density ratio, and mean sentence length. These statistics had been used to investigate simplification as a recurrent feature of translated in Laviosa-Braithwaite's (1996: 114-120) comparison of Anglo-European translated and non-translated texts, the results of which demonstrated that the vocabulary ranges, information loads, and syntax of the translated texts were simplified in comparison with those of non-translated texts in the same language.

Laviosa-Braithwaite's findings with these measures were as follows. Type/token/*100* ratios were generally lower among the translated texts (1996: 125, 128, 129, 132, 133, 141, and 145). There was a tendency for lexical density ratios to be lower throughout (1996: 125, 128, 129, 141, and 145). Mean sentence lengths were lower in most of the types of translations included in Laviosa-Braithwaite's corpus, with the exception of narrative prose texts (1996: 125, 128, 132, 133, 134,

137, 141, and 146). The findings of Laviosa-Braithwaite (1996) were considered to constitute significant support for the hypothesis of simplification.[158]

In the present study, it was expected that the measure of standardized type/token ratios (henceforth type/token/*n* ratios) would show that the translated texts had comparatively limited vocabulary ranges, as seen in their lower ratios of types to tokens.[159] Lexical density ratios were expected to show that the translated texts had lower information loads, as seen in their lower proportions of content words to total running words. Mean sentence lengths were expected to prove to be shorter among translated texts, showing them to be less syntactically complex.

4.3.2 Simplification: Test of Significance and Summary of Results

The Z-test was used to determine statistical significance for comparative type/token ratios, lexical densities, and mean sentence lengths. The test statistics for these measures tended to indicate that while the research hypothesis could be accepted for the results with the English corpus, the null hypothesis should be retained for the results with the French corpus. For the measures used, the proportional difference between translated and non-translated texts in both languages of translation was not significant, and the null hypothesis is retained (see Appendix X). The significance levels were as follows.

---

[158] Using the t-test, statistical significance was found for the lower lexical densities and mean sentence lengths, but not for the lower type/token/*n* ratios. It should be noted, however, that in qualitative corpus linguistics research, statistical significance is not tantamount to confirmation or denial of hypotheses. Stubbs (1995 *passim*; also 2001, 223-224) points out that while the t-test assumes a random sample, no corpus is compiled randomly. Laviosa-Braithwaite (1996: 147 *ff*) acknowledges that the t-test results "do *not* describe the significance of the hypotheses themselves." See also Huntsberger and Billingsley (1987: 298), cited in Laviosa-Braithwaite (*ibid.*).

[159] "Type/token/*n* ratio" is shorthand for statistically standardized type/token ratio: *n* stands for the interval at which the ratio is standardized. Where the rate of **statistical standardization*** is specified, *n* is replaced with the value of the standardization rate. For instance, if the type/token ratio is standardized every 50 words, it will be referred to as the "type/token/*50* ratio."

For type/token ratios in English, the p-value was 0 (very highly significant), and the proportions were greater than zero and smaller than those of the non-translated texts as predicted in the alternative (research) hypothesis. These findings strongly support the research hypothesis, since the chance of error is essentially nil. For type/token ratios in French, we found the complete opposite. The evidence does not support the research hypothesis because the difference is negative (the type/token ratio of the translated texts is greater than that of the non-translated texts) while the research hypothesis says that it will be positive. At a p-value of 1, we definitely retain the null hypothesis.

For lexical densities, it should be recalled that the research hypothesis is that the value of the translated texts will be smaller than that of the non-translated texts, but larger than 0. In English, the difference is in favour of the research hypothesis and the p-value is 0. This finding is highly significant, as the chance of error is infinitesimal. In French, we again found the opposite: the difference in proportions is actually smaller, and there is a higher lexical density in TT. The p-value is 1 (i.e. the chance of error is nearly 100%), and the null hypothesis is retained.

For mean sentence lengths, the research hypothesis is that the value for TT will be less than for ST but greater than 0. In English, the difference was a negative number, which is evidence against the research hypothesis. Furthermore, at a p-value of 1, the chance of error approaches 100%. In French, on the other hand, the value is above zero, supporting the research hypothesis, and is highly significant, with a risk of error of 1 in a billion (p-value less than 0.0001).

Based on the test statistics, the research hypothesis seems valid for the following measures: type/token ratios in English, lexical densities in English, and mean sentence lengths in French. However, we retain the null hypothesis for the following measures: type/token ratios in French, lexical densities in French, and mean sentence lengths in English. A summary of the descriptive statistics and our interpretation of their implications for the research hypothesis follows.

Type/token/*n* Ratios

The English non-translated corpus had a type/token/*50* ratio of approximately seventy-eight percent (about 78 different words per 100 running words), while the English translated corpus had a type/token/*50* ratio of approximately seventy-five percent (about 75 different words per 100 running words). The type/token/*n* ratio of the English translated corpus was thus indeed lower as predicted, by almost four percent. The non-translated French corpus had a type/token/*50* ratio of approximately seventy-seven percent (about 77 different words per 100 running words), while the translated French corpus had a type/token/*50* ratio of approximately seventy-nine percent (about 79 different words per 100 running words). The type/token/*n* ratio of the translated corpus was therefore in fact two percent higher, counter to the outcome predicted.[160] In general, the type/token/*n* ratios show reverse tendencies in each language, a result which does not appear to support our hypothesis.

[160] Since Laviosa-Braithwaite (1996) had calculated statistical significance for type/token/*n* ratios, we were curious to see how our results would compare with hers if we calculated these inferential statistics at this point in our study. Our results were as follows: English corpus: $p = .144$ (>.05); French corpus: $p = .175$ (>.05). Independent-sample t-tests thus did not show the differences in the translated and non-translated corpora to be statistically significant in either language. Given the relatively small size of our corpora, however, it is difficult to say whether this result (that there is no statistical significance for either corpus) should be qualified as a failure of the hypothesis or not. We continue to feel that the descriptive statistics and raw numbers can tell us more.

Lexical Density Ratios

The calculation of the lexical density ratios in each corpus is given in Table 11 below. The translated English corpus had an overall lexical density ratio that was lower than the corresponding non-translated corpus, as predicted, by slightly less than three percent. The translated French corpus had an overall lexical density ratio that was nearly three percent higher, counter to predictions. The lexical density ratios of the translated English and French corpora show reverse tendencies. These findings cannot be construed as support for our hypothesis. The total number of running words and content words in each corpus is shown below, since these figures are necessary for the calculation of their proportion in terms of lexical density ratios.

Table 11
Calculation of lexical density ratio

EST	
Running words (N: tokens in text)	61,699
Function words	23,934
Content words (L: Running words minus total function words)	37,765
Lexical density ratio = 100 x L/N	
100 X 37,765 ÷ 61,699 = **61.2 %**	
ETT	
Running words (N: tokens in text)	60,435
Function words	25,041
Content words (L: Running words minus function words)	35,394
Lexical density ratio = 100 x L/N	
100 X 35,394 ÷ 60,435 = **58.6 % (-3%)**	
FST	
Running words (N: tokens in text)	60,725
Function words	31,179
Content words (L: Running words minus function words)	29,549
Lexical density ratio = 100 x L/N	
100 X 29,549 ÷ 60,725 = **48.7 %**	
FTT	
Running words (N: tokens in text)	60,798
Function words	29,583

Content words (L: Running words minus function words) 31,215
Lexical density ratio = 100 x L/N
100 X 31,215 ÷ 60,798= **51.3 % (+3%)**

Sentence Lengths

The mean sentence length of the translated English corpus was 15% higher. However, the mean sentence length of the translated French corpus was nearly 12% lower. These are reverse tendencies: the predicted decrease did occur in French, but not in English. The translated texts in the two languages thus showed opposing trends, meaning that no language-independent recurrent feature of translation was observed, and that our hypothesis of simplification was again not supported.

Summary of Results

The results obtained with the above measures are summarized in Table 12. They consistently run counter to expectation. The differences found between the standardized type/token ratios of the translated and non-translated corpora were not consistent in both languages: in English, the translated texts had a nearly four percent lower vocabulary range (as predicted and statistically significant), while in French, the same range was nearly two percent higher (counter to predictions and not statistically significant).The lexical density ratio was nearly three percent lower in the translated English corpus (as predicted and statistically significant), but was higher by the same amount (3%) in the translated French corpus (counter to predictions and not statistically significant). Mean sentence lengths, which had been predicted to be lower in both languages, were nearly twelve percent lower in

French (statistically significant), but about fifteen percent higher in English (not statistically significant).

Table 12
Simplification measures: summary of results

English (overall corpus: specialized and non-specialized texts)				
Simplification Feature	Measure	Non-translated (EST)	Translated (ETT)	Difference TT
Vocabulary range	Type/token/*50* ratio	78.4%	74.8%	-3.6%
Information load	Lexical density ratio	61.2%	58.6%	-2.6%
Syntactic complexity	Mean sentence length	21.62	24.93	+3.3 (+15 %)
French (overall corpus: specialized and non-specialized texts)				
Simplification Feature	Measure	Non-translated (FST)	Translated (FTT)	Difference TT
Vocabulary range	Type/token/*50* ratio	77.2%	79.1%	+1.90%
Information load	Lexical density ratio	48.7%	51.3%	+2.7%
Syntactic complexity	Mean sentence length	26.18	23.06	-3.1 (-12%)

4.3.3 Simplification: Interpretation of Results

The differences seen in the results for the simplification measures appear to be related in some way to the characteristics of each language of translation, and to the fact that each language has the other as a source language, rather than to the translated or non-translated status of our corpora. The hypothesis of simplification predicted that compared to non-translated texts, translated texts would be observed to have consistently smaller vocabulary ranges, lighter information (lexical) loads, and simplified syntax, regardless of the TL. Contrary to this prediction, it appears here that the results in a TL depend on the features of its SL: the vocabulary range of translated texts appears to be slightly smaller in English but larger in French, the

information load is lower in English but higher in French, and the syntactic complexity is higher in English but lower in French. These results suggest that if simplification is indeed a feature of translation, it is not seen solely in impoverishment of vocabulary, reduced information loads, or simpler sentence construction. Other measures may need to be explored.

Our hypothesis of simplification as a consistently recurrent feature of translation does not, in short, appear to be supported, and our results contradict those of Laviosa-Braithwaite (1996). Since this is a new area of research, we can only speculate as to why our findings were so different from those of the previous study.

The literature offers one or two clues. Halliday's definition of lexical density (e.g. Halliday 1994: 351) has been widely applied as a means of distinguishing spoken from written language, particularly in tests of readability (e.g. Biber 1988, Hunt 1966, or Zakaluk and Samuels, eds. 1988). A lower lexical density ratio is generally identified with spoken style, whereas a higher lexical density ratio is identified with written style. This might help explain why lexical density ratios were higher in French translation but not in English: it may be that clearly distinguishing oral from written style is considered more important in French than it is in English. Of course, an association of lexical density with the oral-to-literate continuum does not explain why lexical density ratios were generally higher in the British translated texts studied in Laviosa-Braithwaite (1999). Only further study with larger corpora, more text types, and more languages could begin to adequately explain the differences in the findings.

To use mean sentence length as a measure of simplification is to assume that shorter sentences have simpler syntax while longer sentences have more complex syntax.[161] It followed that if our translated sentences had proved to be shorter and more numerous than our non-translated sentences, this might have constituted evidence of comparative simplification of the translated texts.

However, mean sentence lengths proved to be distinctly different in the two translated corpora, and they moved in opposite directions from their respective non-translated corpora: translated English sentences were longer, translated French sentences shorter. These results do not suggest that translated syntax is recurrently simplified; rather, they hint that TT mean sentence length will tend to be directly influenced by SL sentence length, regardless of the language pairs involved.

[161] This is in keeping with Hunt's (1966) theory that the "T-unit," that is, the independent clause and all its subordinates, lengthens and grows in complexity with the development of syntactic maturity in reading and writing. See Gaies (1980).

5. Investigating Levelling-out

To the best of our knowledge, levelling-out, the fourth translation universal proposed by Baker (1996: 184-185), has not heretofore been systematically investigated in corpus-based study. As mentioned in the introductory chapter, we hypothesize that translated texts will show observable levelling-out.

We propose to use a new method of measurement to investigate levelling-out: the application of various readability indices to our corpora. The readability indices selected are described in detail in Sections 5.1.1 to 5.1.6 below. We also propose to use a different set of materials (an appropriate set of sub-corpora) for the application of readability indices to the investigation of levelling-out.

Our specific hypothesis (see Section 1.6.4) is that translated texts will generate more homogeneous sets of scores,[162] and that they will also show a central tendency to be more "readable" by the standards of a given readability index, than will non-translated texts.[163] A number of predictions follow from this hypothesis.

For the purposes of investigating levelling-out in our corpora, we predict that the readability scores of translated texts will have lower standard deviations (lower dispersion above and below their mean scores) than the scores of non-translated texts, showing the translated texts to be more homogeneous. Furthermore, translated texts will tend to score closer to the centre of a given independently-established readability continuum (e.g. "target," "average," or "optimal"; Gunning 1964;

---

[162] If the sets of scores are regarded solely as ranges of "internally-generated" values (Baker 1996: 177, 184-185; Laviosa-Braithwaite 1996: 134-136), and are not placed within the continuum of a readability index, it is predicted that compared to the scores of non-translated texts, the upper and lower scores of translated texts will establish narrower ranges of values.

[163] If the sets of scores are regarded from the point of view of a readability index, the scores are therefore "placed within the continuum" of that index: they are referred to an independently defined, "pre-established" set of references, as in Shlesinger (1989: *passim*).

Tremblay 2000; Björnsson 1968), while non-translated texts will tend to score closer to one of the peripheries of that readability continuum (e.g. "very easy to read" or "very difficult to read"). In other words, compared to non-translated texts, translated texts will generate a narrower range of scores: their set of scores will have a lower standard deviation, indicating greater homogeneity. Furthermore, the scores of translated texts will be closer to the centre of a readability index, indicating that greater readability is their central tendency.[164]

Finally we will replicate a brief "side study" carried out by Laviosa-Braithwaite (1996: 143-146) and cited in Baker (1996: 184) as an additional method of investigating levelling-out in the present study. Following Laviosa-Braithwaite, we will calculate the standard deviations for type/token/*n* ratio and mean sentence length, noting whether, as hypothesized, they indicate that the translated corpora are homogeneous. The "main" corpora (EST, ETT, FST, FTT) will provide the values used; this will allow comparison with the other results for levelling-out obtained with the non-specialized sub-corpora.

5.1 Readability in the Non-specialized Sub-corpora

Levelling-out can theoretically be observed through any measurable characteristic of texts. Ideally, the instrument used should be both valid and reliable, consistently measuring the same characteristic in all of our texts. The best measure of levelling-out would therefore be a validated test designed to apply to a range of written texts. This test should consistently qualify and quantify recognizable

[164] Although there is a different set of names for the upper and lower peripheries of each readability index, the continuum of readability can be generally assumed to go from "extremely easy to read" to "extremely difficult to read." Note that different readability indices postulate different populations as readers.

characteristics of written texts, and should produce understandable information about them.

The readability index—an instrument used in educational psychology and in technical writing practice as a means of assessing the degree of skill required to read a given text or set of texts—easily meets these criteria. There are readability indices that have been validated for English and French. Readability indices satisfy Shlesinger's (1989) precondition that the "equalizing effect" of translation should be measured using a generally recognized, "pre-established" continuum (Shlesinger 1989: 96-97; 170-171).[165] They also make it possible to follow Baker's (1996: 184) suggestion that levelling-out should be measured with sets of numerical values, such as those generated by readability indices.

There are a number of readability indices (e.g. Fry 1989; Gaies 1980; Henry 1987; Klare 1988; Macdonald *et al.* 1987; Pigeon 2002; Roberts *et al.* 1994; Russell, Pamela 1993; Shrock 1995-2003; Tremblay 2000; Venezky 1984). Those that we have selected for use in the present study (on the basis of criteria introduced in Section 2.4.2, last paragraph) will be described in detail in the sections which follow. They include five indices applicable to English-language texts, and three indices applicable to French-language texts. Three of the indices (the Flesh Reading Ease and Flesch-Kincaid Grade scales, and the Fry Graph) apply to English only. Their use is apparently widespread. An automated calculation of the Flesch Reading Ease and Flesh-Kincaid Grade indices is bundled with the Microsoft Word 2000 Office software package, which at the time of this writing was one of the world's

[165] The "oral-to-literate" continuum was used for this same purpose in Shlesinger's study (1989: *passim*). For the purposes of the present study, we are proposing to apply the well-established continuum of "readability," which ranges from "very easy to read" to "very difficult to read."

leading word processors. The Fry graph is in the public domain and is popular among primary and secondary school teachers as a pedagogical tool, for preparing materials that provide instruction in reading skills.

One of the indices applicable to English language texts, the Gunning-Fog Index, has been adapted by the Institut Canadien des Actuaires (henceforth **ICA***). for use with French texts. In the present study, the version of the Gunning-Fog Index adapted for use with French by ICA is referred to as the ICA Index. There is one readability index, Björnsson's Lix, that was developed for a number of languages of European origin. The Lix has been applied to both French- and English-language texts in the present study. The readability index developed by Georges Henry in cooperation with his professor and mentor, Gilbert de Landsheere, is referred to as the Henry-de Landsheere Index in the present study. It was developed and validated for application to European French-language texts.

In addition to appropriate methods for measuring levelling-out, we need materials that are suited to them. Four non-specialized sub-corpora (see Section 3.1.1) have been selected for the purposes of the present study.[166] Their design satisfies the requirements of a number of the readability indices used in the present study: they include only whole (non-excerpted) texts, they are matched with each other in length, and they are uniformly non-specialized.[167]

The larger "main" corpora (EST, ETT, FST, FTT) are not suitable for application of readability indices for two reasons, one having to do with the limited

---

[166] Note that the non-specialized sub-corpora are also used in the present study as control (reference) corpora for testing the possible influence of text register on explicitation: see Sections 4.2.1 and 4.2.2.

[167] Readability indices are highly sensitive to specialized vocabulary. Many readability indices (including those used in the present study) rate texts as harder to read if they contain words that are not on a basic vocabulary list, that are above a certain length in letters, or that have more than two or three syllables. It is therefore important not to mix specialized and non-specialized texts when measuring readability.

resources available for the present study, and the other with the textual characteristics of the corpus. We will briefly discuss each of these reasons in turn.

First, the majority of readability scores (with the exception of those automatically calculated by word processing software such as Word 2000) are manual formulas.[168] For a lone researcher, running these indices on corpora totalling nearly 250,000 words—accurately and without the use of automated tools—would have been a physical impossibility. We therefore considered it a necessity to instead select what we judged to be a suitable sample from each of the overall corpora of specialized and non-specialized texts.

Second, readability formulas are most often applied to one kind of text at a time: specialized texts are usually compared with specialized texts, non-specialized texts with non-specialized texts.[169] Our "main" corpora contain a mix of types. They contain both specialized and non-specialized texts, the latter being defined from the translator's point of view as those with a relative lack of specialized terminology.[170] The non-specialized texts have word counts that vary greatly from corpus to corpus.[171] It was therefore decided that, for the purposes of measuring levelling-out as a hypothetical recurrent feature of translation, we would exclude "specialized"

---

[168] Many of the variables of these formulas, such as syllable counts, must be calculated manually.

[169] Technical writers often apply readability indices to their work (e.g. Pigeon 2002; Tremblay 2000). Specialized texts will score as more "difficult to read" than will non-specialized texts. That is, a set of specialized texts will have a range of scores that is arrayed closer to the "difficult to read" extreme of the continuum of readability.

[170] The specialized texts in the corpora are scientific, commercial, and legal; there are no literary texts. We restricted our selection criteria to terminological features, although it could be argued that non-specialized and specialized texts are also distinguished by style. In practice it is very difficult to identify stylistic features that are exclusive to one single text genre.

[171] The number of texts identified as specialized in each corpus is as follows: Scientific (EST 6/ETT 4; FST 14/FTT 2); Commercial (EST 13/ETT 3; FST 2/FTT 4); Legal (ETT 10). These specialized texts do not have comparable word counts: the highest total word count of specialized texts is found in FST (44,000 words), while the smallest total word count of specialized texts is found in FTT (7,000 words).

texts from this part of the study, and would apply the selected readability indices only to the non-specialized sub-corpora.[172]

Samples smaller than those commonly used in corpus-based study are necessary for many of the readability indices applied in the present study. For instance, the instructions for many readability indices require samples of *n* words, representing only a few passages (often less than a page), to be taken at evenly spaced intervals from several parts of a corpus (usually from the beginning, middle, and end). For example, the Flesch Reading Ease formula is applied to strictly delimited samples of 100 words (Section 5.1.1), and the Henry-de Landsheere formula is likewise applied to samples of exactly 250 words (Section 5.1.5).

It must be stressed that we are making use of these readability indices for the purpose of investigating levelling-out, and by comparing the degree of homogeneity (or the degree of dispersion) of the translated and non-translated sub-corpora, through their readability scores. Our primary purpose is not to discover how much skill is required for reading the texts in our corpora, but to see whether the readability scores of the translated corpora do indeed "level out" by being more homogeneous. These are two distinct uses of readability indices, and a familiar example will serve to illustrate the difference between them.

When a language skills test, such as an objective test of reading comprehension, is administered to a group of students and subsequently graded, the result is a set of scores that (assuming the test is valid and reliable) reflect the different levels of achievement of each student (Heaton 1975: *passim*). When the

[172] Note that not all of the unspecialized texts in our overall corpora were gathered into the non-specialized sub-corpora, since the latter had to be matched as closely as possible by total number of words. The total number of general (non-specialized) texts in the overall corpora is as follows (EST 43/ ETT 36; FST 31/FTT 31). See Section 3.1.1 for the design of the non-specialized sub-corpora.

scores are added up and divided by the number of students, the result is the mean score, which presumably reflects the group's overall level of achievement in reading comprehension (Heaton 1975: 168-169). The mean score denotes the group's central tendency (e.g. beginning, intermediate, advanced).

Taken alone, the mean score tells us nothing about the degree of diversity or homogeneity of the individual scores within the group (Heaton 1975: 170-171). In other words, the mean score by itself tells us nothing about the degree of dispersion of the group scores (e.g. to what extent each student diverges from the mean). However, if the standard deviation is calculated, it is possible to determine whether the group score set is homogeneous or dispersed (*ibid.*).

The present research pertains to the characteristics of texts (translated vs. non-translated) rather than to the skills of humans. However, the application is the same: we will be determining the central tendency, and (more importantly, for our purposes) the degree of dispersion, of a group of score sets. In other words, for each set of texts (translated and non-translated), we will determine a central tendency—how "readable" each set of texts is—and in doing so we will achieve our primary purpose of determining and comparing each text set's degree of homogeneity, by calculating the standard deviation of each set of scores. For the purposes of the present study, the hypothesis of levelling-out as a recurrent feature of translation predicts that the two translated sub-corpora will consistently generate readability scores that are comparatively more homogeneous, i.e. that have a smaller spread of readability scores above and below their mean value, and therefore a lower standard deviation, than the two non-translated sub-corpora. We also predict that the two

translated non-specialized sub-corpora will each have a central tendency to score as more "readable."

We will now turn to a detailed description of each of the selected readability indices, since a knowledge of their application will clarify the results we obtained from applying these measures to samples taken from our four non-specialized sub-corpora.

5.1.1 Flesch Reading Ease and Flesch-Kincaid Grade indices

The Flesch Reading Ease formula assigns scores on a scale of 0 to 100. The higher the score, the more readable the text. The designated standard level of reading difficulty is a score of 60 to 70. Texts with scores dropping below 60 are considered progressively more difficult to read; those with scores above 70 are deemed easier to read. Only a hypothetical university graduate reads texts at a score level of 0-30; American fourth graders are believed to read at a score level of 90-100 (Klare *et al.* 1954: 163).

To apply the formula the way it was originally developed in the 1950s, samples of about 100 words per individual text are selected at evenly-spaced intervals throughout the corpus. One then calculates the mean number of syllables per word and the mean number of words per sentence. The Flesch Reading Ease score is subsequently calculated using the following multiple regression formula in order to keep the resulting scale within a 100-point range (Flesch 1948 and 1948b; Klare and Buck 1954: 98-9):

Reading Ease = 206.835 – (1.015 x ASL) – (84.6 x ASW)
where:

ASL = mean (average) sentence length (the number of words divided by the number of sentences)

ASW = mean (average) number of syllables per word (the number of syllables divided by the number of words).

The Flesch-Kincaid Grade Level index rates texts by U.S. grade-school level. A score of eight signifies that an eighth-grader, reading at the expected level for an American of that age and level of education, should understand the entire text. The standard level of reading difficulty for this measure is a grade level of 7-8. The Flesch-Kincaid formula (Kincaid *et al.* 1975) converts Flesch Reading Ease Scores to U.S. grade-school levels and is calculated as:

(0.39 x ASL) + (11.8 x ASW) - 15.59

The Flesch Reading Ease and Flesch-Kincaid Grade formulas are no longer restricted to the above manual counts on short samples. For corpus-based study of many thousands of words stored on a computer hard drive, scores can be obtained using Microsoft Office Word 2000, the widely-available word processing software which automatically includes a calculation of Flesch and Flesch-Kincaid readability scores in its grammar and spelling check. Although this calculation is, according to the software's Help file, based on probability, and while the results are therefore unlikely to be due to chance, the output is not produced in such a way as to allow us to perform tests of statistical significance on it.

At the time our research was carried out, the automated Microsoft Office Word 2000 output was available only for English texts. We therefore obtained Flesch and Flesch-Kincaid readability scores for our English non-specialized sub-corpora (translated and non-translated), each of which we compiled into a single

machine-readable document.[173] Since each English non-specialized sub-corpus is about 15,000 words, the resulting Flesch and Flesch-Kincaid scores were produced for a total 30,000 words. The results obtained using the Flesch and Flesch-Kincaid readability indices are reported in Section 5.2.1.

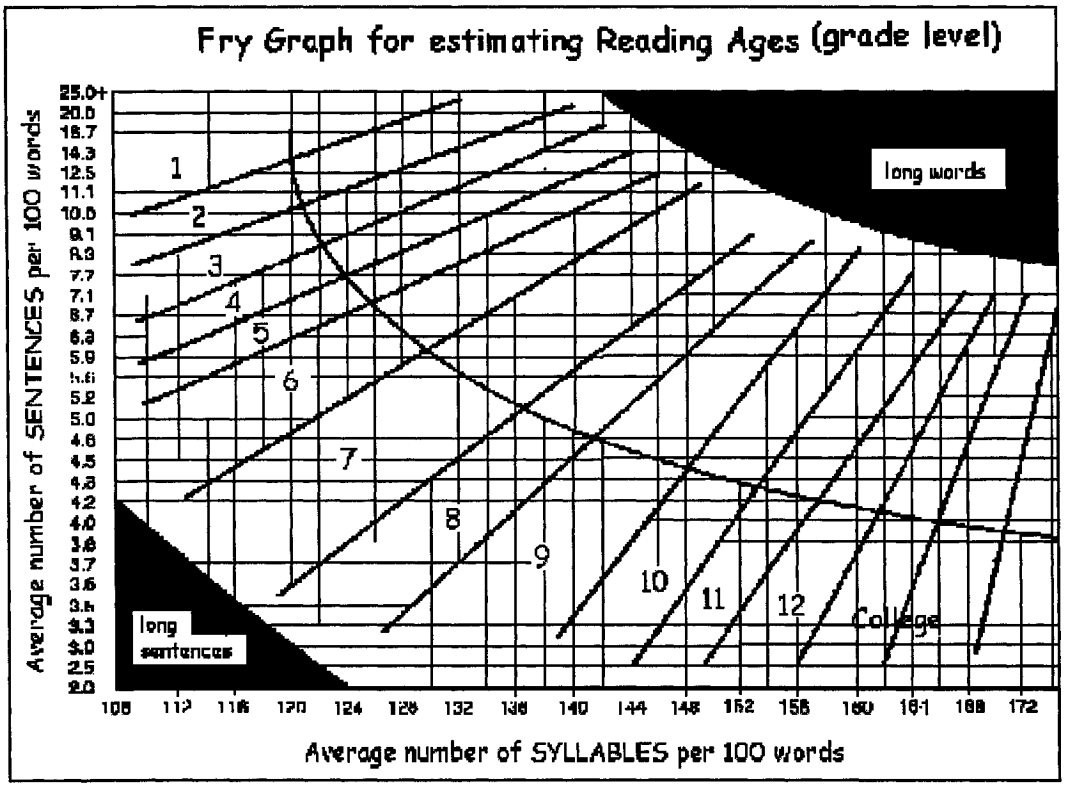
5.1.2 Fry readability graphs

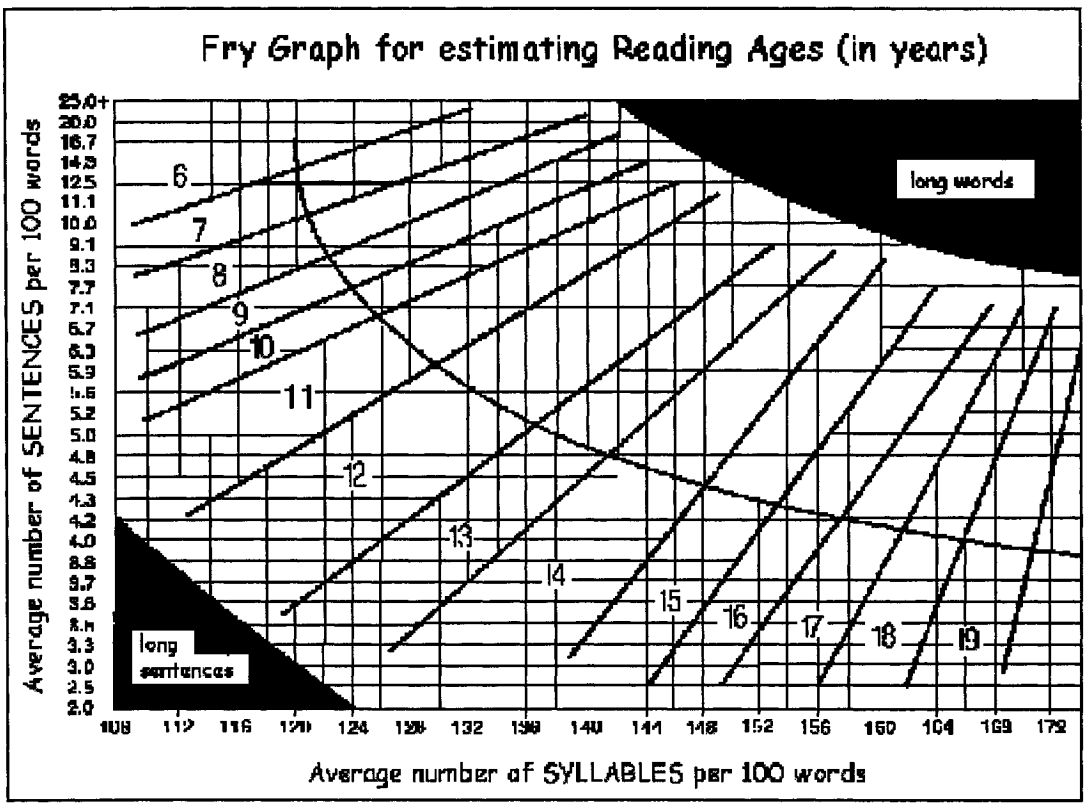
These graphs, which are intended for use with English texts only, are designed for easy use by teachers. They are widely published, are in the public domain, and appear to be quite popular among educators.[174] To use the Fry Graphs, one calculates the mean number of sentences per 100-word text and the mean number of syllables per hundred words. There are two versions of the graph: one which gives the (American) grade level, the other which gives the age level at which a hypothetical reader will first comprehend the entire text. The two Fry Graphs are shown below.

---

[173] In order to obtain overall scores for a corpus, it is necessary to compile all the texts it contains into a single plain text format file (with the extension .txt). The Microsoft Office Word 2000 grammar and spell checking function can then be applied to this single very large document. This makes it possible to quickly calculate readability scores for whole corpora, instead of for small samples taken at evenly-spaced intervals.

[174] Along with the Flesch Reading Ease and Flesch-Kincaid indices, Fry's Readability Graphs are included in a suite of automated readability formulae called *Readability Plus* (see http://www.micropowerandlight.com/rd.html) which we opted not to purchase and run since it could not be applied to French texts. The graphs themselves (one giving readability scores by grade level, the other by age of hypothetical reader) are in the public domain and are available at http://school.discovery.com/schrockguide/fry/fry2.html.


Fry Graph for estimating Reading Ages (grade level)
Average number of SENTENCES per 100 words
Average number of SYLLABLES per 100 words
long words
long sentences
1
2
3
4
5
6
7
8
9
10
11
12
College


Fry Graph for estimating Reading Ages (in years)
Average number of SENTENCES per 100 words
Average number of SYLLABLES per 100 words
long words
long sentences
6
7
8
9
10
11
12
13
14
15
16
17
18
19

In the present study, we applied the Fry Readability Graphs to four samples of fifteen randomly-selected 100-word texts taken from each of the non-specialized sub-corpora (at least 10% of each sub-corpus). The two sets of samples have a total 6,000 words (3,000 words per set). Calculating the scores for all the individual texts allows us to subsequently calculate each group's mean score, as well as the standard deviation for each set of scores.

Although no optimum grade or age level is given (due to the intended use of these graphs, which is for teachers conducting classes for a range of ages), we will consider a Grade level of 12 to be optimum for the purposes of the present study. The results obtained using the Fry readability graphs are reported in Section 5.2.2.

5.1.3 Gunning-Fog Index

This formula provides an index of relative levels of reading difficulty. It is often used by technical writers (see for instance Pigeon 2002 and Roberts *et al.* 1994). Scores range between 0 and 20 and are pegged to American educational levels: a score of 20 indicates a reading difficulty that is Ph.D.-level. The "target" readability scores for Gunning-Fog are between 11 and 15 (Gunning 1964). The Gunning-Fog Index is calculated on 100-word sample segments selected at regular, evenly-spaced intervals throughout a document (or corpus).

The "average" (mean) sentence length (commonly designated by the acronym ASL, which we will use to designate this score henceforth) is calculated by counting the sentences in the segment (including fragments of sentences that start within the selected sample), counting the number of words in the segment that have three or more syllables, adding the two numbers (mean sentence length + words

with 3+ syllables), and then multiplying that sum by 0.4. This is thus an entirely manual count.

In the present study, we applied the above Gunning-Fog Readability Index to samples of fifteen 100-word texts taken at evenly-spaced intervals from each of the English non-specialized sub-corpora (at least 10% of each sub-corpus). The two sets of samples have a total 6,000 words (3,000 words each). Calculating the scores for all the individual texts allowed us to subsequently calculate each group's mean score, and thereby to calculate the standard deviation for each set of scores. The results obtained using the Gunning-Fog Index are reported in Section 5.2.3.

5.1.4 Lix readability formula

Validated for a number of European languages including English and French, and easily adapted for use with electronic corpora, the Lix[175] readability formula is a useful addition to our roster. The manual formula, which is calculated using 100-word samples, is quite simple:

$$Lix = Lo + Ml$$

Where Lo = the number of long words (containing six or more letters)

Ml = the arithmetic mean of the sentence lengths

Lix scores are meant to be interpreted as follows, with an expected score range of about 35 points, from an extreme low of 20 points to an extreme high of around 55 points.

[175] Lix stands for "LäsbarhetsIndeX," Swedish for "readability index."

Very easy	20-30	Children (primary school)
Easy	30-35	Young adults (secondary school)
Optimal	**35-50**	Adult (high school graduate)
Difficult	50-55	Adult (undergraduate)
Very Difficult	55+	Adult (specialized)

The formula is easily adapted so as to eliminate the necessity for taking 100-word samples, as follows:[176]

Adapted LIX =
(ASL) + 100 x (Number of long [6+-letter]words/Number of words)

As with the Flesch and Flesch-Kincaid readability formulas (Section 5.1.1), the Lix formula can be calculated on whole corpora, rather than on the short samples necessitated by manual counts. The values for each of the variables used in the Lix are obtained automatically. The Lix is furthermore applicable to texts in French, an advantage for the present study.

We therefore calculated Lix readability scores for all four of the non-specialized sub-corpora, two sets in each language (a total 63,817 words). Since we also wished to calculate the mean score and the standard deviation for each of the non-specialized sub-corpora in both languages, we did not use single compiled documents (as was done with the Flesch and Flesch-Kincaid readability formulas; see Section 5.1.1), but instead used sets of individual texts for our calculations. The results obtained with the Lix are reported in Sections 5.2.4 (English) and 5.2.6 (French).

[176] The elements of this adapted Lix formula are all easily obtained using the WordList feature in WordSmith Tools.

5.1.5 *Henry-de Landsheere French readability index*

One of the most thoroughly validated readability measures for French is the Henry-de Landsheere formula, a well-known educational tool and an index used by students of journalism in Europe (Vandendooren 1999). The user of the Henry-de Landsheere formula is instructed to take text samples of exactly 250 words each.[177] Three variables are calculated. The first variable is the mean sentence length. The second variable is the number of words that are *not* present on Gougenheim's list of "fundamental French" words, which is a basic vocabulary list (Gougenheim *et al.* 1967: 69-113).[178] This second variable is referred to as the AG or "absents de Gougenheim" (Henry 1987, *passim*). The AG are counted and calculated as percentages:

AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)

To better reflect a basic vocabulary for francophone Canadian adults, we modified the Gougenheim list slightly, by adding to it the words that designate Canada, or Canadians, by country or province (*canadien [ne –s]*, *québecois [e]*, and so on).[179] We excluded other proper nouns and acronyms from this count.[180]

The third variable is the DEXGU or "nombre d'indicateurs de dialogue." These are calculated by counting the number of "dialogue indicators" (such as exclamation marks, opening French quotation marks or *guillemets*, and first- and

---

[177] "Dans des blocs de moins de 250 mots, une phrase longue pourrait fausser tout le résultat," explains Vandendooren (1999: Chapter 4, page 2).

[178] The list is reproduced in Henry 1987: 192-195.

[179] The words *français(e –s)* figure among the Gougenheim's list of basic vocabulary for French pupils, so we added the analogous Canadian words.

[180] Gougenheim's list includes some proper names, but does so on the basis of the outcome of tests, and not according to any system applicable to our modification of the count. For instance, the names of some months (*août, décembre, juillet, octobre*) appear on the Gougenheim list, while the rest (*avril, février, janvier, juin, mars, septembre, novembre*) do not.

second-person French pronouns), and by then converting that number to a percentage:

DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots)

A score of 40 is considered optimal. Vandendooren (1999) gives the following interpretation scale to be used by journalists who are targeting an adult public that reads at the grade level of Secondaire supérieur 11-12 (within the educational systems of France and Belgium):

Trop difficile	10-35
Bonne lisibilité	35-45
Lisibilité parfaite	**40**
Trop facile	45-70

Because intensive manual counts are necessary for application of the Henry-de Landsheere French Readability Index, we used randomly-selected samples of six 250-word texts (about 10% of each sub-corpus) taken from each of the non-specialized French sub-corpora. The two sample sets had a total 6,000 words (3,000 each). Calculating the scores for all the individual texts allowed us to subsequently calculate each group's mean score, as well as the standard deviation for each set of scores. The results obtained using the Henry-de Landsheere French Readability Index are reported in Section 5.2.5.

5.1.6 ICA French readability index

The only readily available Canadian index for French readability is the one used by ICA. The scores are calculated by adding the value of the mean sentence length to the percentage of words with four syllables or more, and multiplying the sum by 0.4:

(L + M) x 0.4

where L = mean sentence length, and

where M = the percent of words with 4+ syllables

It must be stressed that this formula has not been validated directly. It is adapted from the English-validated Gunning-Fog Index, and it takes into account the observation made by de Landsheere that French words are generally longer (Tremblay 2000). However, since the ICA formula is in regular use by actuaries in Canada, and since it is much simpler to apply than the Henry-de Landsheere formula, we opted to apply it to samples from each of the non-specialized French sub-corpora.[181]

The ICA scale is regularly applied to specialized documents (such as insurance contracts), but it is intended to show their "degree of transparency" (readability) for the general public (Tremblay 2000: 1). Documents with a score of 13 or more are considered to be "...difficiles et complexes et nécessit[a]nt une scolarité de niveau collégial ou universitaire."[182]

The ICA interprets the scores as follows. The illustrative quotes are cited from Tremblay (2000: 2).

ICA French Readability Score Interpretation

6 and less: Very easy (« comme les bandes dessinées »)

9-10: Target Score (« genre *Sélection du Reader's Digest* »)

13+ : Difficult (« par exemple, les revues spécialisées »)

[181] We could not find a way to apply it to the entire corpus of non-specialized French (translated and non-translated), however. Automating a syllable count for each word in the sub-corpora was beyond our limited knowledge of computer programming.

[182] "difficult and complex, requiring post-secondary education" (Tremblay 2000: 2; trans. D. A. Williams).

Logically, if one is to apply it to non-specialized texts, the ICA scale may be expanded to put it in keeping with the more usual five-point continuum of readability:

6 and less:	Very easy
7-8:	Easy
9-10:	**Target Score** (general public, grade 11-12)
11-12	Difficult (general public, grade 11-12)
13+ :	Very difficult (post-secondary)

The ICA formula is calculated on 100-word sample segments selected at evenly-spaced intervals throughout a document (or corpus). One of the variables for the ICA formula (i.e. the percent of words with 4+ syllables) requires intensive manual counting. We therefore selected fifteen 100-word texts (about 10% of each sub-corpus) from each of the two non-specialized French sub-corpora, for the purposes of calculating the formula. The two sets of samples have a total 6,000 words (3,000 words each). Taking samples composed of individual texts allowed us to subsequently calculate both the mean score and the standard deviation for each set of scores. The results obtained using the ICA French Readability Formula are reported in Section 5.2.7.

5.2 Results

The standard deviation of the readability scores provided evidence of greater dispersion, not greater homogeneity, in the translated sub-corpora, although this was not true of all of the scores, as will be seen below. However, by comparing the mean readability scores to the designated optimal score or optimal score range for each index, and by calculating the difference in terms of percentage of overall

predicted range of scores, we found evidence that in both languages, the translated sub-corpora consistently scored "harder to read" as a central tendency. These results are presented in Tables 13 to 24.

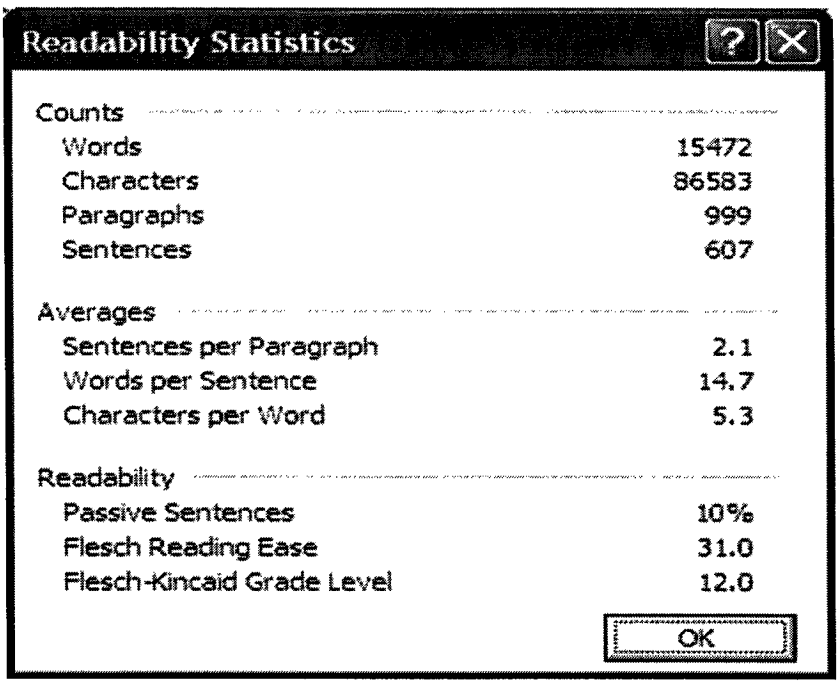
We also calculated the standard deviation for the values of two measures used in the study of simplification as a recurrent feature of translation (in Section 4.3). The results for this "control test" are presented in Tables 25 and 26.

5.2.1 *English: Flesch and Flesch-Kincaid*

Since Microsoft Office Word 2000 output was available for English texts only (see Section 5.1.1), we used it solely as an indicator of the central tendency of the English non-specialized sub-corpora. The Microsoft Office Word 2000 output for the non-translated English texts is shown in Figure 1 below.[183]

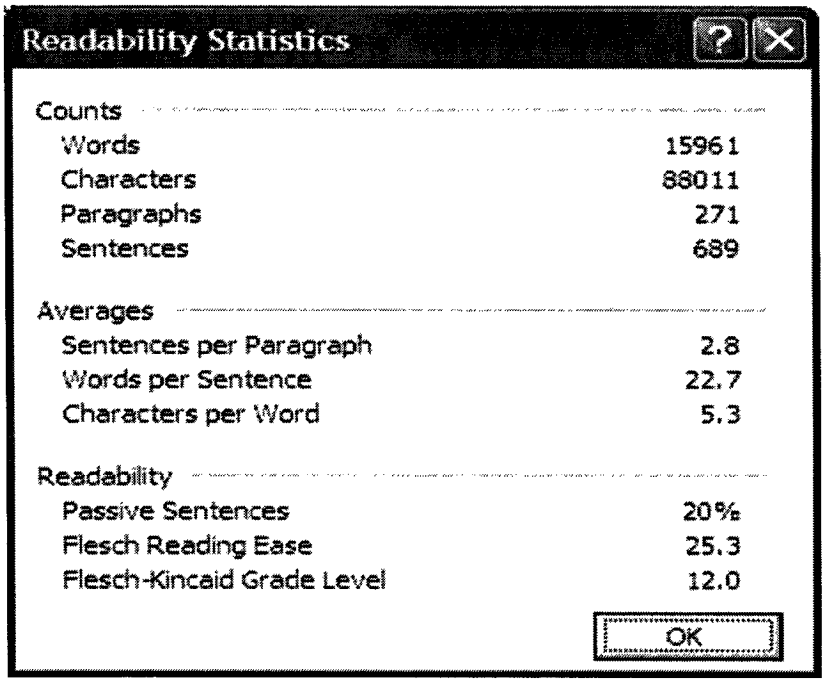
[183] The number of words counted by Microsoft 2000 is slightly lower than the number of tokens calculated by WordSmith because Word 2000 calculates these numbers differently.

Figure 1: Screenshot of Word 2000 Readability Scores, EST



The Microsoft Office Word 2000 output for the translated English texts is shown in Figure 2 below.

Figure 2: Screenshot of Word 2000 Readability Scores, ETT



The Flesch Reading Ease scores assigned by Microsoft Office Word 2000 indicated that the translated English sub-corpus was "harder to read" on the Flesch Reading Ease scale. The non-translated English sub-corpus scored 31, while the translated English sub-corpus scored 25.3, or "harder to read" by 5.7% of the possible score range.[184]

Microsoft Office Word 2000 will assign Flesch-Kincaid Grade Level scores only up to Grade 12, it should be noted. This output does not provide an accurate scale for the purposes of present study, since the top score is well above Grade 12

[184] Since each readability formula is interpreted relative to its own scale, it is helpful to know how the scores obtained differ by percentage of each scale. We therefore give both the absolute differences and the percentage differences. The latter are obtained by subtracting the bottom from the top potential score points in the scale, and then dividing the difference by that number. The percentage difference is then calculated, and in the case of the Flesch Reading Ease formula, the scale is 1-100 (the higher the score, the higher the readability), making the difference in score points equal to the percentage difference. Better demonstrations of this calculation appear below, with other scale ranges.

for many of the texts included in the present corpora. However, a patch exists which allows users to calculate manually the correct Flesch-Kincaid Grade Level of a text based on the Microsoft Office Word 2000 output.[185]

We ran the manual calculations necessary for implementing the patch, and were thus able to determine that the non-translated English sub-corpus had an overall grade level readability score of 12.6, while the translated English sub-corpus had the considerably higher overall score of 15 (which is more extreme from the point of view of the readability continuum, and is therefore a noteworthy difference in terms of the measurement of levelling-out). This is a difference of nearly three (2.78) grade levels, a full 20% of the scale.[186] The translated English sub-corpus therefore had Flesch Reading Ease and Flesch-Kincaid Grade Level scores that were closer to the "very hard to read" extreme end of the readability continuum than were those of the non-translated English sub-corpus.

Along with its calculation of Flesch and Flesch-Kincaid scores, the grammar checker of Microsoft Office Word 2000 also provides the percentage of passive-voice sentences in a document. The Microsoft Office Word 2000 output showed that there were 10% more sentences in the passive voice in the translated texts (see screenshots above). Style guides and writing manuals often counsel against unnecessary use of the passive voice in English, claiming it to be more

---

[185] There is a bug in Microsoft Office Word 2000 which must be manually corrected before a true Flesch-Kincaid grade level readability score can be obtained. A 2002 posting to the *Mimetics Discussion List*, a listserv associated with the *Journal of Memetics - Evolutionary Models of Information Transmission,* notes that the Microsoft implementation of Flesch-Kincaid seems to "invoke a ceiling function" that slots all "hard [to read]" texts at Grade 12, regardless of their actual readability score. Lynch recommends the following simple "linear" formula as a manual patch: FK = 13.26 + 0.248*ASL - 0.139*FRE, where FK = Flesch Kincaid Grade Level; ASL = Average [mean] Sentence Length (obtained by Word); and FRE = Flesch Reading Ease (obtained by Word). Implementing Lynch's manual patch, we obtained the following Grade Levels for our English corpus: EST 13.26 + (0.248 x 14.7) – (.139 x 31) = **12.6** (12.5966); ETT 13.26 + (0.248 x 22.7) – (.139 x 25.3) = **15** (15.3729).

[186] We calculated the percentage difference as follows: 15.3729 - 12.5966 = 2.7763 ÷ 16 (the grade scale of Grade 1 to "College" senior year) = .1735 x 100 (to calculate the percentage) = 17.35%.

difficult to read than the active voice.[187] This result is therefore in keeping with the Flesch and Flesch-Kincaid scores obtained via the Microsoft Office Word 2000 output.

It was not possible to calculate the standard deviation of the above Flesh and Flesch-Kincaid scores, because the Microsoft Office Word 2000 grammar and spellchecker was run on two single documents, each containing a compilation of all the texts in the non-specialized sub-corpora. As discussed below, several other indices were judged to be more suited to the purpose of comparing the homogeneity of the group scores for each sub-corpus. In Section 5.4, we interpret the above findings.

5.2.2 English: Fry Graphs

Fry Graph readability scores were calculated for the fifteen 100-word texts selected from each non-specialized sub-corpus. The standard deviation of the individual text scores was then calculated using the mean of the total scores for each sub-corpus. The results are presented in Tables 13 and 14 below.[188]

With the Fry Graphs, the readability scores of the translated English sub-corpus have a standard deviation of 3.19, which is close in value to the standard deviation of the non-translated non-specialized English sub-

[187] The *Canadian Style* 13.05, for instance, advises writers and translators to "give preference to the active voice" because the passive voice "tends to be wordy and impersonal."

[188] In Table 13, note that scores labelled "College" in the Fry Grade graph cover three grade years (see Section 5.1.2). For the purposes of calculating scores, we have assigned the appropriate values above Grade 12 (i.e. Grade 13, 14, or 15) to each of the "College" grade years. Scores that are off the Age chart are calculated as 21 for the purposes of the present study. The standard deviation has been calculated using the Age scores only.

corpus (3.26). The comparative degree of homogeneity of the two sub-corpora is therefore very close.

However, according to the Fry Graphs, the translated and non-translated non-specialized sub-corpora have divergent central tendencies: their readability levels are not the same. The translated non-specialized sub-corpora require the reading skills of a 21-year-old adult (Table 13), while the non-translated non-specialized sub-corpora require the reading skills of a 16-year-old teenager (Table 14). For each sub-corpus, there is thus a difference in the age scores of five years, which is 35.71% of the possible score range for the Fry Reading Age graph. In sum, the translated texts scored as considerably harder to read.

The Fry Graphs appear to distinguish translated from non-translated texts in an interesting and unexpected way: they appear to have a continuum with extremes that do not necessarily represent the characteristics of translated texts. The upper extremes of the Fry Graphs proved inadequate, in the present study, for scoring translated texts. Approximately 67% of the TTs had a variable (sentences per 100 words or syllables per 100 words) that was "off the scale" on one of the Fry Graphs (Table 14).

Furthermore, the overall Age score for the translated sub-corpus was off the Fry readability scale (Table 14). This suggests that the Fry Graphs might be included among instruments used to distinguish translated texts from non-translated texts in future study (see Section 6.5). We will discuss our interpretation of the above findings in Section 5.4.

Table 13
Fry scores and standard deviation (non-specialized EST)

Sample text	Sentences per 100 words	Syllables per 100 words	Individual scores (grade/age)	Mean (age)	(d)	(d^2)
CB1	5	146	9/14	16	-2	4
CB3	4	133	8/13	16	-3	9
CP11	6	147	7/12	16	-4	16
CD2	3	171	15/20	16	4	16
CP16	4	130	8/13	16	-3	9
CP12	5	138	8/13	16	-3	9
CP13	4	155	11/16	16	0	0
CP16	4	130	8/13	16	-3	9
CP19	5	129	7/12	16	-4	16
CP20	4	151	10/15	16	-1	1
PW5	4	173	15/20	16	4	16
PW6a	5	176	15/20	16	4	16
PW6b	5	188	15/21	16	5	25
SC1	4	235	15/21	16	5	25
PW10	5	134	7/12	16	-4	16
TOTAL	(67)	(2336)			$\sum d^2 = 171$	
MEAN	4.46	155.73			$Sd = \sqrt{171/16} = \sqrt{10.68}$	
Mean Score	Grade: 11 / Age: **16**				**Sd = 3.3**	
Total Off Scale	**0**					

Table 14
Fry Scores and standard deviation (non-specialized ETT)

Sample text	Sentences per 100 words	Syllables per 100 words	Individual score (grade/age)	Mean (age)	(d)	(d^2)
CT1	3	189 (off scale)	15/21	21	0	0
DT4	5	146	9/14	21	-7	49
DT7	5	181 (off scale)	15/21	21	0	0
DT12	10	144	6/11	21	-10	100

DT16	4	178	15/21	21	0	0
CC5	5	174	15/21	21	0	0
CC2	2 (off scale)	117	NA/21	21	0	0
CC3	4	148	10/21	21	0	0
CC7	5 (off scale)	162	12/(21)	21	0	0
CC10	5 (off scale)	179	15/(21)	21	0	0
CC13	2 (off scale)	170	15/(21)	21	0	0
CC17	5 (off scale)	163	12/17	21	-4	16
CC9	4 (off scale)	174	14/21	21	0	0
CC15	4 (off scale)	174	14/21	21	0	0
DT12	5 (off scale)	152	NA/14	21	-7	49
TOTAL	(68)	(2451)			$\sum d^2 = 214$	
Mean	4.53	163.4			Sd = $\sqrt{214/21}$ = $\sqrt{10.19}$	
Mean score	**Grade: 12 / Age: (21) (Age is off scale)**				**Sd = 3.19**	
Total Off Scale	**Variable: 10 (66.7%) Score: 2 (13.3%)**					

5.2.3 *English: Gunning-Fog Index*

Gunning-Fog Readability Index scores were calculated for the fifteen 100-word texts selected from each non-specialized English sub-corpus. The standard deviation of the individual text scores was then calculated using the mean of the total scores. The results are presented in Tables 15 and 16 below.

With Gunning-Fog, scores for the translated English sub-corpus have a standard deviation of 4.25, which is 1.58 points greater than 2.67, the standard deviation of the Gunning-Fog scores for the non-translated English sub-corpus. This shows that the Gunning-Fog Readability scores for the translated sub-corpus are more dispersed and therefore less homogeneous.

The translated sub-corpus has a "Ph.D level" Gunning-Fog mean score, signifying that it requires the reading skills of a highly educated adult. The translated sub-corpus as a whole thus has a different central tendency: its Gunning-Fog score (20) is two points higher than the Gunning-Fog score of the non-translated sub-corpus (18). The score of the translated sub-corpus is therefore higher by 10% of the possible score range for the Gunning-Fog Index.

However, it should also be noted that both of the non-specialized English sub-corpora have mean Gunning-Fog scores which indicate that their texts are too "hard to read" for their ostensible purpose. Both sub-corpora have mean scores that are well above the "target" Gunning-Fog difficulty level for mass audiences such as the Web surfers who are presumably addressed by the **GOL*** texts in question. Furthermore, since 15 is considered the maximum optimal Gunning-Fog score,[189] comparatively more (6.66%) of the individual TTs are actually somewhat easier to read by the standards of the Gunning-Fog readability index (see Tables 15 and 16). This is in keeping with the finding that the translated sub-corpus, with its higher standard deviation, is less homogeneous.[190]

The Gunning-Fog Readability Index rates each sub-corpus in terms of a pre-established continuum (that of readability). However, if we set aside the continuum as a reference, we can also view the sets of scores produced by each sub-corpus as score ranges that are "internally-generated" (Laviosa-Braithwaite 1996: 134-135). Counter to our predictions, the internally-generated ranges of scores for our TTs

[189] Optimal range is considered to be 11-15 for Gunning-Fog scores.

[190] In other words, the higher standard deviation reflects the divergence of the disparate individual texts in the translated sub-corpus, which, as a whole, scores "harder to read" (mean score = 20) in the Gunning-Fog readability continuum, but which also has more individual texts (73.33%) that are closer to the designated "optimum." These are two different ways of viewing the same thing: greater dispersion of scores in the translated sub-corpus.

move "toward the fringes" (Baker 1996: 177), instead of "gravitat[ing] toward the centre," as Baker puts it in her description of levelling-out (*ibid.*).[191] Our interpretation of the above findings will be discussed in Section 5.4.

Table 15
Gunning-Fog Index scores (Non-specialized EST)

	3+ syllables	ASL	Score	Mean	(d)	(d^2)
CB1	17	20	15	18	-3	**9**
CB3	21	25	18	18	0	0
CD2	22	33.3	22	18	4	16
CP11	20	16.66	15	18	-3	**9**
CP12	22	20	17	18	-1	1
CP13	28	25	21	18	3	9
CP17	13	20	13	18	-5	**25**
CP18	25	20	18	18	0	0
CP19	22	20	17	18	-1	1
CP20	27	20	19	18	1	1
PW5	42	20	25	18	7	49
PW6	29	20	17	18	-1	1
PW6b	23	20	20	18	2	4
PW10	24	20	18	18	0	0
SC1	16	25	16	18	-2	4
Mean	(23.4)	(21.66)	18		$\sum d^2 = 129$	
					$Sd = \sqrt{129/18} = \sqrt{7.16}$	
				Sd = 2.67		
Percent of individual scores outside the maximum optimum of 15			**80%**			
Number of scores in the optimum Gunning-Fog range (11-15)			**3**			

[191] Considered purely from the point of view of the pre-established continuum, however, it could be argued that the translated sub-corpus does appear to show some evidence of a slight "equalizing effect" (Shlesinger 1989: *passim*). Recall that the range of readability scores considered optimal for the Gunning-Fog index is 10-15. As can be seen in Tables 16 and 17, the translated sub-corpus has one more score in this "optimal" range than does the non-translated sub-corpus. If the other descriptive statistics are disregarded, this could be interpreted as a relative "sliding" toward readability on the part of the translated texts.

Table 16
Gunning-Fog Index scores (Non-specialized ETT)

	3+ syllables	ASL	Score	Mean	(d)	(d^2)
CC2	23	49	29	20	9	81
CC3	13	36	20	20	0	0
CC5	14	23	15	20	-5	25
CC7	15	20.8	14	20	-6	36
CC9	20	24.75	18	20	-2	4
CC10	20	24	18	20	-2	4
CC13	21	55	30	20	10	100
CC15	20	25	18	20	-2	4
CC17	31	22.8	14	20	-6	36
CT1	23	32.33	22	20	2	4
DT2	18	21.4	16	20	-4	16
DT4	15	22.2	15	20	-5	25
DT7	37	20	23	20	3	9
DT12	33	10.6	17	20	-3	9
DT16	32	25.5	23	20	3	9
					$\sum d^2 = 362$	
Mean	22.33	27.49	**20**		Sd = $\sqrt{362/20}$ =$\sqrt{18.1}$	
				Sd = 4.25		
					TT +1.58	
Percent of individual scores outside the optimum of 15			**73.3%**			
Number of scores in the optimum range (11-15)			**4**			

5.2.4 English: Lix formula

A Lix score was calculated for all of the individual texts in each non-specialized English sub-corpus. Using WordSmith's WordList, we obtained the values of both of the variables of the Lix formula, (i.e. the mean sentence length

and the number of words that have six letters or more).[192] Using the mean Lix score for each sub-corpus, we subsequently calculated and compared their standard deviations (see Tables 17 and 18; the statistical workup for the Lix Readability scores is shown in Appendix VI).

The standard deviation of the individual Lix scores of the translated English sub-corpus (Table 18) was 1.28 points lower than the standard deviation of the individual Lix scores of the non-translated English sub-corpus (Table 17). This indicates that the Lix scores of the translated sub-corpus are comparatively homogeneous.

The Lix readability scores indicate that the translated English sub-corpus is harder to read than the non-translated sub-corpus by 2.80 score points, or 8% of the expected range of 35 points. The comparative central tendency (i.e. the overall degree of readability) of the translated sub-corpus thus appears to be closer to the "very difficult" to read extreme of the Lix readability continuum. See Section 5.4 for a discussion of our interpretation of the above findings.

Table 17
Lix Standard Deviation (Non-specialized EST)

	Score	Mean 64.44	(d)	(d^2)
1. CBEST1	58.57		-5.87	34.45
2. CBEST2	66.41		1.97	3.88
3. CBEST3	66.48		2.04	4.16
4. CDEST2	73.27		8.83	77.96
5. CPEST7	76.82		12.38	153.26
6. CPEST8	60.78		-3.66	13.39
7. CPEST11	64.75		0.31	.09
8. CPEST12	53.71		-10.73	115.13
9. CPEST13	54.76		-9.67	92.92

[192] The exact formula, which has been adapted for use in corpus-based study, was as follows: LIX = (ASL) + 100 x (Number of long words/Number of words). See Section 5.1.4.

10. CPEST15	61.88		-2.56	6.55
11. CPEST16	61.49		-2.95	8.7
12. CPEST18	69.08		4.64	21.52
13. CPEST19	56.45		-7.99	63.84
14. CPEST20	67.19		2.75	7.56
15. PWEST1	72.05		7.61	57.91
16. PWEST3	82.44		18.00	324
17. PWEST4	76.45		12.01	144.24
18. PWEST5	84.97		20.53	421.48
19. PWEST6	66.73		2.29	5.24
20. PWEST8	52.47		-11.97	143.28
21. PWEST9	47.35		-17.09	292.06
22. PWEST10	62.49		-1.95	3.80
23. PWEST11	52.93		-11.51	132.48
24. SCEST1	57.24		-7.2	51.84
Mean	64.44			2179.74: **Sd 5.81**

Table 18
Lix standard deviation (Non-specialized ETT)

	Score	Mean	(d)	(d^2)
1. CCETT2	65.35	67.24	-1.89	3.57
2. CCETT3	60.92	67.24	-6.32	39.94
3. CCETT5	59.45	67.24	-7.79	60.68
4. CCETT6	60.94	67.24	-6.3	39.69
5. CCETT7	61.05	67.24	-6.19	38.31
6. CCETT8	70.08	67.24	2.84	8.06
7. CCETT9	67.02	67.24	-0.22	.04
8. CCETT10	49.45	67.24	-17.79	316.48
9. CCETT11	69.09	67.24	1.85	3.42
10. CCETT12	82.92	67.24	15.68	245.86
11. CCETT13	71.02	67.24	3.78	14.28
12. CCETT14	70.52	67.24	3.28	10.75
13. CCETT15	77.11	67.24	9.87	97.41
14. CCETT17	65.35	67.24	-1.89	3.57
15. CCETT18	73.03	67.24	5.79	33.52
16. CSETT3	59.95	67.24	-7.29	53.14
17. CTETT1	70.36	67.24	3.12	9.73
18. DTETT1	66.87	67.24	-0.37	0.13
19. DTETT2	63.92	67.24	-3.32	11.02
20. DTETT3	66.23	67.24	-1.01	1.02
21. DTETT4	65.16	67.24	-2.08	4.32
22. DTETT5	72.63	67.24	5.39	29.05
23. DTETT6	71.38	67.24	4.14	17.13

24. DTETT7	73.45	67.24	6.21	38.56
25. DTETT8	71.56	67.24	4.32	18.66
26. DTETT11	68.28	67.24	1.04	1.08
27. DTETT12	52.31	67.24	-14.93	222.90
28. DTETT13	67.22	67.24	-.02	.0004
29. DTETT16	74.01	67.24	6.77	45.83
30. DTETT17	70.71	67.24	3.47	12.04
Mean	67.24			**1380.19** **sd 4.53 ETT-1.28**

5.2.5 *French: Henry-de Landsheere index*

The scores obtained using the Henry-de Landsheere Readability Index are presented in Tables 19 and 20. See Appendices VII and VIII for the complete Henry-de Landsheere Index workup.

Table 19
Henry-de Landsheere scores and standard deviation (Non-specialized FST)

	Score	Mean	(d)	(d2)
Text 1	39	41.33	-2.33	5.42
Text 2	37	41.33	-4.33	18.74
Text 3	43	41.33	1.67	2.78
Text 4	42	41.33	0.67	.44
Text 5	44	41.33	2.67	7.12
Text 6	43	41.33	1.67	2.78
Total	248	41.33		$\sum d2 = 37.28$
MEAN	41.33			
Standard deviation				$Sd = \sqrt{37.28/41.33}$ $= \sqrt{0.90}$ Sd = .94

Table 20
Henry-de Landsheere scores and standard deviation (Non-specialized FTT)[193]

	Score	Mean	(d)	(d2)
Text 1	36	40.83	-4.83	23.32
Text 2	47 ("trop facile")	40.83	6.17	38.06
Text 3	41	40.83	0.17	0.02
Text 4	47 ("trop facile")	40.83	6.17	38.06
Text 5	43	40.83	2.17	4.70

[193] Extreme scores are noted (e.g. "trop facile"; "trop difficile"). The non-translated texts had no extreme scores.

Text 6	31 ("trop difficile")	40.83	-9.83	96.62
Total	245			$\sum d2 = 200.78$
MEAN	40.83			
Standard deviation				$Sd = \sqrt{200.78/40.83}$ $= \sqrt{4.19}$ $Sd = 2.04$
Difference translation				TT +1.1

The non-specialized French translated texts have Henry-de Landsheere scores with a standard deviation of 2.04, which is 1.10 points greater than that of the comparable non-translated texts (sd: 0.94). Since this indicates that the Henry-de Landsheere scores for the translated texts are more dispersed compared to those of the non-translated texts, it follows that the scores of the translated texts are less homogeneous.

Judging from the mean Henry-de Landsheere scores for each non-specialized French sub-corpus (non-translated 41.33; translated 40.83), the central tendencies of the two sub-corpora—their levels of readability—are very similar. Furthermore, the difference between the two mean scores (0.05 score points) is less than 1 percent (0.83%) of the possible score range of the entire Henry-de Landsheere index (i.e. less than 1 percent of the Henry-de Landsheere index range of 10 to 70 points).[194]

However, half of the individual text scores of the translated sub-corpus (i.e. the scores for French TTs #2, #4, and #6) have extreme scores that are outside the optimal range, being interpreted by the standards of the Henry-de Landsheere score as either "too easy" or "too difficult" for an adult population to read.[195] In contrast,

[194] Vandendooren's interpretation scale for the Henry-de Landsheere Index (see Section 5.1.5) gives a lowest possible score of 10, and a highest possible score of 70. The range of potential score points is thus 60.

[195] It should be recalled that by the standards of the Henry-de Landsheere index, the optimal range (of "bonne lisibilité") is considered to be 35-45 score points. See Section 5.1.5.

none of the non-translated text samples has an extreme score; all of the non-translated texts score within the optimum "good to perfect" readability range for this index. Against the standard of the continuum established by the possible range of Henry-de Landsheere scores (i.e. 10 to 70), the French translated texts have more scores located in the extreme ranges, while all the scores of the non-translated texts are within the middle, optimum range (i.e. 35 to 45).

With the Henry-de Landsheere formula, we have thus found that the translated non-specialized texts have with scores that are close to those of the comparable non-translated texts, but that they have arrived at that mean along a "bumpier" path: individual translated texts tended to have more extreme values compared to individual non-translated texts.

There initially appeared to be no feasible way to obtain scores for every individual text in each 16,000-word French sub-corpus, but we did develop a way of partially "automating" the calculations so as to obtain scores for each sub-corpus as a whole. We managed this by using the features of two different software programs.

WordSmith's WordList feature automatically calculates mean sentence lengths, and the search function of Microsoft Office Word 2000 quickly finds those "dialogue indicators" that are punctuation marks.[196] Instead of manually counting the words that are "absent" from the modified Gougenheim list ("les absents du Gougenheim," or AG), one can obtain an accurate "alternate" AG by first adding up

[196] WordList counts the number of dialogue indicators that are pronouns, but our copy of the software would not locate exclamation marks or *guillemets*. The search function of Word 2000 quickly finds these dialogue indicators, however, so we were able to count them by clicking through the corpus.

the number of tokens for what we would take the liberty of calling "les *présents* du Gougenheim": the words in each sub-corpus that *do* appear on the Gougenheim list.

This is done by comparing each of the words on Gougenheim's list to an alphabetical Frequency List of the words in the corpus,[197] adding up the total number of tokens of the words that match, and then subtracting this number from the total token count, to obtain an initial count of AGs. Since proper names are excluded from the count, the total number of proper names in the corpus must then also be counted, and their number subtracted from the initial count of AGs obtained above.

This partially automated calculation of AG as a percentage can be formulated as follows:

AG = T x (100 ÷ Tokens), where T = (tokens – proper names) – "présents"

The workup for the partially automated calculation of Henry-de Landsheere readability scores is shown in Appendix VIII. By following the above procedure, we were able to obtain scores for each non-specialized French sub-corpus as a whole.

The results were as follows. The non-translated French sub-corpus (15,947 words) had an "automated" Henry-de Landsheere score of 46 (i.e. slightly "easier to read"); the translated French sub-corpus (15,977 words) had an "automated" Henry-de Landsheere score of 45, or one point lower (i.e. slightly "harder to read"). The difference is 1.66% of the possible score range.[198]

---

[197] Obtained using WordSmith's WordList function.

[198] The score of 45 for the translated sub-corpus is approximate: because the AG was off the scale for Secondaire supérieur with a DEXGU of less than 1 (and therefore counted as 0), we had to extrapolate the curve.

Recalling that a text with a score higher than 45 is considered "trop facile" (too easy) to read, we note that the scores for each sub-corpus as a whole show a slightly different central tendency compared to the cumulative mean scores for the six 250-word samples in each sub-corpus (mean sample scores: non-translated 41.33; translated 40.83).

The above "automated" calculation places the central tendency of each sub-corpus in the liminal space between "bonne lisibilité" and "trop facile," while the manual count with the shorter samples places both sub-corpora well within the "bonne lisibilité" range of scores.

5.2.6 French: Lix formula

This user-friendly readability instrument has been validated for (European) French, allowing us to apply it to our non-specialized French sub-corpora. The mean sentence length, and the number of long words (i.e. words that have six letters or more), were both obtained with WordSmith's WordList feature. Using these values, the Lix score was calculated for all individual texts.[199] For each non-specialized French sub-corpus, the mean Lix scores and their standard deviations were calculated. The results are presented in Tables 21 and 22. The calculation of the Lix Readability scores is shown in Appendix VI.

Table 21
Lix (Non-specialized FST)

	Score	Mean	(d)	(d^2)
1. CCFST2	84.66	70.27	14.39	207.07
2. CCFST7	68.06	70.27	-2.21	4.88
3. CCFST9	77.21	70.27	6.94	48.16
4. CCFST11	76.21	70.27	5.94	35.28
5. CCFST12	93.47	70.27	23.2	538.24

[199] The formula was as follows: LIX = (ASL) + 100 x (Number of long words/Number of words). See Section 5.1.4.

6. CCFST13	65.52	70.27	-4.75	22.56
7. CCFST14	71.69	70.27	1.42	2.01
8. CCFST15	74.40	70.27	4.13	17.05
9. CCFST18	76.53	70.27	6.26	39.18
10. CSFST3	67.54	70.27	-2.73	7.45
11. CTFST1	65.42	70.27	-4.85	23.52
12. DTFST1	67.20	70.27	-3.07	9.42
13. DTFST2	61.99	70.27	-8.28	68.55
14. DTFST3	65.00	70.27	-5.27	27.77
15. DTFST4	69.09	70.27	-1.18	1.39
16. DTFST5	71.30	70.27	1.03	1.06
17. DTFST6	73.27	70.27	3	9
18. DTFST7	66.16	70.27	-4.11	16.89
19. DTFST8	69.67	70.27	-0.6	.36
20. DTFST9	73.62	70.27	3.35	11.22
21. DTFST10	64.11	70.27	-6.16	37.94
22. DTFST11	61.39	70.27	-8.88	78.85
23. DTFST12	54.01	70.27	-16.26	264.38
24. DTFST13	63.61	70.27	-6.66	44.35
25. DTFST14	60.72	70.27	-9.55	91.20
26. DTFST16	77.24	70.27	6.97	48.58
27. DTFST17	76.32	70.27	6.05	36.60
28. PWFST2	63.77	70.27	-6.5	42.25
29. PWFST3	78.72	70.27	8.47	71.74
	70.27 (mean)			**1806.95** **sd = 5.07**

Table 22
Lix (Non-specialized FTT)

	Lix	Mean	(d)	(d^2)
1. CBFTT1	73.84	73.12	.72	.51
2. CBFTT3	71.06	73.12	-2.06	4.24
3. CCFTT1	101.49	73.12	28.37	804.85
4. CCFTT3	54.73	73.12	-18.39	338.19
5. CCFTT5	60.01	73.12	-13.11	171.87
6. CCFTT6	94.71	73.12	21.59	466.12
7. CCFTT7	128.15	73.12	55.03	3028.30
8. CCFTT9	66.86	73.12	-6.26	39.18
9. CCFTT10	78.82	73.12	5.7	32.49
10. CDFTT1	86.05	73.12	12.93	167.18
11. CDFTT2	74.25	73.12	1.13	1.27
12. CPFTT1	75.38	73.12	2.26	5.10

13. CPFTT3	79.18	73.12	6.06	36.72
14. CPFTT4	75.36	73.12	2.24	5.01
15. CPFTT5	65.67	73.12	-7.45	55.50
16. CPFTT6	63.21	73.12	-9.91	98.20
17. CPFTT8	67.16	73.12	-5.96	35.52
18. CPFTT12	64.04	73.12	-9.08	82.44
19. CPFTT14	62.76	73.12	-10.36	107.32
20. CPFTT15	41.22	73.12	-3.19	1017.61
21. CPFTT16	93.94	73.12	20.82	433.47
22. PWFTT1	69.57	73.12	-3.55	12.60
23. PWFTT9	55.25	73.12	-17.87	319.33
24. PWFTT11	58.62	73.12	-14.5	210.25
25. SGFTT1	66.77	73.12	-6.35	40.32
	73.12 (mean) **TT +2.85**			7513.59 **sd 10.13** **FTT+5.06**

The standard deviation of the individual Lix scores of the translated texts (Table 22) was 5.06 points higher than the standard deviation of the individual Lix scores of the comparable non-translated texts (Table 21). This shows that the scores of the translated sub-corpus are on the whole comparatively dispersed, and therefore less homogeneous.

The Lix readability scores indicate that the translated French texts are harder to read than the comparable non-translated texts by 2.85 score points, or 8.14% of the expected range of 35 points. The comparative central tendency (i.e. the overall degree of readability) of the translated sub-corpus thus appears to be closer to the "very difficult to read" extreme of the Lix readability continuum. See Section 5.4 for our interpretation of the above findings.

5.2.7 *French: ICA Formula*

This formula was developed as a tool for assessing the readability levels of documents written for the general public by Canadian actuaries (Tremblay 2000: 1). The expected score range is between 6 and 13 (seven points). Scores above 13 correspond to the level of reading difficulty of specialized journals ("revues spécialisées" Tremblay 2000: 2). A score of 9-10 on the ICA scale can be considered optimal, since it is labelled as the target score range (Tremblay 2000: 2-3).

The ICA formula is applied to text excerpts (samples) of 100 words each. As with most readability indices, samples are generally taken from several evenly-spaced parts of a document. We therefore took fifteen 100-word samples, five each from the beginning, middle, and end of the two non-specialized French sub-corpora (10% of the words per sub-corpus).

For each set of non-translated and translated text samples, the ICA Readability scores and their standard deviations were calculated. The results are presented in Tables 23 and 24. The text samples and the values used in the calculation of the ICA formula are presented in Appendix IX. [200]

Table 23
ICA scores and standard deviation (Non-specialized FST Samples)

Sample #	Score	Mean	(d)	(d^2)
1.	24.8	14.83	9.97	99.40
2.	12.51	14.83	-2.32	5.38
3.	16.7	14.83	1.87	3.49
4.	15.5	14.83	0.67	0.44
5.	14.4	14.83	-0.43	0.18

[200] For the ICA formula, recall that the variables are each sample's ASL, plus the number of words with four or more syllables that each sample contains. See Section 5.1.6.

6.	16.12	14.83	1.29	1.66
7.	10.66	14.83	-4.17	17.38
8.	23.2	14.83	8.37	70.05
9.	10.8	14.83	-4.03	16.24
10.	9.93	14.83	-4.9	24.01
11.	10.8	14.83	-4.03	16.24
12.	10.4	14.83	-4.43	19.62
13.	11.36	14.83	-3.47	12.04
14.	16.8	14.83	1.97	3.88
15.	18.53	14.83	3.7	13.69
TOTAL	222.51			303.7
MEAN	14.83			
Number of extreme high (+13) scores	8 (53%)			√ 303.7/14.83 = √20.47 = **Sd 4.52**
Number of "target" (9-10) scores	5			

Table 24
ICA Scores and Standard Deviation (Non-specialized FTT Samples)

Sample #	Score	Mean	(d)	(d^2)
1.	13.2	15.06	-1.86	3.45
2.	10	15.06	-5.06	25.60
3.	38.66	15.06	23.6	556.96
4.	10.5	15.06	-4.56	20.79
5.	14.26	15.06	- 0.8	.64
6.	11.1	15.06	-3.96	15.68
7.	11.6	15.06	-3.46	11.97
8.	14.8	15.06	-0.26	.006
9.	12	15.06	-3.06	9.36
10.	17.33	15.06	2.27	5.15
11.	10.33	15.06	-4.73	22.37
12.	10.4	15.06	-4.66	21.71
13.	17.46	15.06	2.4	5.76
14.	10.11	15.06	-4.95	24.5
15.	24.2	15.06	9.14	83.53
TOTAL	225.95			807.47
MEAN	15.06			
Number of extreme high (+13) scores	7 (47%) **TT – 6%**			Sd = √807.47/15. 06 = √ 53.61 = **Sd 7.32**
Number of "target" (9-10) scores	5			**TT + 2.8**

The ICA scores for the sets of translated samples have a standard deviation (7.32) that is 2.8 points greater than the standard deviation of the non-translated French samples (4.52). This shows that the ICA scores for the translated sub-corpus are more dispersed compared to those of the non-translated sub-corpus. It follows that the ICA translated French scores are less homogeneous.

As far as readability (the central tendency measured) goes, none of our samples has a score of "easy" or "very easy" to read on the ICA scale. Considering their non-specialized nature and the stated GOL goal of addressing the general public, these texts appear to be surprisingly difficult for the general public to read, by the standards of the ICA index. The mean scores for both sets of samples (non-translated and translated) are unambiguously in the "very difficult" ICA index range (see Section 5.1.6).

With ICA, there is only a very slight tendency for the translated samples to be more difficult to read than the non-translated samples. The mean scores (translated samples 15.06; non-translated samples 14.83) are quite close. The translated mean score is higher by less than a quarter (0.23) of a point, which is only .03% of the expected range of scores.[201] Note however that the calculation of the ICA index entails taking a relatively small total sample, and that the small size may be reflected in this correspondingly small range of scores.

As their central tendency shows, the internally-generated range of our two sub-corpora is skewed toward the upper extreme of the ICA readability continuum. The lowest-scoring sample texts from each sub-corpus (non-translated 9.93;

[201] It should be recalled that the expected range in values as indicated on the interpretation scale for the ICA index is 7 points. See Section 5.1.6.

translated 10) are on "target" readability, according to the ICA scale. The total number of samples that score in the target range is exactly the same in the two sub-corpora (non-translated 5; translated 5, or about one third of the samples). When the scores that fall into the upper extreme range (of 13+ points) are compared, we see that there are eight among the non-translated samples and seven among the translated samples, a difference of about six percent (see Table 24). Approximately half of both the translated and the non-translated text samples are in the "very hard to read" range.

Nonetheless, when the top ICA scores are compared by sample set (top non-translated sample score 24.8; top translated sample score 38.66), we see that there is one translated sample with a score that is 13.88 points higher than the top score for the non-translated samples. This difference of nearly 14 points is double the expected score range, a high extreme that is in keeping with the higher standard deviation of the translated sample set, which is clearly less homogeneous. See Section 5.4 for our interpretation of these findings.

5.3 Type/token/n Ratios and ASL: Standard Deviation

For the four overall corpora (which include both specialized and non-specialized texts) used in the present study, we calculated the standard deviation of the statistics generated by two of the simplification measures: type/token/*n* ratios and ASL (mean sentence length).[202] The results are summarized in Table 25 and Table 26. Both of the translated corpora (ETT and FTT) had lower standard

[202] As has been noted above (Section 2.4.2), Laviosa-Braithwaite (1996: 144, 146) found that compared to the non-translated texts in the part of her corpus that was non-fiction, the translated texts generally had scores with lower degrees of dispersion above and below their median value.

deviations than those of the non-translated corpora (EST and FST). We thus found that the standard deviation of the scores for translated corpora does appear to be comparatively more homogeneous. See Section 5.4 below our interpretation of these findings.

Table 25
Standard deviation, type/token/*n* ratios (All Texts)

	Number of texts	Total	Mean	Standard deviation
EST T/T/*350*	64	3324.95	52.77	7.93
ETT T/T/*350*	54	2753.22	51.94	7.28 **(ETT – 0.65)**
FST T/T/*250*	48	2512.33	53.45	6.85
FTT T/T/*250*	34	1798.19	54.49	5.74 **(FTT -1.1)**

Table 26
Standard deviation, mean sentence length (All Texts)

	Number of texts	Total	Mean	Standard deviation
EST	64	1483.8	23.55	7.93
ETT	54	1358.32	25.62	7.28 **(ETT – 0.65)**
FST	48	1351.88	28.76	6.85
FTT	34	940.48	28.49	5.74 **(FTT -1.1)**

5.4 Levelling-out: Test of Significance and Interpretation of Results

We used the the F-test (for comparing variances) to determine the statistical significance of our findings. The alternative (i.e. research) hypothesis is that σ^2_x / σ^2_y is larger than 1 (x>y). The F-value is the ratio of two sample variances. To support the research hypothesis, the F-value must be larger than 1.

In the case of the Fry scores, the F-value is over 1, but the p-value is .47 (larger than 1), so we retain the null hypothesis (see Appendix X). The Gunning Index F-value is less than 1, which does not support the research hypothesis, and we retain the null hypothesis with a p-value of 0.959131. For the English Lix scores, the F-value is above 1, and we can reject the null hypothesis at a fairly strong confidence level (higher than 90%; p-value 0.098032). With all three of the French indices (Henry-de Landsheere, ICA, Lix), none of the evidence supports the alternative hypothesis, and there is no significance. All of the F-values are less than 1, and all of the P-values approach 1 (a nearly 100% risk of the findings being due to chance). For all but one of the readability indices, then, the difference in variance was not acceptably significant, and we were unable to accept the alternative hypothesis (see Appendix X). Except for the findings of significance with the Lix index in English, these test statistics indicate that the results summarized below might have been due to chance. These results do not appear to support the research hypothesis predicting greater homogeneity for the scores of the translated texts. However, they leave open the question of the marked central tendency indicated by the scores. This is further discussed below.

In what follows, we will summarize the results for homogeneity and central tendency. On the basis of these results, we will then discuss the the interpretation we have made concerning our research hypothesis.

5.4.1 Homogeneity of the Score Sets

The comparative standard deviation of the translated non-specialized English text samples per readability index was as follows:

-0.07 Fry Readability Graphs (English)
+1.58 Gunning-Fog Readability Index (English)
-1.28 Lix (English)

The comparative standard deviation of the translated non-specialized French text samples was as follows:

+1.10 Henry-de Landsheere French Readability Index (French)
+2.80 ICA (French)
+5.06 Lix (French)

The type/token/*n* ratio and mean sentence length standard deviations of the translated "main" corpora (i.e. of all of the translated texts, specialized and non-specialized) were lower than the standard deviations of the non-translated overall corpora by exactly the same values:

-0.65 ASL and type/token/*n* ETT
-1.10 ASL and type/token/*n* FTT

The lower standard deviations of the translated texts can be interpreted as evidence of greater homogeneity among the translated texts, an apparent confirmation of our hypothesis.

5.4.2 Central Tendency of the Score Sets

The comparative central tendency per readability index of the translated non-specialized English text samples is presented in Table 27 below, where the designation of "harder to read" is represented by the plus (+) sign.

Table 27
Readability scores compared to index optimum (Non-specialized Texts)

Index	Optimal Index Range	Mean Score, Non-translated Texts	Mean Score, Translated Texts	Difference, Translated Texts (percent of range)
Flesch Reading Ease (English only)	60 to 70 (out of 100)	31 (hard to read)	25 (harder to read)	+ 5.7%

Fry Readability Graphs (English only)	Grade 12 (out of 15)[203]	16 (hard to read)	21 (harder to read)	+35.7%
Gunning-Fog Readability Index (English only)	11-15 (out of 20)	18 (hard to read)	20 (harder to read; maximum possible score)	+10%
Lix (English)	35-50 (out of a range of 20 to 55+)	64.44 (hard to read)	67.24 (harder to read)	+ 8%
Henry-de Landsheere French Readability Index	40 (out of a range of 10 to 70)[204]	41.33	40.83 (slightly harder to read)	+0.8%
ICA (French only)	9-10 (out of a range of 6 to 13+)	14.83 (hard to read)	15.06 (harder to read)	+8.1%
Lix (French)	35-50 (out of a range of 20 to 55+)	70.27 (hard to read)	73.12 (harder to read)	+0. 03%

There is an obvious (if unexpected) difference in both languages: the translated texts score as consistently harder to read, and this is true even where the mean score for both translated and non-translated texts is considerably more difficult than the Index optimal score (e.g. Gunning-Fog, Lix French). In other words, both from the point of view of the independently defined continuum of readability and from the point of view of the internally-generated score range, the translated texts in both languages have an observably consistent tendency to score "harder to read," contrary to our predictions.

5.4.3 Homogeneity: Interpretation Regarding Levelling-out

Standard deviation values that are comparatively low indicate lower dispersion, and therefore greater homogeneity. Negative values thus support the present study's hypothesis. Negative standard deviation values were found with

[203] This Fry Readability Graph optimum is designated for the purposes of the present study only; see Section 5.1.2.

[204] Somewhat counterintuitively, in the Henry-de Landsheere scale, a score of 10 is the absolute periphery of difficulty (10 = extremely difficult to read), while a score of 70 is the absolute periphery of ease of reading (70 = extremely easy to read).

the Fry Readability Graph scores and the Lix scores in English (the latter of which was statistically significant and therefore unlikely to be due to chance according to the F-test), and with the standard deviation of the mean sentence lengths and type/token/*n* in both languages.[205]

Conversely, positive values indicate the opposite of homogeneity: greater dispersion of score values, a result that fails to support the hypothesis of levelling-out. Positive values were obtained with the Gunning-Fog Readability Index in English, and with the Henry-de Landsheere French Readability Index, the ICA Index, and the Lix Index in French. Interestingly, none of these scores tested as significant, leaving open the possibility that they are due to chance.

Our overall results for standard deviation values can thus be interpreted as follows. The fact that standard deviation is comparatively lower for the translated English text score values (with the exception of the translated Lix English scores, which are statistically significant and therefore highly unlikely to be due to chance) appears to support our hypothesis of levelling-out, but also fails to test consistently as significant. However, the fact that the standard deviation of the translated French score values is comparatively higher (but statistically *not* significant) apparently fails to support our hypothesis while possibly being due to chance. Since the general motivation of the present study is to locate features that recur clearly and consistently in translated texts, regardless of their language, these findings prompt us to conclude that the present study's hypothesis of levelling-out is not supported.

---

[205] For tests of significance with type/token ratios and mean sentence lengths, see Chapter 4.

5.4.4 Central Tendency: Interpretation Regarding Readability

It should be recalled that readability index scores indicate the level of skill required for reading a given text (see Section 2.4.2). The hypothesis of levelling-out as formulated for the purposes of the present study (see Section 1.6.4) predicts that translated texts will test out as more "readable": they will tend as a group to have more readability scores in the designated optimal range of a given readability index. It was predicted that translated texts would generally score in the central range of the readability continuum, and that non-translated texts would in contrast tend to score more toward the periphery (i.e. that non-translated texts would score as either "easier" or "harder" to read than would translated texts).

Our overall results for the central tendency of readability show the contrary: translated texts appeared to show a general tendency to score farther away from the designated optimal score or score range than did non-translated texts. It follows that translated texts could be construed as tending to score more often toward the periphery of a given readability continuum, counter to our predictions.

The results for central tendency can thus be interpreted as not supporting the present hypothesis of levelling-out. However, where central tendency is concerned, our translated texts apparently tend to be more difficult to read in both languages. We may have stumbled upon a new recurrent feature, one not predicted in Baker's original hypothesis. This would be an interesting topic of future study.

Although it would be impossible to provide an adequate explanation for the differences in readability scores without more research, we might very tentatively speculate that the translated texts scored as "harder to read" because they are

perhaps somewhat more complex, having been "double-processed," first as written texts, and later as translated texts, and complexity may accrue with each re-writing.

6. Conclusion

In the present study, we have seen how discussion of what have been called "translation universals" has led to the opening up of a new field of research within the discipline of Translation Studies: the cross-language search for consistently recurring features of translated texts. This topic has recently been the subject of a number of pioneering studies.

To this field of research, we have contributed our investigation of all four of the "translation universals" (normalization, explicitation, simplification, and levelling-out) that were originally hypothesized in Baker (1996). Three major empirical studies (Kenny 1999b, Olohan and Baker 2000, and Laviosa-Braithwaite 1996), have each previously investigated one of these features in English, providing a methodological model for investigating normalization, explicitation, and simplification as proposed recurrent features of translation. The present study has built upon the methods used in these previous studies, applying them to materials of a type not yet investigated, namely non-literary Canadian texts in both English and French. The investigation carried out in the present study has furthermore been extended to the fourth feature proposed in Baker (1996): levelling-out, which to the best of our knowledge has not previously been the subject of systematic, corpus-based empirical study.

In what follows, we will summarize the findings of the present study, comparing them, where possible, with the results of the previous major studies carried out on this topic. We will assess the measures used in the present study, the

measurements they produced, and what they may say about the population of texts included in the present corpora. We will then list the possible contributions made by the present research to this field of study. We will go on to interpret our findings in the light of common assumptions about the nature of translation. Finally, we will discuss avenues for future research from three perspectives, considering (1) what our results say, and leave unsaid, about translation commonplaces, (2) what questions might be particularly appropriate for future research on this topic, and (3) what measures might possibly prove fruitful in further study of this topic.

6.1 Summary of findings

In the present study, the measure used to investigate normalization as a recurrent feature of translation (that is, the comparative number of transient coinages attempted in translated versus non-translated texts; see Section 4.1.1) appears to provide support for the hypothesis of normalization as it was formulated for the present research. There were fewer attempted coinages in our translated corpora, and the number of transient coinages found between translated and non-translated corpora in each language had exactly the same proportional differences.

The measure used to investigate explicitation as a recurrent feature of translation (that is, the comparative number of explicitating elements in translated versus non-translated corpora; see Section 4.2.1) appears to offer some evidence in support of the hypothesis of explicitation as it was formulated for the present research, although with the French measures, there is a risk that these findings are due to chance. There were generally (with one exception) greater numbers of

explicitating elements among our translated texts, in both languages. An unexpected finding was that in specialized French, the frequency of occurrence of the impersonal pronoun *on* appeared to distinguish translated from non-translated texts in our Canadian corpus; this might be a useful measure for future study (see Section 6.5).

The three measures used to investigate simplification as a recurrent feature of translation (type/token/*n* ratio, lexical density ratio, and mean sentence length, see Section 4.3.1) do not appear to offer consistent evidence in support of the hypothesis of simplification as it was formulated for the purposes of the present research. The results appear to depend on the vocabulary and grammar of the particular languages involved, and not on the translated or non-translated status of a corpus. In other words, attempting to observe simplification through these measures instead produced evidence of apparent distinctions between the languages, rather than between the translated and non-translated corpora.[206] These results suggest that if simplification is indeed a consistently-recurring feature of translation, it is not seen solely in impoverishment of vocabulary, reduced information loads, or simpler sentence construction. In future research, it would be advisable to devise other measures more likely to be "universal."

Standard deviation of readability index scores does not appear to offer evidence in support of the hypothesis of levelling-out as it was formulated for the present research. The results suggest that there may on the contrary be greater

[206] It should be recalled that higher type/token/n ratios indicate larger overall vocabularies, while higher lexical densities indicate higher information loads. In light of Hunt's (1977) T-unit theory, longer ASL (mean sentence lengths) indicate more complex syntax. The findings thus vary by language: vocabulary range is smaller in (translated) English but larger in (translated) French, information load is lower in (translated) English but higher in (translated) French, and syntactic complexity is higher in (translated) English but lower in (translated) French.

dispersion, not greater homogeneity, among translated texts in our Canadian corpus. However, the type/token/*n* ratios and mean sentence lengths of the translated texts appeared to be slightly more homogeneous in both, although since the difference was only about one standard deviation point in either language, this evidence is not conclusive. We do not believe that the process of translation has had a "levelling-out effect" on the texts included this particular Canadian corpus.

This part of the study offered unexpected findings. The overall readability index scores demonstrated that our translated texts consistently scored as "harder to read" in both languages. Additionally, the translated texts in both languages had "internally-generated" score ranges that included more anomalous scores (i.e. more scores in the "very hard" or "very easy" to read peripheries). Finally, when type/token/*n* ratios and mean sentence lengths were used to investigate levelling-out (in addition to their primary use in the present study as measures of simplification, see Section 4.3), our translated texts proved to have more homogeneous vocabulary ranges and mean sentence lengths. These findings point to interesting topics of future study (see Section 6.5).

6.2 *Test of significance and interpretation*

F-Tests and Z-tests showed that our findings for normalization in both languages and for explicitation in English were statistically significant, that is, unlikely to be due to chance. The statistical picture for Simplification was clear even though our findings were mixed: those findings that appeared to confirm the research hypothesis were not statistically significant, leaving open the possibility that they were due to chance, while the findings that apparently failed to confirm

the research hypothesis were statistically significant and therefore unlikely to be due to chance. All of the readability indices showed the same central tendency, namely that our TTs tend to require a higher level of skill on the part of the reader.

With Levelling-out, the findings were statistically significant for only one of the indices, and in one language, leaving open the possibility that the findings for five out of six of the indices (and for all those indices that were applied to French) were due to chance. Given these facts, we are prompted to conclude that levelling-out as a concept is probably artificially distinguished from the hypothesized recurrent feature of normalization.

The findings discussed in the previous section add new evidence to the pool of corpus-based knowledge (which is admittedly still very small) concerning characteristics that may distinguish translated texts. The corpora previously studied in the literature consisted mainly of literary texts, and included English as the only language of translation. As is appropriate for a study of features thought to be "universal"—true of translation in any language—evidence has now been added from corpora in an additional European language of translation, and from a different part of the world, namely Canada.

Evidence has been found to support two of the four hypotheses in the present study. The present study's hypotheses that normalization and explicitation are recurrent features of translation into both English and French in our corpora appear to be supported. However, the hypotheses that simplification and levelling-out are recurrent features of translation into both English and French in our corpora appear to be unsupported.

The findings of the present study add support to those of Kenny (1999b) that normalization may be a recurrent feature of translation, and also lend further support to the findings of Olohan and Baker (2000) that explicitation may be a recurrent feature of translation.

However, it remains to be explained why the findings concerning simplification as a recurrent feature of English and French translation do not support those obtained by Laviosa-Braithwaite (1996) with English texts translated from multiple SLs.

The present study's hypothesis that levelling-out is a recurrent feature of translation also appears to be unsupported, although testing this hypothesis has led to the discovery that the translated texts in our English and French corpora appear to have a strong tendency to be "more difficult to read" by the standard of various readability indices.

In what follows, we will summarize how we have interpreted the results for each of the hypothesized recurrent features of translation. Specifically, we will discuss our inferences—our understanding of what the results seem to say about the features of the texts included in the present study—in the light of our experiences with the corpora and the measures used.

6.2.1 Normalization

The results with the investigation of normalization in the translated texts are tentatively interpreted as evidence that greater "conservatism" may be a norm in our sample of Canadian translation, and that the translated texts may therefore tend to be normalized regardless of language. It is also possible that there are normative

differences: we interpret the fact that there were significantly more attempted coinages in English than in French as evidence that the two languages studied may have different sociolinguistic norms concerning the acceptability of coining words.

As mentioned above, the findings appear to support the present study's hypothesis of normalization. The translated texts in the present corpora do seem to have wording that is normalized (that is, for the purposes of the present study, wording that has been kept more conservative) compared to the non-translated texts, at least in terms of the number of coinages ventured. On the basis of these findings, we very tentatively infer that, in the translation community whose work is sampled in the study, a subjective social and behavioral norm of greater linguistic conservatism may have led to an observably lower frequency of attempted coinages in the translated corpora.

An objection might be raised that the relatively few attempted coinages found in the corpora could all be the work of one individual writer or translator.[207] Unlike its predecessor (Kenny 1999b), the present study is not designed in such a way as to allow a direct link to be made between the translated text and its producer. In Kenny (1999b), the author and translator of each of the literary source texts and target texts was known. In the present study, as noted in Section 3.3, "Corpus compilation," most of the authors (with very few exceptions) are anonymous, and none of the translators can be identified.

This means that it is theoretically possible, although unlikely, that all of the approximately 60,000 words gathered in each corpus were produced by one single

[207] There are, as noted in earlier chapters, about 50 texts in each corpus (see Chapter 3), but far fewer than 50 coinages in each corpus (see Chapter 4).

author or translator working for a number of different government departments (see Section 3.1). However, as noted in Section 4.1, we were able to check the general distribution of the attempted coinages in each corpus (using the WordSmith Tools Distribution Plot function), and could verify that the attempted coinages were all distributed across a reasonable number of texts.

Therefore, even if the attempted coinages found in the present corpora had been produced by a single author and a single translator in each language, our findings concerning the differences between translation and non-translation would still hold, since the theoretically-possible "single" translator would still have ventured to coin considerably fewer words than the (theoretically also possible) single institutional writer in each language.

It should be noted that the type of text gathered is no doubt a factor in the overall low numbers of transient coinages found. The present corpora are all non-literary texts posted to Government of Canada Web sites for public consumption. As such, the texts included in the present study cannot be considered to be as generally "creative" as literary texts would probably be. It follows that specific features of the texts included in the present study, such as their wording, would also be less "creative," and that the small overall number of transient coinages found in the present corpora is to be expected.

6.2.2 Explicitation

The results with the investigation of explicitation in the translated texts are tentatively interpreted as evidence that regardless of the register of the texts in question, the translations sampled in the present study may tend to contain more

elements of syntactic disambiguation. As mentioned above, these findings appear to support the our hypothesis of explicitation. There do appear to be a greater number of some of the optional syntactic elements observed (i.e. optional *that*, *which*, and *ne*) in both the English and the French translated texts, as predicted in the study. On the basis of these findings, we tentatively infer that, in the translation community whose work is sampled in the present corpora, there may be a tendency to adopt the general strategy of shifting translated texts toward greater syntactic disambiguation by "spelling out" more elements of sentence structure. This tendency may have led, in turn, to the observable use of more optional syntactic elements in the translations.

There may be two reasons why explicitation is observable in our translated corpora. First, a source text is usually analyzed before a target text is produced.[208] There is therefore a certain amount of cognitive "pre-organization" of a translation, an analytical step that is not necessarily taken during the writing of an "original" text. Since a translated text has necessarily been written twice (first when it was formulated in the original, and second when it was re-written in translation), it may be that conventional markers of written style tend to rise, not fall, in translation.

Second, translation programs at Canadian universities emphasize formal training in grammar and writing style, while such training is currently lacking in

---

[208] The breadth of the analysis probably varies, depending on translator's training and experience. The claim is often made in the literature that student translators are more likely than trained, experienced (and paid) professionals to analyze a text narrowly, point by point, rather than by taking the whole of the text into account. See for instance Baker (1992), Chesterman (1998), Delisle (1993), Gerloff (1986), Gile (1995), González Davies *et al.* (2001), Hervey and Higgins (1992), Ivanova (1998), Jääskeläinen (1993), Jääskeläinen and Tirkkonen-Condit (1991), Larson (1998), Lörscher (1991), Kussmaul (1995), Nord (1997), Poirier (2003), Robinson (1997c), Séguinot (1991), and Tirkkonen-Condit, ed. (1991). Nonetheless, the fact remains that on some level, the source text must be analyzed before a target text can be produced, and that the knowledge gained in analysis is incorporated into the target text. Furthermore, this analytical step is not involved in the writing of an "original" text.

many other Canadian university programs.[209] It follows that non-translators (who include highly trained policy analysts and subject specialists) may have received fewer hours of formal training in writing skills. In short, differences in training may be reflected in differences in surface features of texts.

We have thus identified two possibly "explicitating" contextual factors: the re-processing necessary for translation, and the specific training of Canadian translators.[210]

The objection might be raised that the present findings should be considered indicative, not of syntactic explicitation, but rather of differences in register among the translated and non-translated texts. After all, in manuals prescribing the rules of grammar and written composition in each language, a more formal overall text register is often identified as being marked by the use of those very syntactic elements whose frequencies have served as measures of explicitation in the present study (see Literature Review).

However, it is unlikely that the results of this investigation could have been produced by chance differences in register among the individual texts included in the study. Since the non-translated and translated texts included in the present study were directed at the same type of target audience (i.e. the Canadian general public), it is probable that they have a comparable range of register. Although it is difficult to objectively assign categories of register to individual texts, we can

---

[209] We are assuming, based on the gist of our correspondence with government employees during the gathering of our corpora, that some university education—at minimum a Bachelor's degree—is possessed by most of the government employees and contract workers who produce written texts (both translated and non-translated) for the Canadian government. See the Web sites of Canadian universities to compare the curricula of translation programs with those of most other programs in the Arts and Sciences. Journalism schools (e.g. Ryerson, Carleton) are an exception, in that they tend to emphasize hands-on writing exercises. As of 2004, Ryerson has added a basic grammar course to its first-year Bachelor's degree.

[210] As we suggest below, these two "explicitating" factors may also have played a role in the present study's findings for simplification as a recurrent feature of translation (see Section 6.2.3 below).

nonetheless state with a fair amount of confidence that most of the texts included in the present corpora are likely to have a reasonably similar range of register.

Furthermore, "register" can be considered synonymous with what Quirk *et al.* (1985: 25-27) call the speaker or writer's "attitudinal varieties," which range from "very informal" to "very formal" (1985: 27). These attitudes of register are "presupposed": they are driven by what Quirk *et al.* call " field of discourse," which is the type of activity that is "engaged in through language" according to the speaker or writer's "profession, training, and interests" (Quirk *et al.* 1985: 23-24). Examples of fields of discourse that "presuppose" register are scientific and technical writing, journalism, "bureaucratic" writing (1985: 24) or "legal" writing, such as "legal statutes" (1985: 25). These fields correspond to the styles of writing prevalent within various "professions"—that is, within various sectors of an economy.

If the text's register (i.e. the "attitudinal variety" and "field of discourse" of its writer) is indeed "presupposed" by the economic field in which its author produces it, then we may state with a fair amount of certainty that the texts included in the present corpora probably fall within the restricted range of register that corresponds to the formal end—"neutral" to "very formal"—of Quirk *et al.*'s range of register (1985: 27). The "attitudinal" range of the texts included in the present corpora is far narrower than the one proposed by Quirk *et al.* (1985: 27) for a set of "fields of discourse" found in a modern economy.

Precisely because our sample is designed to be narrow, the range of register of the texts included in the present study is narrow. The present corpora include sample texts taken from only one large main source. They are texts that have been

selected (if not written and/or translated) for publication by a single institutional "author." They are texts that are ostensibly aimed at a single mass audience, the Canadian general public (Government of Canada 2002). This is additional indication that their "authors" probably have "attitudes" (of the sort postulated by Quirk *et al.* 1985) that fall within a range that will tend to be restricted to the "more formal" end of the postulated continuum (from very informal to very formal) of attitude-driven register (Quirk *et al.* 1985: 27).

It is unlikely that the present corpora contain significant amounts of text that could be qualified as "very informal." The texts included in the present corpora are (insofar as register, as defined by Quirk *et al.*, is concerned) similar enough that they should, all other features being equal, contain similar numbers of such markers of formality as the optional syntactic elements used as measures in the present study. The fact that our findings show that the translated and non-translated texts do not contain similar numbers of these elements allows us, we would argue, to tentatively conclude that the features of the translated and non-translated texts are not identical, and that syntactic explicitation—as seen in the greater numbers of explicitating elements found in the translated texts included in present study—is a feature, present in both languages, of the translated texts included in this study.

It should also be noted that the present definition of text register (a definition that follows Quirk *et al.* 1985: 23-27) as a quality which is predicated on field of endeavor, rather than on a writer's idiosyncracies and personal style, precludes inferring that the findings should be dismissed as the possible work of single writers and translators.

It is true, as noted above (see Section 6.2.1) that the design of our corpora is such that it cannot be known how many writers and translators produced the texts studied. Nonetheless, even in the unlikely event that all of the texts in each of the four corpora should prove to be the work of a single freelancer working for all of the government departments represented, that lone freelancer would still, according to the definition provided by Quirk *et al.* 1985 (above), have to have produced texts with registers that varied according to their "field" (in this case, the department and profession or economic activity). In other words, since the field presupposes the register, it makes no difference, as far as register is concerned, how many people worked on the texts in our corpora. When those texts are translations, they are still likely to contain more instances of the explicitating elements identified.

Although it was extremely difficult to determine what role register had played in the frequency of these elements, the short "control tests" using the non-specialized sub-corpora, which consisted of those texts in our overall corpora that were least likely to be marked for formality, showed that removing the specialized texts did not decrease the apparent explicitation of the translated texts, but in fact appeared to increase it. The difference in frequency of use of the explicitating elements was in most cases even more marked among the non-specialized texts than in our overall corpora, allowing us to tentatively infer that translated texts which are not marked for register will also tend to be more explicitated than comparable non-translated texts.

6.2.3 *Simplification*

The results do not appear to support the hypothesis of simplification as formulated for the purposes of the present study. This finding is tentatively interpreted as evidence that compared to non-translated texts, the Canadian translations included in our study do not have comparatively restricted vocabularies, are no less densely packed with information, and have sentences that are no less complex.

We note that the findings are reversed in the two languages: translated English texts have lower vocabulary ranges and information loads, but longer, possibly more complex sentences, while translated French texts have larger vocabularies and information loads, but shorter and possibly less complex sentences.

There may be two reasons for these findings. First, their extensive formal training in writing skills may make it comparatively more likely that the Canadian translators whose work is included in our sample will employ a broad vocabulary, sustain a "dense information feed," and write complex sentences in a translated text. They may also tend (as a group) to exaggerate certain characteristics of the language in which they are writing. For instance, what has previously been interpreted as simplification of translation (Laviosa-Braithwaite 1996) may instead be an emphatic overall adherence to the stylistic norm of writing "plain and simple" English. Likewise, the above findings may reflect a preference in French for a large vocabulary (i.e. for "*la synonymie*"), and for sentences that are syntactically uncluttered but informationally dense. Second, the general cognitive "pre-processing" necessary for translation may also have contributed to an exaggeration

of style in each language. These two points would make interesting topics for future study (see Section 6.5).

6.2.4 *Levelling-out*

The results are interpreted as failing to consistently support our version of the hypothesis of levelling-out. Given a pre-established continuum, it appears that the Canadian-sourced translations used in our study may tend to produce sets of scores with fewer values close to the designated middle range of values, and with more values in the extreme range. They may also have more individual scores that diverge widely from their internally-generated mean.

We infer that the translated texts we have sampled possibly do not "level out," and have demonstrated that they will tend to score in ranges designated as "harder to read" than non-translations. This, we note, is in keeping with most of the other findings. More complex vocabularies and sentence structures (found in lieu of signs of simplification), and greater use of optional syntactic elements (found in the present corpora and interpreted as signs of explicitation), were features observed among the translated texts. It follows that these texts should require more advanced language skills in reading.

Furthermore, two of the features found in the translated texts—the longer mean sentence lengths (in lieu of simplified syntax) and the greater use of the optional elements (i.e. explicitation)—automatically entail a lengthening of the T-units (Hunt 1977) of the translated texts. By standards of readability theory that have long been accepted, lengthier T-units (defined as an independent clause plus any subordinate clauses attached to it) are harder to read, no matter what words they

contain or how they are otherwise structured. The translated texts included in the present corpora can therefore be inferred to be syntactically more complex (rather than simplified) because they test as harder to read.

The results for normalization—that fewer attempted coinages were found in translated texts—are in fact the only findings that would *not* tend to increase reading difficulty in texts. Since, according to the definition of readability (as it is understood by users of readability instruments), coined words are necessarily unfamiliar, they must figure among the list of words that are "harder to read" because they take slightly longer for the reader to process.

Finally, it must be acknowledged that the corpora of non-specialized texts used to investigate levelling-out are small by the current standards of corpus linguistics, even for special-purpose corpora. However, the non-specialized corpora are quite large by the standards of readability indices, many of which were designed to be run on only a few sets of texts of approximately 300 words.[211] We can therefore assume that the findings for readability (and therefore with levelling-out) are likely to be fairly accurate.

6.3 Contributions of this research

The present study has, we believe, fulfilled its chief objectives (see Section 1.3). Our specific objectives were to conduct an empirical study of recurrent features of translation within a broad scope of inquiry and, wherever possible, to refine the methodology used. These goals have, we believe, been met. We have conducted a systematic, empirical search for the hypothesized recurrent features of

[211] It should be recalled that the general practice is to take samples of 100 words from the beginning, middle, and end of a text that would be too long to process manually.

translation, and we have counted, qualified, and described the patterns found in the results as clearly as possible. Both our materials and our methods have broadened the scope of inquiry into this topic. Our corpora (the materials studied) have been gathered from sources outside the social, geographic, and linguistic boundaries represented by corpora used in previous studies. Our research, unlike that of previous studies, investigates not just one, but all of the "translation universals" originally hypothesized in Baker (1996). Although we have replicated three measures (type/token/*n* ratios, lexical density ratio, and mean sentence length) commonly used in corpus linguistics and in the field of readability instruments, we have also adapted Kenny's (1999b) measure of normalization for the purposes of the present study of non-literary texts. In addition, we have adapted the measurement of readability, through the use of the readability index, and put it to innovative use in the investigation of levelling-out in translated texts.

Our general goal was to contribute to the knowledge of those features which may universally distinguish translation from other types of writing, in order to advance the ongoing theoretical discussion, in Translation Studies, of the topic of "translation universals." This objective has also been met: new empirical evidence, observed first-hand in authentic texts, is offered by the present research as to whether and/or how the four hypothesized recurrent features of translation may manifest themselves in the texts we have studied.

Apart from meeting its objectives, the present study has made contributions to both theory and practice. To theory, the contribution is made in chiaroscuro: light has been shed on some areas, while others remain in darkness. Some of our findings are indeed negative, but these may be considered interesting in themselves.

Tymoczko (1998: 657) maintains that where a search for "universals" instead finds differences, this knowledge is also of value to Translation Studies researchers. With further study and increasingly enlarged corpora, separate cumulative "snapshots" may begin to merge into a global picture whose appearance we cannot yet begin to imagine. To this ongoing but very new research effort, all findings are potentially valuable.

Practical contributions have also been made. The usefulness of corpus-based methodology to Translation Studies has been demonstrated further. The English and French corpora gathered for the purposes of the present study could provide material for the teaching of translation, as well as for future research in corpus-based Translation Studies (Zanettin 2001: *passim*, Bowker 2002: 15-18, 20-21; 2003: 73, 75-76, Bowker and Pearson 2002: 137-221, Ulrych 2002: 199, and Granger and Petch-Tyson 2003: *passim*).

Furthermore, interdisciplinary uses may be found for the present corpora, including study of vocabulary, of grammar, and (if compared with similar corpora from other eras) study of language change and development. Our corpora could also be used in language instruction and the teaching of cultural studies (McEnery and Wilson 1996: 87-114, Botley *et al.*, eds. 2000: *passim*), as well as for cross-linguistic comparisons of languages, for comparative study of stylistics (Leistyna and Meyer 2003), for the development of software tools for corpus analysis (Aijmer and Altenberg 2004), and for the study of language variation (Mauranen and Kujamäki 2004).

6.4 Assumptions

A number of commonly held ideas about translation have formed part of the general intellectual context in which our hypotheses have been tested. Certain assumptions about the nature of translation follow from these ideas. Below, we will discuss these assumptions from the perspective of the present study.

Evidence of enduring notions about translation may be found in Robinson (1997), a collection of theoretical texts written on the nature of translation between the mid-fifth century B.C. and the late nineteenth century A.D. An overview of the implications of these notions for the practice of translation can be found in Munday (2001: 18-29) and in Chesterman (1995). Five key ideas, which "come up again and again" in the history of translation theory and practice, are listed in Chesterman (1997: 7-14). These assumptions are

(1) that translation is "directional," going from a source text to a target text,

(2) that a translated text is "equivalent" to a source text,

(3) that "untranslatability" makes such equivalence impossible, because meaning is inherent in the individual words of a text,

(4) that because this is so, translation will always be somewhat erroneous, being either "literal" (word-for-word; or what Toury might call "adequate") or "free" (what Toury might call "acceptable"),

(5) and finally, that all writing is derivative ("all-writing-is-translating": Chesterman 1997: 13-14), because meaning is negotiated, being the product of interaction among elements in a text, and among speakers.

It does not take much scrutiny to see that Chesterman's above five "supermemes" of translation (1997) form a continuum, a scale with opposing qualities ("original" and "non-original") at each end:

←Original – Equivalent – Imperfectly equivalent – Conventional→

In the present study, the notion of the "continuum" has been an important one. For instance, the idea that any given text will have the quality of being more or less readable is seen in a hypothetical readability range, from "very easy" to "very difficult" to read (Section 2.4). It was the existence of an independently-established "oral to literate" continuum that gave Shlesinger (1989) the original inspiration for the hypothesis that the process of translation may have an "equalizing" effect on texts; the present hypothesis of levelling-out was based in part on that inspiration.

Likewise, the set of "supermemes" of translation proposed by Chesterman (1997) forms another continuum that is of interest to the present study. Chesterman's first supermeme (that translation is "directional," going from a source text to a target text) assumes that there is direct transference: an "original" is being "reproduced," and the translation is a reproduction. Chesterman's fifth supermeme (that all writing is translating) similarly assumes that translation is the opposite of "original": translation is assumed to be "non-original." These supermemes are considered in theory to be ideas which co-exist, and which may therefore be put together (Chesterman 1997: 15-16; 48-49). When we put the opposing ends of Chesterman's continuum together, a key assumption underlying all of the above supermemes is revealed. This assumption may be stated as follows: "Translation is a copy of an original; translation is derivative."

From this metaphor of original and copy, four specific assumptions follow, namely that, quite like print copies made of an oil painting, translations must to some extent have been drained of the creativity, subtlety, complexity, and individuality of the original. These specific assumptions, we would argue, have formed part of the intellectual context in which the hypotheses tested in the present study were formed. Applying the results of the present study to each of these assumptions may therefore tell us something about them.

Assumption #1: Translations, being derivative, lack creativity.

Since there were in fact some coinages in both translated corpora, and since a coined word is necessarily a created word (however ephemeral or transparent-seeming), the results of the present study (see Section 4.1.2) show that this pervasive assumption may not be entirely true. However, it may not be entirely untrue, either: since more words were coined in each of the non-translated corpora, the results of the present study hint that the translated texts we have studied may tend to contain fewer examples of creative use of language.

Assumption #2: Translations, being copies of an original, are less subtle.

Since in both languages, the translated corpora contained higher numbers of "explicitating" syntactic elements, the results of the present study (see Section 4.2.2) tentatively indicate that this pervasive assumption may have some basis in truth. Connections (in the case of the present study, syntactic connections) that might have been left implicit may more often, in translation, be made plainly obvious.

Assumption #3: Translations are copies which have, compared to the original, lost some of their complexity.

The results of the present study (see Section 4.3.2) offer some evidence that this assumption, despite its ubiquity, may not be entirely true. At least where richness of vocabulary and overall syntactic structuring are concerned, one feature which may not be "lost in translation" is the complexity of the translated texts we have studied.

Assumption # 4: Translations, being derivative, exhibit uniformity.

Since in both languages, our translated corpora had significantly more scores that were close to one extreme end of the pre-established continuum of readability, and since they also generated a statistical middle range with more scattered scores and in more irregular patterns than were found in the non-translated texts, the results of the present study (see Section 5.4) seem hint that this pervasive assumption could be mistaken. Readability indices, which score texts according to what in Translation Studies has long been called the norm of acceptability, measure certain very basic and typical features of a language, such as mean sentence- and word- lengths. The readability scores obtained with both sets of translated texts appear to indicate that the assumption that the translations we have studied would exhibit a general tendency to be more uniform than texts that are not translated was possibly unfounded.

6.5 Suggestions for future research

In what follows, we will list a number of questions that have been left unanswered by the results of the present study. We will then go on to discuss how these unanswered questions might be investigated in future study of this topic.

6.5.1 Questions for future research

Limited time and resources restricted both the scope of the present research and the content of the present corpora. A number of closely related questions remain available for future study. These are listed and discussed below.

1. How could the hypotheses be refined?

Baker (1996: 180) notes that where hypothesized "translation universals" are investigated in corpora, "the process of refining the definition will go hand in hand with that of verifying the feature." This question is inherent in the present method; it is one that will continue to be asked, overtly or implicitly, at the end of all such studies. Much more evidence must be gathered before we can begin to assess whether the relatively small samples of language used in the present study have uncovered tendencies that truly reflect those of the larger population.

2. With improved measures, would the same results be found?

Iterative corpus-based studies of translation are likely to find many ways of improving on the measures used in the present study. One of the chief reasons for undertaking the present research was to make it possible for future researchers to ask this question.

3. With different samples, would the same results be found?

Again, this question is inherent in the corpus-based method: studies such as the present one are carried out in the hope that further research projects will continue investigating the hypothesized recurrent features of translation, using ever-larger corpora.

Ideally, a pooled international archive of translated and non-translated texts would be created at a number of research sites around the world. If a standardized protocol for recording the attributes of the archived texts were followed systematically,[212] the result would be an expanding matrix from which corpora of many different types could be extracted according to the needs of a given research project.

We wondered whether formal training in writing skills had played a role in our findings. Future studies might fruitfully gather corpora of texts produced by formally-trained translators in order to compare them with corpora produced by "self-taught" translators. Such study would be facilitated if the attributes recorded with each text in these corpora included the training background of the translator. The Student Translation Archive at the University of Ottawa (Bowker 2003: 170; Bowker and Bennison 2003), which records attributes in such a way as to allow the skills development of an entire cohort to be tracked throughout their training, would be of interest to the researcher wishing to isolate what might be called "universal translation skills" from skills specific to dealing with English, French, and Spanish source-language interference.

Both the effect of formal training in writing skills and the "pre-organization" of the text might be observed in translations whose surface features included more markers of "correctness" (normalization) and more disambiguating elements (explicitation). Normalization could thus be hypothesized to be an unconsciously-

[212] See Bowker (2003: 170) for an example of a list of attributes recorded along with each text in an archive. See also Bowker and Bennison (2003).

made display of writing skills, while explicitation could be hypothesized to be a subconscious reinforcement—a clarification—of the text's structure.

4. With other target languages included in the corpus design, would the same results be found?

Where the topic is "universal," the goal must ultimately be to cover the globe. Canada's multicultural communities offer an ideal potential resource for studies whose scope is widened to include many more languages. For instance, corpora of texts rendered into the translator's mother tongue might be compared with corpora of texts rendered into the translator's second language.

5. With multiple source languages included in the corpus design, would the same results be found?

Because the corpora were designed to include only one SL for each translated corpus in the present study, we were unable to estimate the possible influence of different SLs on each feature.[213] Research with corpora designed to include multiple SLs and one TL, (modelled, perhaps, on the corpus design used in Laviosa-Braithwaite 1996 or in Eskola 2004) might be suggested as one way of answering this question. An alternative might be to study multiple "bidirectional and parallel" corpora designed as in Bernardini and Zanettin (2004), or a combination of parallel corpus and comparable corpora, designed as in Pápai (2004).

6. With combined interpretation/translation corpora, will the same results be found?

Translation, the act of reporting that which has been said in another language, can be performed both orally and in writing. Shlesinger (1989: 14-15) suggests that

[213] We recognize that studies should reflect the growing context, especially in Europe, of translation from and into multiple languages. This invites a comparison between first- and second-language translation, that is, between translations made in the translator's mother tongue, versus translations made into a language that is not the translator's mother tongue.

the characteristics of a text are determined more by its "situational context" and by five "basic parameters" (degree of planning, shared knowledge, lexis [vocabulary], degree of involvement, and the role of non-verbal features) than by its medium or "channel" (that is, by whether it is spoken or written). To better understand the characteristics of translation as an act of language, it might at some point be well to remove the "oral/written distinction," taking large tagged samples of both interpreted (spoken) and translated (written) texts.[214] This would contribute to the ongoing search for translation-centred (rather than language-structured) differences between translation and other acts of language use.

Specifically, a Canadian legal corpus could be modelled on Shlesinger (1989): recordings could be made in Law Courts of both translated (simultaneously interpreted) and non-translated (verbatim) texts, which could then be transcribed, tagged, and investigated for recurrent features of the translated texts.

7. Are the features that have so far been identified artificially distinguished?

In the more recent literature on the topic of "translation universals" (e.g. Pápai 2004), there is some indication that the distinction among these four features is unnecessary: their delineation may not be as clear cut as depicted in those empirical research projects that have been carried out so far.[215]

Normalization and levelling-out appear to be concepts that may be linked, since both hypothesize that translations adhere more closely to norms (social or statistical) and therefore "resemble" one another more than do comparable sets of non-translations. In both hypotheses, it is predicted that the translated texts will in

[214] Collecting oral corpora requires much dedication, and we acknowledge that such a research project could not be undertaken without a sizeable budget and experienced staff.
[215] This criticism, although not (yet) aimed at the present study, would apply to it.

some way exaggerate their tendencies, and that this exaggeration will form a visible pattern.

Explicitation and simplification also appear to be concepts that are in some way related. As Pápai (2004) has pointed out, to repeatedly include an optional syntactic element in a text is to increase its redundancy (explicitation) while simultaneously lowering its type/token ratio (simplification).

As further links are uncovered, what are on these pages depicted as separate "features" may in future study begin to merge. To translate is to report content; to "author" (a non-translated text) is to think it up for the first time. It may be that when a text is first created, the cognitive effort of focussing on the topic results both in greater idiosyncrasy in the use of content words and in a greater willingness to make new ones up. When a text is re-written in another language, that is, when it is translated, it may be that the need to reduce ambiguity and render meaning as clearly as possible produces the set of traits that we have grouped under explicitation and simplification. Furthermore, a universal social need to make the foreign familiar, acceptable, and readable in translation may contribute to what we have called normalization.

In fact, we would very tentatively propose that all four of the hypothesized recurrent features of translation might, in some future study, be grouped under one heading, "standardization," with multiple measures implemented to reveal its many facets.

8. Are there more features than the ones identified so far?

All of the above-proposed expanded research projects would no doubt offer opportunities to explore the final, fascinating question of whether there might be more "translation universals" or recurrent features of translation. Such features would perhaps become evident in time, if the number of research projects continued to increase, allowing evidence on this topic to accumulate.

We see no reason why such studies should not continue to be undertaken iteratively, since the topic of recurrent features of translation is in its infancy, and since the material obstacles to its empirical investigation are steadily being dismantled.

The computer technology that originally put corpus-based study of translation within reach of individual (and possibly underfunded) researchers has, we note, improved considerably in the period of time (2001-2004) that it took for the present study to be contemplated, conceived, designed, researched, and written. The necessary hardware and software have undergone extensive improvement. The processing speed, memory, and storage capacity of the personal computer have all increased dramatically. Key "corpus crunching" software, such as the WordSmith Tools suite, has been made much more user-friendly, and would now require considerably less investment of initial training time on the part of the novice corpus researcher. Text analysis tools such as HyperPo and TAPoRware can help scholars who want to study the content (i.e. the meaning, rather than the surface linguistic features) of texts in large corpora. Easily-captured, public-domain data are now widely available on the Internet, both in English and in many other languages.

Furthermore, the increased ubiquity of communications technologies (cell phones, email, text messaging) is steadily increasing the means and ease with which texts can be transmitted, gathered, and shared.

6.5.2 *Measures for future research*

In general, when texts are gathered for future investigation of recurrent features of translation, we would suggest that a systematic protocol be followed for recording the attributes of each text. One particularly useful annotation would describe any technical writing tools used in the production of the text. This information would then be easily retrieved and sorted, allowing future researchers to note any correlation between the surface features of translated texts and the implementation of such technologies. For instance, with translated texts, it could be noted whether technologies such as translation memory, on-line dictionaries, or terminological databases had been used in the preparation of the document. With non-translated texts, it could be noted whether tools such as on-line dictionaries or document content management software (e.g. AuthorIT®) had been used. [216]

Since the use of translation memory is expected to become much more widespread in the Canadian translation community in the coming decades, the capacity of well-designed corpora to distinguish the features of texts translated with and without translation memory could provide a basis for very interesting research. It would also be interesting to test the idea that reliance on translation

---

[216] The practical functioning of document content management software is similar to that of translation memory: parts of "legacy" documents can be "reused" in new documents. Content is stored in "chunks," by topic, in a "relational database." See for example <http://www.author-it.com/index.mv?different> at Furl: http://www.furl.net/members/sissela.

memory will tend to foster repeated use of certain segments of text, the exaggerated presence of which might come to set translated texts apart.

Below, we will propose possible ways of measuring the specific features of normalization, explicitation, simplification, and levelling-out in future research.

1. Possible future study of normalization

In the present study, we postulated a continuum ranging from "normative" to "creative" in translated versus non-translated text vocabularies. In future study, it might be interesting to postulate a continuum that instead ranges from "conventional" to "unconventional." Transient coinages with a meaning as opaque as that of "alterned" and "deviatoric," both of which were found in the translated English corpus used in the present study (see Section 4.1.2), might be counted as occurrences ranking toward the "unconventional" extreme. The results of such a study might prove especially applicable in the field of translation pedagogy.

A diachronic study of attempted coinages in translation would be of interest. If properly designed, such a study might allow researchers to compare how many "viable coinages" (i.e. how many possibly permanent new words) were formed in corpora of translated versus non-translated texts, over a given period of time. Such a study would, of course, require very large and well-designed corpora (on the scale and of the quality of TEXTUM), not to mention the skills of at least one expert bilingual lexicographer.

Working with tagged corpora might make a number of new measures of normalization possible. Such measures could be inspired by the prescriptive rules given in grammars and style manuals against language use that is overtly labelled

as erroneous, such as split infinitives, punctuation errors, dangling prepositions, misuse of conjunctions, and in French, *anglicismes*. The comparative frequency of rule-transgression might show any diverging tendencies in translated versus non-translated texts.[217]

For investigation of stylistic normalization in translated corpora, it should be noted that Laviosa-Braithwaite (1996: 161) has proposed using KWIC concordancers and Mutual Information (a statistical measure of the association between two words in a text, also known as collocation: see Scott 2004) to look for ironic phrases, (semantic) prosodic clashes, and creative metaphors. Kenny (2001) is an ideal source of inspiration for future study of normalization and creativity in literary translation.

2. Possible future measures of explicitation

With a tagged corpus, it would be possible to compare the ratio of *that/which* ODC to "zero" ODC. Based on the findings of the present investigation, we speculated (Section 6.2.2) that the explicitation found in our translated corpora may reflect the fact that many Canadian translators are formally trained, and that this training may prompt them to make the construction of their sentences as explicit as possible. This explanation could be tested by comparing texts produced by formally trained translators with texts produced by "self-trained" translators, to see which group of texts contained more frequent use of explicitating elements.

[217] All the potential measures of normalization named here were searched in the present corpora. No differences were found using them, usually because few instances of these errors were found in either sub-corpus. However, in a larger corpus, these measures might prove more productive.

We had the unexpected finding (Section 6.2.2) that French specialized translations contained a very high number of occurrences of the impersonal pronoun *on,* compared to specialized non-translated French texts. This finding prompts us to suggest that it might be fruitful to study the comparative frequency of use, in specialized translated and non-translated texts (in a number of languages), of passive voice constructions, and of impersonal or indefinite pronouns (e.g. *on* in French, *one* and/or the authorial *we* in English, *den, det, dess* in Swedish). Such a study might help future research projects to distinguish among markers of explicitation, markers of formality, and markers of text specialization. Needless to say, such a study would be undertaken much more easily with a reliably tagged corpus. With a tagged French corpus, one might be able to find what is noted by Folkart (1991: 132-134) to be the "buttressed," deep-structure-echoing syntax that may be tangible evidence of syntactic explicitation in translation.

Any possible link between explicitation and register might be explored by correlating the use of optional *that* with contractions and other markers of informality, to determine whether translated corpora tend to use fewer such markers, as suggested in Olohan (2001: 429) and in Hansen and Teich (2001: 3).

Verb tenses may also have different patterns of use in translated texts, according to Chuquet (2003), who has suggested that translated English texts will contain more occurrences of the simple past. Specifically, Chuquet believes that the *imparfait* is overused in "contemporary" French (2003: 116) and maintains that it is systematically translated with the English simple past (116-118), leading to a "loss of indeterminacy" (115)—that is, to a "gain in explicitness" (111) in English

translated texts. Chuquet is assuming that the simple past will prove to be more frequent than the past perfect and past progressive in large corpora of English translated texts. This raises the question of whether different verb tenses might tend to be used with different frequencies in specialized and non-specialized translated texts in any language.[218] This question, as well as Chuquet's equation of "determinacy" with explicitation, could provide inspiration for future study.

3. Possible future measures of simplification

We have speculated above (in Section 6.2.3) that Canadian translators' formal training in writing may affect the degree of complexity of the translated texts that they produce. It would be interesting to compare type/token ratio, lexical density ratio, and mean sentence length in the work of formally trained translators versus that of translators who had not received such training.

We have also wondered whether the extra step of "pre-processing" a source text has an effect on the complexity of the translated text. Corpora of translated and non-translated first drafts, gathered and compared with translated and non-translated final drafts, might show differences in complexity at different stages of a translated text's development, making it possible to observe whether vocabulary and sentence structure become more or less complex with revision. Such a study would be especially interesting if a distinction between specialized and non-specialized texts were built into the corpus design, making it possible to determine whether vocabulary, style, and sentence structure vary by text type and content.

[218] However, if it were indeed discovered that one of the past tenses tended to dominate in translation, scholars interested in recurrent features of translation might interpret such a finding as evidence of a type of grammatical simplification, *contra* Chuquet (2003), rather than attributing it to syntactic explicitation.

By comparing translated and non-translated content word vocabularies (using lexical density ratio or some other measure involving content words) in tagged corpora, it might also be found that simplification was at least partially semantic (in other words, simplification might entail a loss of meaning), although this would be surprising, given the typically rich vocabulary of the bilingual or polyglot translator.[219]

4. Possible future measures of levelling-out

The field is wide open for future researchers to devise interesting new measures for this least-studied hypothetical feature of translation. Shlesinger's (1989: 14-42) various parameters of the oral-literate continuum appear to be a particularly interesting potential source of inspiration for new measures of levelling-out.

The findings from several areas of the present study could also be considered a basis for future study of levelling-out. We noted in Section 4.2.3 that specialized translated French texts had considerably higher occurrences of the impersonal pronoun *on* compared to specialized non-translated French texts. A future hypothesis of levelling-out as a recurrent feature of translation might include greater frequencies of impersonal pronouns as a measure of differences between specialized translated and non-translated corpora.

Given the differences found in "internally generated" score ranges versus scores that are compared to independently established reference ranges, future studies might do well to distinguish the hypothesis of levelling-out from that of

[219] Since translators are necessarily bilingual or polyglot, many probably possess total vocabularies (the sum of all the words they know in both/all their languages) that are larger than those of most monolinguals. See Kolers and Paradis (1980), Paradis (1985), and Paradis (1984).

"convergence" as defined by Laviosa-Braithwaite (1996). In Sections 5.2.2 and 5.4, we reported that a large number of translated English texts had variables with values that were off the scale of the Fry Readability Graphs. The Fry Graph scores (for Grade and Age) appear to apply exclusively to non-translated texts, and do not seem adapted to translated texts.

This unexpected feature of the Fry Graphs suggests a possible appropriation of their use. In future study, these graphs could perhaps be implemented as a means of further establishing the distinctive characteristics of translation. They might prove useful for research aimed at assisting national security agencies in their attempts to develop methods of distinguishing translated texts from non-translated texts, in situations where the status of a text is not otherwise known. Such research would constitute a type of "forensic" linguistics that would focus on markers distinguishing groups of texts rather than on traces of individual authorship.

In sum, future study of levelling-out might include a combination of readability indices, type/token ratios, and mean sentence lengths as a useful set of measures for discerning differences between translated and non-translated corpora.

6.6 Concluding remarks

The combination of corpus-based methodology with the theory of "translation universals" offers exciting opportunities for the discipline of Translation Studies. Separating translated texts from source texts (and thereby elevating the status of translated texts to objects of study in their own right) creates an advantageous new perspective. Scholars and practitioners alike stand to benefit

from the accurate description of translation that this approach affords. Assumptions about *what* translation actually is, which previously had to remain in the domain of the theoretical, can now be tested. Evidence of *why* translation works—a fresh line of inquiry in itself—can be added to the translation teacher's traditional roster of instructions on *how* to translate. Since comparison with non-translated texts is a key component of research on this topic, studies of this type may also be considered useful outside the field of Translation Studies, in those areas of linguistics, modern languages, and information studies that delve into the various forms and uses of human communication.

Bibliography

For easier access, this list of works consulted is divided into three subject sections:
1. Corpora and Research Methods
2. Grammar, Word-formation, and Readability
3. Translation Studies
Works covering two section categories are listed in both.

Authors whose names contain letters that are not part of the Modern Latin (Roman) alphabet are cited in the alphabetical order usual for their language. For instance, Øverås is cited at the end of the Translation Studies section, after the authors whose names begin with the letter "Z," and Pápai is listed in the same section after the authors whose names begin with the letters "Pa." German dieresis (e.g. Lörscher) is treated as a diphthong.

Wherever possible, Web sites have been archived at FURL <http://www.furl.net/members/sissela>. The only exceptions are those Web sites that were removed from the Internet before Web archiving became available, and which were not cached by Google. These are treated below as recommended in the *MLA Style Manual.*

Corpora and Research Methods

Aijmer, Karin and Bengt Altenberg, eds. (2004). *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)Göteborg 22-26 May 2002.* Amsterdam: Rodopi.

---------- (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman.

Aland, K., ed. (1975). *Vollstandige Konkordanz zum griechischen Neuen Testament.* Berlin: De Gruyter.

Aston, Guy, ed. (2001). *Learning with Corpora.* Houston: Athelstan.

Aston, Guy and Lou Burnard (1998). *The BNC Handbook: Exploring the British National Corpus with Sara.* Edinburgh: Edinburgh University Press.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler (1992). "Corpus Design Criteria." *Literary and Linguistic Computing* 7:1, pp. 1-16.

Babbie, Earl R. (2004). *The Practice of Social Research*. Belmont (CA): Wadsworth Publishing Company. Tenth edition.

Baker, Mona (2003). "The Translational English Corpus (TEC)." Accessed August 2004: <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>.

Biber, Douglas (1993). "Representativeness in Corpus Design." *Literary and Linguistic Computing 8*:4, pp. 243-257.

---------- (1990). "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation." *Literary and Linguistic Computing 5*:4, pp. 257-269.

---------- (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

---------- (1986). "Spoken and Written Textual Dimensions in English." *Language 62*: 2, pp. 384-414.

Biber, Douglas, S. Conrad, and R. Reppen (1994). "Corpus-based Approaches to Issues in Applied Linguistics." *Applied Linguistics 15*: 2, pp. 169-189.

Bilingual Canadian Dictionary Project. *Corpus TEXTUM*. Ottawa: University of Ottawa, 1988-1996.

British National Corpus (2002). "What is the BNC?" Accessed August 2004: <http://www.natcorp.ox.ac.uk/what/index.html>.

Botley, Simon Philip, Anthony Mark McEnery and Andrew Wilson, eds. (2000). *Multilingual Corpora in Teaching and Research*. Amsterdam: Rodopi.

Bowker, Lynne (2003). "Teaching Translation Technology: Towards an Integrated Approach." *Tradução & Comunicação 12*, pp. 65-79.

---------- (2002). *Computer-aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Bowker, Lynne and Peter Bennison (2003). "Student Translation Archive and Student Translation Tracking System: Design, Development and Application." In Zanettin, *et al.*, eds., pp. 104-117.

Bowker, Lynne and Jennifer Pearson (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Casebeer, Ann L. and Marja J. Verhoef (1997). *Chronic Diseases in Canada 18*:3, pp. 130-135. Published under the auspices of Health Canada, Population and Public Health Branch (PPHB).

Chaski, Carole E. (2001). "Empirical Evaluations of Language-based Author Identification Techniques." *The International Journal of Speech, Language and the Law: Forensic Linguistics 8*:1, pp. 1-65.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Coste, Didier (1988). "Pour une histoire littéraire négative." In Pym, ed., pp. 30-41.

Coulthard, Malcolm, ed. (1986). *Talking about Text.* University of Birmingham: English Language Research. Discourse Analysis Monograph No. 13.

Creswell, John W. (2003). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks (CA): Sage Publications. Second edition.

Cutting, Joan (2000). *Analysing the Language of Discourse Communities*. Amsterdam: Elsevier.

Daoust, François (2003). "SATO-4: Système d'Analyse de Texte par Ordinateur." Accessed August 14, 2003: <http://www.ling.uqam.ca/sato/outils/sato.htm> . Text is an introduction to SATO-4 by the software's developer.

Denzin, Norman K. and Yvonna S. Lincoln, eds. (2000). *Handbook of Qualitative Research*. Thousand Oaks (CA): Sage Publications. Second edition.

Dodd, Bill, ed. (2000). *Working with German Corpora*. Birmingham: University of Birmingham Press.

Government of Canada (2002). "Government On-line: Serving Canadians Better; Gouvernement en directe : Mieux servir les Canadiennes et les Canadiens." <http://www.gol-ged.gc.ca/index_e.asp> and <http://www.gol-ged.gc.ca/index_f.asp>.

Granger, Sylviane and Stephanie Petch-Tyson (2003). *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam: Rodopi.

Grant, Tim and Kevin Baker (2001). "Identifying Reliable, Valid Markers of Authorship: A Response to Chaski." *The International Journal of Speech, Language and the Law: Forensic Linguistics* 8:1, pp. 66-79.

Hansen, Silvia and Elke Teich (2001). "Multilayer Analysis of Translation Corpora: Methodological Issues and Practical Implications." *Proceedings: Eurolan (Romania) 2001, Summer Institute on "Creation and Exploitation of Annotated Language Resources"; Workshop: Multi-layer Corpus-based Analysis*. Accessed June 2003 <http://www.coli.uni-sb.de/~hansen/hansenteich.pdf> and <http://www.racai.ro/EUROLAN-2001/page/resources/workshops/corpus/>.

Huntsberger, David V. and Patrick. P. Billingsley (1987). *Elements of Statistical Inference.* Boston: Allyn and Bacon.

Johansson, Stig (1997). "Using the English-Norwegian Parallel Corpus: A Corpus for Contrastive Analysis and Translation Studies." In Lewandowska-Tomaszczyk and Melia, eds., pp. 282-296.

Johansson, Stig and Knut Hofland (2000). "The English-Norwegian Parallel Corpus: Current Work and New Directions." In Botley, McEnery, and Wilson, eds., pp. 134-147.

Johansson, Stig and Signe Oksefjell, eds. (1998). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.

Lamothe, Gilles (2005). "Comparing Two Proportions, Comparing Means, and Comparing Variances." *Notes for an Introductory Statistics Course: A Study Guide*. Accessed 2005: < http://aix1.uottawa.ca/~glamo058/mat2375.html>.

Lee, David (2001). "Software, Tools, Frequency Lists, etc." *Bookmarks for Corpus-based Linguists*. Accessed August 2002 and May 2004: <http://devoted.to/corpora>.

Leech, Geoffrey N. (1992). "Corpora and Theories of Linguistic Performance." In Svartvik, ed., pp. 105-122.

---------- (1991). "The State of the Art in Corpus Linguistics." In Aijmer and Altenberg, eds., pp. 8-29.

---------- (1966). *English in Advertising: A Linguistic Study of Advertising in Great Britain*. London, Longman.

Lewandowska-Tomaszczyk, Barbara and Patrick James Melia, eds. (1997). *Practical Applications in Language Corpora. PALC '97 Proceedings*. Lódz: Lódz University Press.

Ljung, Magnus, ed. (1997). *Corpus Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*. Amsterdam: Rodopi.

Mahoney, Anne (2002). "Review of Computer-aided Translation Technology: A Practical Introduction, L. Bowker." *Bryn Mawr Classical Review July 2002*. Accessed September 2002 <http://ccat.sas.upenn.edu/bmcr/2002/2002-07-28.html>.

Malmkjaer, Kirsten (1998). "Love Thy Neighbour: Will Parallel Corpora Endear Linguists to Translators?" *Meta 43*:4. Accessed June 2002 <http://www.erudit.org/revue/meta/1998/v43/n4/003545ar.html>.

Marcinkevičienė, Ruta (1997). "Hapax Legomena as a Platform for Text Alignment." In *Straipsnis atspausdintas: Proceedings of the Third European Seminar "Translation Equivalence" Montecatini Terme, Italy, October 16-18, 1997*. Kaunas, Lithuania: Centre of Computational Linguistics, Vytautas Magnus University, pp. 125 - 137. Accessed September 2003 <http://donelaitis.vdu.lt/publikacijos/hapax.htm>.

McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

---------- (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McMenamin, Gerald R. (2002)."Style Markers in Authorship Studies." *The International Journal of Speech, Language and the Law: Forensic Linguistics 8*:2, 93-97.

Nilsson, Kristina and Lars Borin (2002). "Living Off the Land: The Web as a Source of Practice Texts for Learners of Less Prevalent Languages." In the

Proceedings of LREC 2002. Third International Conference on Language Resources and Evaluation Las Palmas, Spain: ELRA II, pp. 411-418.

Norman, Geoffrey R. and David L. Streiner (2003). *PDQ Statistics*. Hamilton, Ontario: BC Decker. Third edition. *PDQ* Medical Series.

Partington, Alan (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Penas Ibáñez, Beatriz, ed. (1996). *The Intertextual Dimension of Discourse*. Zaragoza: Universidad de Zaragoza.

Percy, Carol E., Charles F. Meyer, and Ian Lancashire, eds. (1997). *Synchronic Corpus Linguistics: Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. Amsterdam: Rodopi.

Popper, Karl Raimund (1972; 1971). "My Solution of the Problem of Induction." In Popper (1972b), pp. 1-32. First published in *Revue Internationale de Philosophie 25*: 95-96, pp. 167-197.

---------- (1972b). *Objective Knowledge - An Evolutionary Approach*. Oxford: Clarendon Press.

Pym, Anthony (1998). *Method in Translation History*. Manchester: St. Jerome Press.

---------- (1996). "Multilingual Intertextuality in Translation." In Penas Ibáñez, ed., pp. 207-218.

Pym, Anthony, ed. (1988). *L'internationalité littéraire*. Paris: Noesis.

Rockwell, Geoffrey and Lian Yan (2004). *TAPoRware 1.0*. Accessed October 2004: <http://taporware.mcmaster.ca/~taporware/about.shtml>.

Russell, Bertrand (1959; 1912). *The Problems of Philosophy*. London: Oxford University Press. First published in the Home University Library, 1912. Public domain. HTML version accessed April 2004: <http://www.fh-augsburg.de/~harsch/anglica/Chronology/20thC/Russell/rus_pro0.html>.

Russell, Pamela (1993). "Evaluation : A Holistic Perspective." *Technostyle 11*: 2, pp. 86-97.

Salkie, Raphael (2000). "Unlocking the Power of the SMEMUC." In Botley, McEnery, and Wilson, eds., pp. 148-156.

---------- (1997). "Naturalness and Contrastive Linguistics." In Thelen and Lewandowska-Tomaszczyk, eds., pp. 297-312.

Sampson, Geoffrey (1996). "From Central Embedding to Corpus Linguistics." In Thomas and Short, eds., pp. 14-26.

Sanders, Donald H. (1990). *Statistics: A Fresh Approach*. New York: McGraw Hill. Fourth edition.

Sanders, Donald H., Robert J. Eng, and A. Franklin Murph (1985). *Statistics: A Fresh Approach*. New York: McGraw-Hill. Third edition.

Schmied, Josef (1993). "Qualitative and Quantitative Research Approaches to English Relative Constructions." In Souter and Atwell, eds., pp. 85-96.

Souter, Clive and Eric Atwell, eds. (1993). *Corpus-based Computational Linguistics*. Amsterdam: Rodopi.

Scott, Mike (2004). "WordSmith Tools 4: Help." On-line Publication of *WordSmith Version 4 Manual*. Accessed May 2003: <http://www.lexically.net/downloads/version4/html/index.html>.

---------- (1999). *WordSmith Tools* Version 3.00.00.

---------- (1999b). "Type/Token Ratios and the Standardised Type/Token Ratio." Hyperlinked in *WordSmith Tools Help*.

Sinclair, John M. (1996). "Beginning the Study of Lexis." In C.E. Bazell, *et al.*, eds., pp. 410-430.

---------- (1991). "*Corpus, Concordance, Collocation*." Oxford: Oxford University Press.

---------- (1987). "Grammar in the Dictionary." In Sinclair, ed., pp. 104-115.

Sinclair, John M., ed. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Harper Collins.

Stubbs, Michael (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

---------- (2001b). "Texts, Corpora, and Problems of Interpretation: A Response to Widdowson." *Applied Linguistics 22*:2, pp. 149-172.

---------- (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Basil Blackwell.

---------- (1995). "Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Methods." *Functions of Language 2*:1, pp. 1-33. Accessed April 2004: <http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/cause.htm>.

---------- (1994). "Grammar, Text, and Ideology: Computer-assisted Methods in the Linguistics of Representation. *Applied Linguistics 15*: 2, pp. 201-223.

---------- (1986). "Lexical Density: A Technique and Some Findings." In Coulthard, ed., pp. 27-42.

Svartvik, Jan, ed. (1992). *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter.

TAPoRware. See Rockwell and Yan.

Teubert, Wolfgang (2002). "The Role of Parallel Corpora in Translation and Multilingual Lexicography." In Altenberg and Granger, eds., pp. 189-214.

---------- (1999). "Corpus Linguistics: A Partisan View." *TELRI Newsletter 8:*99, pp. 4-19. Cited in Stubbs (2001: 221). Accessed April 2004: <http://tractor.bham.ac.uk/ijcl/teubert_cl.html>.

---------- (1999b) "Korpuslinguistik und Lexikographie." *Deutsche sprache 27*:4, pp. 292-313.

Thiele, Johannes (1987). *La formation des mots en français moderne*. Montréal: Presses de l'Université de Montréal. Traduction et adaptation par André Clas.

Thomas, Jenny A. and Mick H. Short, eds. (1996). *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*. London: Longman.

Tognini Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

---------- (2000). "Corpus Classroom Currency." On-line publication of the Kompiuterinės Lingvistikos Centro (Centre for Computational Linguistics), Vytautas Magnus University, Kaunas, Lithuania. Accessed April 23, 2004: <http://donelaitis.vdu.lt/indexa.html>.

---------- (1996). *Corpus Theory and Practice.* Vellano, Italy: Tuscan Word Centre.

Wichmann, Anne, Steve Fligelstone, Tony McEnery, and Gerry Knowles, eds. (1997). *Teaching and Language Corpora.* Longman, London.

Wilson, Andrew, Paul Rayson, and Tony McEnery, eds. (2003). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World.* München: Lincom-Europa.

Wilson, David R. (2001). "Ways in Which the Language of School Teaching Materials Can Impede Learning." Accessed October 2003: <http://www.tomwilson.com/david/sen/readability.doc>.

Woods, Anthony, Paul Fletcher and Arthur Hughes (1986). *Statistics in Language Studies.* Cambridge: Cambridge University Press.

WordSmith Tools. See Scott, Mike.

Zanettin, Frederico (2001). "Swimming in Words: Corpora, Translation and Language Learning." In Aston, ed., pp. 177-197.

---------- (2000). "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis." In Olohan, ed., pp. 105-118.

Zanettin, F., S. Bernardini, and D. Stewart, eds. (2003). *Corpora in Translator Education.* Manchester: St. Jerome.

Grammar, Word-formation, and Readability

Aarts, Bas and Charles F. Meyer, eds. (1995). *The Verb in Contemporary English: Theory and Description.* Cambridge: Cambridge University Press.

Adams, Barbara (1990). *Hapax Legomena.* Lewiston, New York: The Edwin Mellen Press.

Adams, Valerie (1973). *An Introduction to Modern English Word-formation.* London: Longman.

Agnihotri, R. K. and A. L. Khanna (1992). "Evaluating the Readability of School Textbooks: An Indian Study." *Journal of Reading 35*: 4 (December 1991/January 1992), pp. 282-288.

Aijmer, Karin (1998). "Epistemic Predicates in Contrast." In Johansson and Oksefjell, eds., pp. 277-296.

Aijmer, Karin and Bengt Altenberg, eds. (2004). *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)Göteborg 22-26 May 2002.* Amsterdam: Rodopi.

---------- (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman.

Altenberg, Bengt (2002). "Causative Constructions in English and Swedish: A Corpus-based Contrastive Study." In Altenberg and Granger, eds., pp. 97-116.

---------- (1998). "Connectors and Sentence Openings in English and Swedish." In Johansson and Oksefjell, eds., pp. 115-144.

Altenberg, Bengt, and Sylviane Granger, eds. (2002). *Lexis in Contrast: Corpus-based Approaches*. Amsterdam: John Benjamins.

Austerlitz, Robert Paul, ed. (1975). *The Scope of American Linguistics*. Lisse: Peter de Ridder Press.

Bartsch, Renate (1987). *Norms of Language: Theoretical and Practical Aspects*. London: Longman.

Bauer, Laurie (1983). *English Word-formation*. Cambridge: Cambridge University Press.

Bazell, C. E., J. C. Catford, M.A. K. Halliday, and R. H. Robins, eds. (1996). *In Memory of J. R. Firth*. London: Longmans, Green and Co. Ltd.

Beaman, Karen (1984). "Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse." In Tannen, ed., pp. 45-80.

Bédard, Édith et Jacques Maurais (1983). *La norme linguistique*. Paris: Robert.

Bell, James and Reta E. Johnson (1992). "Effect of Lowering the Reading Level of a Health Education Pamphlet on Increasing Comprehension by ESL Adults." *TESL Canada Journal / Revue TESL du Canada*. 10:1, pp. 9-26.

Bernstein,Theodore, Marylea Meyersohn, and Bertram Lippman (1999). *Dos, Don'ts and Maybes of English Usage*. New York: Gramercy Books.

Berthold, M., F. Mangubhai, and K. Batorowicz (1997). *Bilingualism and Multiculturalism: Study Book*. Distance Education Centre, University of Southern Queensland: Toowoomba, QLD.

Biber, Douglas (1993). "Representativeness in Corpus Design." *Literary and Linguistic Computing 8:*4, pp. 243-257.

---------- (1990). "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation." *Literary and Linguistic Computing 5*:4, pp. 257-269.

---------- (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

---------- (1986). "Spoken and Written Textual Dimensions in English." *Language 62*: 2, pp. 384-414.

Biber, Douglas, S. Conrad, and R. Reppen (1994). "Corpus-based Approaches to Issues in Applied Linguistics." *Applied Linguistics 15*: 2, pp. 169-189.

Birkerts, Sven (1994). *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. Boston: Faber and Faber.

Björnsson, C.H. (1968). *Läsbarhet*. [*Readability*.] Lund: Liber.

---------- (1983). "Readability of Newspapers in 11 Languages." *Reading Research Quarterly 18*: 480-487.

Björnsson, C.H. and Birgit Hård af Segerstad (1979). *Lix på franska och tio andra språk*. [*Readability Index for French and Ten Other Languages*.] Stockholm: Pedagogiskt centrum, Stockholms skolförvaltning [Stockholm School Board].

Blanchard, Jay S., George E. Mason and Dan Daniel (1987; 1983). *Computer Applications in Reading*. Newark, Delaware: International Reading Association.

Bossé-Andrieu, Jacqueline (1994). "Le poids de trois siècles de normativisme linguistique." *Technostyle 11*:3/4, pp. 1-15.

---------- (1993). "La question de la lisibilité dans les pays anglophones et les pays francophones." *Technostyle 11*:2, pp. 73-85.

Botel, Morton and Alvin Granowsky (1974). "A Formula for Measuring Syntactic Complexity: A Directional Effort." *Elementary English 1*: 513-516.

Brinton, Laurel J. (2000). *The Structure of Modern English: A Linguistic Introduction*. Amsterdam: John Benjamins.

Cambridge International Dictionary of English (1995). Cambridge: Cambridge University Press.

Canadian Oxford Dictionary (1998). Don Mills, Ontario: Oxford University Press.

Canadian Style: A Guide to Writing and Editing (1997). Toronto: Dundurn Press Limited.

Carrell, Patricia L. (1987). "Readability in ESL" *Reading in a Foreign Language 4*:1, pp. 21-40.

Cedergren, Magnus (1992). "Kvantitativa läsbarhetsanalyser som metod för datorstödd granskning." ["Quantitative Readability Analysis as a Method of Automated Assessment."] *Technical Report IPLab-55*. Stockholm: Kungliga Tekniska Högskolan. Accessed July 2004: <http://www.lysator.liu.se/~mace/skriv/laesbarhet.html> .

Cherry, Lorinda (1982). "Writing Tools." *IEEE Transactions on Communications*. Vol COM-30: No. 1, pp. 100-105.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Collins COBUILD (2000; 1990). *English Grammar*. London: Harper Collins.

---------- (1992). *English Usage*. London: Harper Collins.

Comrie, Bernard. (1989; 1981). *Language Universals and Linguistic Typology*. Oxford: Basil Blackwell Ltd. Second edition.

Cooper, Charles R. and Lee Odell, eds. (1977). *Evaluating Writing: Describing, Measuring, Judging*. Urbana, Illinois: National Council of Teachers of English.

Coulthard, Malcolm, ed. (1986). *Talking about Text*. University of Birmingham: English Language Research. Discourse Analysis Monograph no. 13.

Crystal, David. (1997; 1980). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell. Fourth edition.

---------- (1987). *The Cambridge Encyclopedia of Language*. Cambridge University Press: Cambridge.

Daoust, François (2003). "SATO-4: Système d'Analyse de Texte par Ordinateur." Accessed August 14, 2003: <http://www.ling.uqam.ca/sato/outils/sato.htm> . Text is an introduction to SATO-4 by the software's developer.

Décsy, Gyula (1987). *A Select Catalog of Language Universals*. Bloomington, Indiana: Eurolingua.

Depecker, Loïc (2001). *L'invention de la langue: le choix des mots nouveaux*. Paris: Armand Colin-Larousse.

Dictionnaire de difficultés grammaticales et lexicologiques (1994; 1949). Bruxelles: Éditions Baude. Hanse, Joseph, rédacteur.

Dictionnaire du français Plus à l'usage des francophones d'amérique (1988). Montréal: Centre éducatif et culturel. Poirier, C., L. Mercier, et C. Verrault, rédacteurs.

Dictionnaire universel francophone en ligne. Paris: Hachette. Accessed June and July 2003: *<http://www.francophonie.hachette-livre.fr/>*.

Diderot, Denis, Pierre Mouchon, et Jean Lerand d'Alembert, éditeurs (1751-1780). *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers. Tome II*. Genève: Pellet. Nouvelle impression en facsimilé publié à Stuttgart-Bad Cannstatt: Frommann, 1966-1967.

Di Sciullo, A.M., P. Muysken, and R. Singh (1986). "Government and Code-Mixing." *Journal of Linguistics 22*, pp. 1-24.

Dixon, R. M. W. (1991). *A New Approach to English Grammar, On Semantic Principles*. Oxford: Clarendon Press.

Doherty, Monika (1998). "Clauses or Phrases: A Principled Account of *when*-Clauses in Translations Between English and German." In Johansson and Oksefjell, eds., 235-254.

---------- (1993). "Parametrisierte Perspektive." *Zeitschrift für Sprachwissenschaft 12*, pp. 3-38.

Dougherty, Ray C., Franca Ferarri-Bridgers, and Lisbeth Dyer (2001). "INTEX Solves Pronunciation and Intonation Problems in Text to Speech Reading Machines." Conference paper: *ACH / ALLC 2001 - New York University, June 13th - June 16th 2001.* Accessed September 29, 2003 : <http://www.nyu.edu/its/humanities/ach_allc2001/papers/dougherty/>.

Dubuc, Robert (1996). *Une grammarie pour écrire: Essai de grammaire stylistique.* Brossard, Québec: Linguatech.

Duran, Luisa (1994). "Toward a Better Understanding of Code Switching and Interlanguage in Bilinguality: Implications for Bilingual Instruction." *The Journal of Educational Issues of Language Minority Students 14*, pp. 69-88. Accessed June 2003: <http://www.ncela.gwu.edu/miscpubs/jeilms/vol14/duran.htm>.

Editors' Association of Canada (2000). *Editing Canadian English.* Toronto: Macfarlane Walter and Ross. Second Edition.

Ellis, Rod (1994). *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

---------- (1985). *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Elsness, Johan (1984). "*That* or Zero? A Look at the Choice of Object Clause Connective in a Corpus of American English." *English Studies 65*: 6, pp. 519-533.

Endicott, A. L. (1973). "A Proposed Scale for Syntactic Density." *Research in the Teaching of English 7*: 512.

Finegan, Edward J., and Douglas Biber (1995). "*That* and Zero Complementisers in Late Modern English: Exploring ARCHER from 1650-1990." In Aarts and Meyer, eds., pp. 241-257.

Firth, J. R. (1968). *Selected Papers of J. R. Firth 1952-1959.* Harlow: Longmans, Green and Co. Edited by F. R. Palmer.

Flesch, Rudolph (1948). "A New Readability Yardstick." *Journal of Applied Psychology 32*, pp. 221-233.

---------- (1948b). "A Readability Formula in Practice." *Elementary English 25*, pp. 344-351.

---------- (1949). *The Art of Readable Writing*. New York: Harper and Brothers.

Fodor, Janet Dean and Fernanda Ferreira (1998). *Reanalysis in Sentence Processing*. Boston: Kluwer Academic Publishers.

Fodor, Janet Dean and A. Inoue (1998). *Attach Anyway: Reanalysis in Sentence Processing*. In Fodor and Ferreira, eds., pp. 101-141.

Fowler, Henry Watson (1965; 1926). *Dictionary of Modern English Usage*. Second edition revised by Sir Ernest Gowers.

Frontier, Alain (1997). *La grammaire du français*. Paris: Belin.

Fry, Edward (1989). "Reading Formulas—Maligned but Valid." *Journal of Reading 32*: 4, pp. 293-297.

---------- (1988). "Writeability: The Principles of Writing for Increased Comprehension." In Zakaluk and Samuels, eds., pp. 77-95.

---------- (1977). *Elementary Reading Instruction*. NY: McGraw Hill.

---------- (1977b). "A Readability Formula that Saves Time." *Classroom Strategies for Secondary Reading*. In Harker, ed., pp. 29-35.

Gage Canadian Dictionary (1983). Toronto: Gage Educational Publishing Company.

Gaies, Stephen J. (1980). "T-unit Analysis in Second Language Research: Applications, Problems and Limitations." *TESOL Quarterly 14*: 1, 54-60.

Gaines, Philip (2003). "Negotiating Power at the Bench: Informal Talk in Sidebar Sessions." *The International Journal of Speech, Language and the Law: Forensic Linguistics 9*:2, 213-234.

Gass, Susan M. and Carolyn G. Madden, eds. (1985). *Input in Second Language Acquisition*. Rowley, Mass.: Newbury House Publishers.

Gentner, Dedre and Susan Goldin-Meadow, eds. (2003). *Language in Mind: Advances in the Investigation of Language and Thought.* Cambridge, MA: MIT Press.

Gibson, Edward (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown.* Ph.D Thesis, Carnegie Mellon University, Pittsburgh, P.A.

---------- (1998). "Linguistic Complexity: Locality of Syntactic Dependencies." *Cognition 68*, pp. 1-76.

Golub, Lester Stanley (1969). "Linguistic Structures in Students' Oral and Written Discourse." *Research in the Teaching of English 3*: 70-85.

Golub, Lester Stanley and Carole Kidder (1974). "Syntactic Density and the Computer." *Elementary English 51*: 1128-1131.

Gougenheim, Georges (1969). *Système grammatical de la langue française.* Paris: Éditions d'Artrey.

Gougenheim, G., P. Rivenc, R. Michea, and A. Sauvageot (1967). *L'élaboration du français fondamental.* Paris: Didier.

Gouvernement du Québec (2002). *Terminogramme 101-102 : Interventions sociolinguistiques et pratiques langagières.* Numéro préparé sous la direction de Monique C. Cormier et Noëlle Guilloton.

Gouvernement du Québec (2002b). *Nouveau grand dictionnaire terminologique.* <http://www.olf.gouv.qc.ca/ressources/gdt_bdl2.html> .

Gouvernement du Québec (2002c). *BDL: La banque de dépannage linguistique.* <http://www.olf.gouv.qc.ca/ressources/gdt_bdl2.html> .

Greenberg, Joseph. H. (1966). "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements." In Greenberg, ed., pp. 73-113.

Greenberg, J. H., ed. (1966). *Universals of Language.* Cambridge, Mass.: MIT Press. Second edition.

Grevisse, Maurice (1990). *Précis de grammaire française.* Paris: Duculot. Twenty-ninth edition.

Gunning, Robert (1968; 1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

---------- (1964). *More Effective Writing in Business and Industry*. Boston: Industrial Education Institute.

Hacker, Diana (2001). *A Canadian Writer's Reference*. Scarborough, Ontario: Nelson Thomson Learning. Adapted from *A Writer's Reference*, Third edition, 1995, St. Martin's Press.

Halliday, M.A.K. (1998). "Corpus Studies and Probabilistic Grammar." In Aijmer and Altenberg, eds., pp. 30-43.

---------- (1994; 1985). *An Introduction to Functional Grammar*. London: Edward Arnold. Second edition.

---------- (1992). "Language as System and Language as Instance: The Corpus as a Theoretical Construct." In Svartvik, ed., pp. 61-77

---------- (1991). "Towards Probabilistic Interpretations." In Ventola, ed., pp. 39-61.

---------- (1991b). "Corpus Studies and Probabilistic Grammar." In Aijmer and Altenberg, eds., pp. 30-43.

---------- (1989). *Spoken and Written Language*. Oxford: Oxford University Press.

---------- (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Halliday, M.A.K. and Ruqaia Hasan (1976). *Cohesion in English*. Harlow, England: Pearson Education Limited. *English Language Series*. Gen. ed.: Randolph Quirk.

Hanse, Joseph, éd. (1994; 1949). *Dictionnaire de difficultés grammaticales et lexicologiques*. Bruxelles: Éditions Baude.

Hapax Legomena v. 1.01 (2003). Eds. Nicolaas Dion & Hansje van Halem. Amsterdam: Rietveld Academie. Accessed October 2003: <http://www.nicolaas.net/hapax/index.php?l=en>.

Harker, W. John, ed. (1977). *Classroom Strategies for Secondary Reading*. Newark, Delaware: International Reading Association.

Harpaz, Yehouda (1998). "Psycholinguistics Blatant Nonsense Examples." Accessed May 2004: < http://human-brain.org/nonsense.html> .

Harrison, C. (1980). *Readability in the Classroom*. Cambridge: Cambridge University Press.

---------- (1986). "Readability in the United Kingdom." *Journal of Reading 29*: 6, pp. 521-529.

Hartmann, Reinhard R.K. (1982). "Contrastive Textology: Comparative Discourse Analysis in Applied Linguistics." *Germanistik 23*, pp. 245-246.

Hatim, Basil and Ian Mason (1990). *Discourse and the Translator*. London: Longman.

Hawkins, J.A. (1986). *A Comparative Typology of English and German*. London: Croom Helm.

Heaton, J. B. (1975). *Writing English Language Tests*. London: Longman.

Heift, Trude (2003). "Morphology: The Analysis of Word Structure." Lecture notes, Linguistics course, Simon Fraser University. PDF file accessed June 2004: <http://www.sfu.ca/~heift/Ling220/lectures.htm> and <http://www.sfu.ca/~heift/Ling220/lecturenotes/Morphology.pdf>.

Henry, Georges (1987; 1975). *Comment mesurer la lisibilité*. Bruxelles: Éditions Labor.

Herriman, Jennifer (2000). "Extraposition in English: A Study of the Interaction Between the Matrix Predicate and the Type of Extraposed Clause." *English Studies 6*, pp. 582-599.

Hockett, Charles F. (1966). "The Problem of Universals in Language." In Greenberg, ed., pp. 1-29.

Hixon, M. W. (1999; 1995). *The Essentials of English Language*. New Jersey: Research and Education Association.

Hunston, Susan and Gill Francis (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Hunt, Kellogg W. (1977). "Early Blooming and Late Blooming Syntactic Structures." In Cooper, *et al.*, eds., pp. 91-104.

---------- (1966). "Recent Measures in Syntactic Development." In *Elementary English*. Champlain, Illinois: National Council of Teachers of English, pp.732-739.

Kane, Thomas S. (1983). *The Oxford Guide to Writing: A Rhetoric and Handbook for College Students*. New York: Oxford University Press.

Kennedy, Graeme (1992). "Preferred Ways of Putting Things with Implications for Language Teaching." In Svartvik, ed., pp. 335-373.

Kidder, Carole L. (1974). "Using the Computer to Measure Syntactic Density and Vocabulary Intensity in the Writing of Elementary School Children." *Dissertation Abstracts International 35*: 3524-A.

Kincaid, J., R. Fishburne, R. Rodgers, and B. Chiasson (1975). "Derivation of New Readability Formulas for Navy Enlisted Personnel." *Branch Report 8—75*. Millington, Tennessee: Chief of Naval Training.

Klare, George Roger (1988). "The Formative Years." In Zakaluk and Samuels, eds., pp. 14-34.

Klare, George Roger, ed. (1969; 1963). *The Measurement of Readability*. Ames: Iowa State University Press. Introduction by George R. Klare.

Klare, George Roger and Byron Buck (1954). *Know Your Reader: The Scientific Approach to Readability*. New York: Hermitage House.

Künig, Ekkehard and Herbert Wiegand Ernst, Hugo Steger, Martin Haspelmath, eds. (2001). *Language Typology and Language Universals - An International Handbook: Volume 1*. Paris: Mouton de Gruyter.

Labov, William M. (1975). "Empirical Foundations of Linguistic Theory." In Austerlitz, ed., pp. 77-133. Also published as *What is a Linguistic Fact?* Lisse: de Ridder, 1975.

Ladmiral, Jean-René (1991). "La langue violée?" *Palimpsestes 6*, pp. 23-33.

Lakoff, George and Mark Johnson (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

---------- (1989). "Image-schematic Bases of Meaning." *RSSI Recherches Sémiotiques/ Semiotic Inquiry 9*: 1-2-3, pp. 109-118.

---------- (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Leech, Geoffrey N. (1992). "Corpora and Theories of Linguistic Performance." In Svartvik, ed., pp. 105-122.

---------- (1991). "The State of the Art in Corpus Linguistics." In Aijmer and Altenberg, eds., pp. 8-29.

---------- (1966). *English in Advertising: A Linguistic Study of Advertising in Great Britain*. London, Longman.

Leech, Geoffrey N. and Jan Svartvik (1975). *A Communicative Grammar of English*. Essex: Longman. Based on *A Grammar of Contemporary English* by Quirk *et al.*

Lehman, Hans Martin (1997). "Zero Relatives: Automatic Retrieval of Zero Elements in a Computerized Corpus." In Ljung, ed., pp. 179-194.

Leistyna, Pepi and Charles F. Meyer, eds. (2003). *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi.

Lemire, Gilles (2001). *Grammaire BEPP*. Accessed June 2003: <http://www.fse.ulaval.ca/fac/Grammaire-BEPP/ >.

Lété, Bernard, Liliane Sprenger-Charolles, and Pascale Colé (2004). "MANULEX: A Grade-level Lexical Database from French Elementary School Readers." *Behavior Research Methods, Instruments, & Computers 36*:1, pp. 156-166 (11). February 1, 2004.

Lively, B.A. and Pressey, S. L. (1923). "A Method for Measuring the 'Vocabulary Burden' of Textbooks." *Educational Administration and Supervision 9*, 389-398.

Lorge, Irving (1966; 1959). *The Lorge Formula for Estimating Difficulty of Reading Materials*. New York: Bureau of Publications, Teacher's College, Columbia University. Second edition.

---------- (1944). "Word Lists as Background for Communication." *Teachers College Record 45*, pp. 543-552.

Lynch, Aaron (2002). "Re: Some Computed Reading Levels of Book Portions." Online posting. Sat 19 Oct 2002 - 13:29:28 GMT. *Mimetics Discussion List*

Archive (Associated with JOM-EMIT). Accessed 2002: <http://cfpm.org/~majordom/memetics/2000/11859.html >.

Macdonald, Nina H., Lawrence T. Frase, Patricia S. Gingrich, and Stacey A. Keenan (1982). "The Writer's Workbench: Computer Aids for Text Analysis." *IEEE Transactions on Communications Vol. COM-30*: No. 1, pp. 105-110.

Maia, Belinda (1998). "Word Order and the First Person Singular in Portuguese and English." *META 43*: 4, pp. Accessed April 2004: <http://www.erudit.org/revue/meta/1998/v43/n4/index.html>.

Martin, Henri-Jean (1989). *Histoire et pouvoirs de l'écrit*. Paris: Librairie académique Perrin. Avec la collaboration de Bruno Delmas; Collection "Histoire et décadence." Réédition Paris: Albin-Michel, 1996, dans la collection Bibliothèque de l'évolution de l'humanité.

---------- (1994; 1989). *The History and Power of Writing*. Chicago: University of Chicago Press. Translated by Lydia G. Cochrane.

McDavid, Virginia (1964). "The Alternation of *That* and Zero in Noun Clauses." *American Speech 39*: 102-113.

McNamara, D.S., Louwerse, M.M. & Graesser, A.C. (2003). "Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension." Accessed October 2003 < http://csep.psyc.memphis.edu/cohmetrix/readabilityresearch.htm/>.

Melchers, Gunnel and Beatrice Warren, eds. (1995). *Studies in Anglistics*. Stockholm: Almqvist and Wiksell. Series: *Stockholm Studies in English 85*.

Milroy, James and Leslie Milroy (1998; 1985). *Authority in Language: Investigating Standard English*. London: Routledge. Third edition.

Multidictionnaire de la langue française (1997). Québec: Québec-Amérique. M-E de Villers, rédactrice en chef.

Noël, Dirk. (1997). "The Choice Between Infinitives and *That*-clauses After *Believe*." *English Language and Linguistics 1*: 2, pp. 271-284.

Nouveau dictionnaire des difficultés du français moderne (1994). Louvain-la-Neuve: De Boeck-Duculot. 3e édition établie d'après les notes de Joseph Hanse avec la collaboration scientifique de Daniel Blampain.

O'Donnell, Roy C. (1974). "Syntactic Differences Between Speech and Writing." *American Speech 49*, pp. 102-110.

Partington, Alan (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Paulussen, H. (1999). *A Corpus-based Contrastive Analysis of English on/up, Dutch op and French sur Within a Cognitive Framework.* Ph.D dissertation. Gent: Faculteit Letteren en Wijsbegeerte Vakgroep Engels Universiteit Gent. Summarized in Laviosa 2002, pp. 96-97.

Pearson, P. David, ed. (1984). *Handbook of Reading Research*. New York: Longman.

Penas Ibáñez, Beatriz, ed. (1996). *The Intertextual Dimension of Discourse*. Zaragoza: Universidad de Zaragoza.

Perren, G. E. and J. L. M. Trim, eds. (1971). *Applications of Linguistics*. London: Cambridge University Press.

Pigeon, Jocelyne (2002). "Editing Advice for Technical Communicators." *Superscript 11*: 1. Edmonton, Alberta: Society for Technical Communication, Alberta Chapter. Accessed July 2003 <http://www.stc-alberta.org/Newsletter/Documents/JulyAug2002.pdf>.

Poplack, Shana (1980). "Sometimes I'll start a sentence in English y termino en espanol." *Linguistics 18*, pp. 581-616.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman. Index by David Crystal.

---------- (1973). *A Grammar of Contemporary English*. London: Longman.

Quirk, Randolph and Sidney Greenbaum (1973). *A University Grammar of English*. London: Longman.

Rabin, Annette T. (1988). "Determining Difficulty Levels of Text Written in Languages Other than English." In Zakaluk and Samuels, eds., pp. 46-76.

Richaudeau, François (1981). *Linguistique pragmatique*. Paris : Éditions Retz.

Riegel, Martin, Jean-Christophe Pellat, et René Rioul (1998; 1994). *Grammaire méthodique du français*. Paris: Presses universitaires de France. 4e édition.

Rissanen, Matti (1991). "On the History of *That*/Zero as Object Clause Links in English." In Aijmer and Altenberg, eds., pp. 272-289.

Robert: Dictionnaire québécois d'aujourd'hui (1992). Saint-Laurent: Dicorobert. Jean-Claude Boulanger, rédacteur en chef.

Roberts, John C. M.D., Robert H. Fletcher, M.D., Suzanne W. Fletcher, M.D. (1994). "Effects of Peer Review and Editing on the Readability of Articles Published in Annals of Internal Medicine." *JAMA 272*, pp. 119-121.

Russell, Bertrand (1959; 1912). *The Problems of Philosophy*. London: Oxford University Press. First published in the Home University Library, 1912. Public domain. HTML version accessed April 2004: <http://www.fh-augsburg.de/~harsch/anglica/Chronology/20thC/Russell/rus_pro0.html>.

Russell, Pamela (1993). "Evaluation : A Holistic Perspective." *Technostyle 11*: 2, pp. 86-97.

Salkie, Raphael (1997). "Naturalness and Contrastive Linguistics." In Thelen and Lewandowska-Tomaszczyk, eds., pp. 297-312.

Salkoff, Morris (1999). *A French-English Grammar: A Contrastive Grammar on Translational Principles*. Amsterdam: John Benjamins.

Santos, Diana Maria de Sousa Marques Pinto dos (1998). "Perception Verbs in English and Portuguese." In Johansson and Oksefjell, eds., 319-332. Summarized in Laviosa (2002), p. 96. Excerpted on line. Accessed April 2004: <http://www.linguateca.pt/Diana/tese.html>.

---------- (1996). *Tense and Aspect in English and Portuguese: A Contrastive Semantical Study*. Ph.D dissertation. Lisbon: Instituto Superior Téchnico, Universidade Técnic de Lisboa. Summarized in Laviosa (2002), p. 96.

Saussure, Ferdinand de (1916: 1965). *Cours de linguistique générale*. Paris: Payot. *Course in General Linguistics*. Chicago: Open Court Classics. Published by

Charles Bally, Albert Sechehaye, and Tullio de Mauro. Originally translated by Wade Baskin (1959). Re-translated by Roy Harris.

Schiffrin, Deborah (1984). *Meaning, Form, and Use in Context: Linguistic Applications.* Washington, D.C.: Georgetown University Press.

Schrock, Kathy (1995-2003). "Fry Readability Graphs." *Guide for Educators.* Accessed June 2002: <http://school.discovery.com/schrockguide/fry/fry2.html>.

Shibatani, M. and S. A. Thompson, eds. (1996). *Grammatical Constructions: Their Form and Meaning.* Oxford: Oxford University Press.

Shopen, Timothy, ed. (1985). *Language Typology and Semantic Description, Volume 3: Grammatical Categories and the Lexicon.* Cambridge: Cambridge University Press.

Skiba, Richard (1997). "Code Switching as a Countenance of Language Interference." *The Internet TESL Journal III*: 10. Accessed June 2003: <http://iteslj.org/Articles/Skiba-CodeSwitching.html >.

Slobin, Dan (forthcoming). "Language and Thought Online: Cognitive Consequences of Linguistic Relativity." Accessed January 2003: <http://ihd.berkeley.edu/slobinpaper-online.pdf>. To appear in Gentner and Goldin-Meadow, eds., *Advances in the Investigation of Language and Thought.* Cambridge, M.A.: MIT Press.

----------(2002). "Verbalized Events: A Dynamic Approach to Linguistic Relativity and Determinism." In Niemeier *et al.*, eds., pp. 107-138.

---------- (1996). "Two Ways to Travel: Verbs of Motion in English and Spanish." In Shibatani *et al.*, eds., pp. 195-217.

---------- (1987). "Thinking for Speaking." *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 435-444.

Storms, G. (1966). "*That*-clauses in Modern English." *English Studies 47*: 249-270.

Surridge, Marie (1984). "L'utilité d'un vocabulaire structuré pour l'étudiant anglophone du français langue seconde." *Canadian Modern Language Review / La revue canadienne des langues vivantes 40*: 5, pp. 563-574.

Swan, Michael (1995; 1980). *Practical English Usage*. Oxford: Oxford University Press. Second edition.

Talburt, John (1986). "The Flesch Index: An Easily Programmable Readability Analysis Algorithm." In *Proceedings of the 4th Annual International Conference on Systems Documentation, ACM Special Interest Group for Design of Communications*. New York: ACM Press, pp. 114-122.

Talmy, Leonard (1991). "Path to Realization: A Typology of Event Conflation." *Proceedings of the Seventeenth Annual Meeting of the Berkeley Linguistics Society*, pp. 480-519.

---------- (1985). "Lexicalization Patterns: Semantic Structure in Lexical Forms." In Shopen, ed., pp. 36-149.

Tannen, Deborah (1982). "The Oral/Literate Continuum in Discourse." In Tannen, ed., pp. 1-16.

Tannen, Deborah, ed. (1984). *Coherence in Spoken and Written Discourse*. Norwood, New Jersey: Ablex.

---------- (1982). *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, New Jersey: Ablex.

Thompson, Sandra Annear (1985). "'Subordination' in Formal and Informal Discourse." In Schiffrin, ed., pp. 85-94.

Thorndike, Edward Lee and Irving Lorge (1944; 1933; 1921). *Teacher's Word Book of 30,000 Words*. New York: Teacher's College of Columbia University.

Tottie, Gunnel (1996). "Grammar and Corpus Linguistics." *ICAME Journal: Computers in English Linguistics 20*, pp. 107-111.

---------- (1995). "The Man Ø I Love: An Analysis of Factors Favouring Zero Relatives in Written British and American English." In Melchers and Warren, eds., pp. 201-215.

Tremblay, Monique (2000). "Les contrats d'assurance : Le long chemin vers la transparence." *Bulletin de l'Institut canadien des actuaires 1-3*.

Ure, Jean (1971). "Lexical Density and Register Differentiation." In Perren and Trim, eds., pp. 443-52.

Valdman, Albert (1979). *Le Français hors de France*. Paris: Champion.

Vandendooren, Joris (2000). *Écriture journalistique.com : Définition d'une "hyperécriture" journalistique à partir de différentes théories de lisibilité et de l'étude du cas du site "Coup d'œil du Net."* Bruxelles : Institut des hautes études des communications sociales de Bruxelles (IHECS), Haute École Galilée. Mémoire présenté pour l'obtention du titre de Licencié en Communication appliquée, Section Presse et Information. Accessed July 2003: <http://www.citizen-kid.net/ecriture/>.

Venezky, Richard L. (1984). "The History of Reading Research." In Pearson, P. David, ed., pp. 3-38.

Ventola, Eija, ed. (1991). *Functional and Systemic Linguistics: Approaches and Uses*. Berlin: Mouton de Gruyter.

Viberg, Åke (2002). "Polysemy and Disambiguation Cues Across Languages: The Case of Swedish *få* and English *get*." In Altenberg and Granger, eds., pp. 119-150.

---------- (1998). "Contrasts in Polysemy and Differentiation: Running and Putting in English and Swedish." In Johansson and Oksefjell, eds., pp. 343-376.

Villers, Marie-Éva de (1988). *Multidictionnaire des difficultés de la langue française*. Montréal: Éditions Québec-Amérique.

Wagner, R. L. et J. Pinchon (1977; 1962). *Grammaire du français classique et moderne*. Paris: Hachette.

Westin, Ingrid (2002). *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.

Wilmet, Marc (1997). *Grammaire critique du français*. Louvain-la-Neuve: Duculot.

Zakaluk, Beverly L. and S. Jay Samuels (1988). "Toward a New Approach to Predicting Text Comprehensibility." In Zakaluk and Samuels, eds., pp. 121-144.

Zakaluk, Beverly L. and S. Jay Samuels, eds. (1988). *Readability: Its Past, Present, and Future*. Newark, Delaware: International Reading Association.

Translation Studies

Álvarez, Román and M. Carmen-ÁfricaVidal, eds. (1996). *Translation, Power, Subversion*. Clevedon: Multilingual Matters Ltd. *Topics in Translation 8.*

Anderson, Benedict (1991; 1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London: Verso. Second edition.

Apter, Ronnie (1987; 1984). *Digging for the Treasure: Translation After Pound*. New York: Paragon House.

Aston, Guy, ed. (2001). *Learning with Corpora*. Huston: Athelstan.

Baker, Mona (2003). "The Translational English Corpus (TEC)." Accessed August 2004: <http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm>.

----------(2000). "Towards a Methodology for Investigating the Style of a Literary Translator." *Target* 12:2, pp. 241-266.

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

---------- (1998). "Réexplorer la langue de la traduction: Une approche par corpus." ["Investigating the Language of Translation: A Corpus-based Approach"] *Meta 43*:4, pp. 480-485.

---------- (1996). "Corpus-based Translation Studies: The Challenges That Lie Ahead." In Somers, ed., pp. 175-186.

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

---------- (1993). "Corpus Linguistics and Translation Studies—Implications and Applications." In Baker *et al.*, eds., pp. 233-250.

---------- (1992). *In Other Words*: *A Coursebook on Translation*. London: Routledge.

Baker, Mona, ed. (1998). *The Routledge Encyclopedia of Translation Studies*. London: Routledge.

Baker, Mona, Gill Francis and Elena Tognini-Bonelli, eds. (1993). *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins.

Bell, Roger (1991). *Translation and Translating*. London: Longman.

Benjamin, Walter (1968; 1923). "The Task of the Translator." *Illuminations*. Frankfurt am Main: Suhrkamp Verlag. English translation by Harry Zohn. Reprinted in Schulte & Biguenet, eds. (1992), pp. 71-82. Also reprinted in Venuti, ed. (2000), pp. 15-25.

Berman, Antoine (2000; 1985). "Translation and the Trials of the Foreign." In Venuti, ed., pp. 284-297. Translated by Lawrence Venuti.

---------- (1985). "L'analytique de la traduction et la systématique de la déformation." In Granel, ed., pp. 65-82.

---------- (1992; 1984). *L'épreuve de l'étranger*. Paris: Gallimard. *The Experience of the Foreign: Culture and Translation in Romantic Germany*. Albany, New York: State University of New York Press. Translated by S. Heyvaert.

Bernardini, Silvia and Federico Zanettin (2004). "When is a Universal Not a Universal? Some Limits of Current Corpus-based Methodologies for the Investigation of Translation Universals." In Mauranen and Kujamäki, eds., pp. 51-62.

Biguenet, John and Rainer Schulte, eds. (1989). *The Craft of Translation*. Chicago: University of Chicago Press.

Blum-Kulka, Shoshana (2000; 1986). "Shifts of Cohesion and Coherence in Translation." In House and Blum-Kulka, eds. (1986), 17-35. Reprinted in Venuti, Lawrence, ed. (2000), pp. 298-315.

Blum-Kulka, Shoshana and Eddie A. Levenston (1983). "Universals of Lexical Simplification." In Faerch, *et al.*, eds., pp. 119-139.

Boisvert, Lionel, Claude Poirier, and Claude Verreault (1986). *La Lexicographie québécoise: bilan et perspectives: actes du colloque organisé par l'équipe du Trésor de la langue française au Québec et tenu à l'Université Laval les 11 et 12 avril 1985*. Québec: Presses de l'Université Laval.

Botley, Simon Philip, Anthony Mark McEnery and Andrew Wilson, eds. (2000). *Multilingual Corpora in Teaching and Research.* Amsterdam: Rodopi.

Bowker, Lynne (2003). "Teaching Translation Technology: Towards an Integrated Approach." *Tradução & Comunicação 12*, pp. 65-79.

---------- (2003b). "Corpus-based Applications for Translator Training: Exploring the Possibilities. In Granger, *et al.*, eds., pp. 169-183.

---------- (2002). *Computer-aided Translation Technology: A Practical Introduction.* Ottawa: University of Ottawa Press.

Bowker, Lynne and Peter Bennison (2003). "Student Translation Archive and Student Translation Tracking System: Design, Development and Application." In Zanettin, *et al.*, eds., pp. 104-117.

Bowker, Lynne and Jennifer Pearson (2002). *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Bowker, Lynne, Michael Cronin, Dorothy Kenny, Jennifer Pearson, eds. (1998). *Unity in Diversity? Current Trends in Translation Studies.* Manchester: St. Jerome.

Brisset, Annie (1996; 1990). *A Sociocritique of Translation: Theatre and Alterity in Québec, 1968-1988.* Toronto: University of Toronto Press. Translated by Rosalind Gill and Roger Gannon.

Brower, Reuben A., ed. (1966). *On Translation.* Cambridge, Massachusetts: Harvard University Press.

Burnett, Scott (1999). *A Corpus-based Study of Translational English.* Manchester: UMIST. Unpublished M.Sc. dissertation.

Camargo, Diva Cardoso de (2003). "An Analysis in Electronic Format of a Parallel Corpus of Journalistic and Technical Texts." *Tradução & Comunicação maio 2003*:12, pp. 15-34.

---------- (2001). "Corpus-based Translation Research on Legal, Technical and Corporate Texts." *Across Languages and Cultures 2*:1, pp. 113-125.

Canada: Department of Justice (1985). *An Act Respecting the Status and Use of the Official Languages of Canada R.S., 1985, c. 31 (4th Supp.: 1988, c. 38,*

assented to 28th July, 1988). Accessed August 2004: <http://laws.justice.gc.ca/en/O-3.01/text.html>.

Chesterman, Andrew (2004). "Beyond the Particular." In Mauranen and Kujamäki, eds., pp. 33-49.

---------- (2000). "Translation Typology." In Veisbergs and Zauberga, eds., pp. 49-62. Accessed April 2004: <http://www.helsinki.fi/~chesterm/2000bTypes.html> .

---------- (1999). "Description, Explanation, Prediction: A Response to Gideon Toury and Theo Hermans." In Schäffner, ed., pp. 90-97.

---------- (1998). "Communication Strategies, Learning Strategies, and Translation Strategies." In Malmkjaer, ed., pp. 135-143.

---------- (1997). *Memes of Translation*. Amsterdam: John Benjamins.

---------- (1995). "The Successful Translator: The Evolution of *Homo Transferens*." *Perspectives: Studies in Translatology 3*:2, pp. 253-270.

---------- (1993). "From 'Is' to 'Ought': Laws, Norms and Strategies in Translation Studies." *Target 5*:1, pp. 1-20.

Chuquet, Hélène (2003). "Loss and Gain in English Translations of the French *Imparfait*." In Granger *et al.*, eds., pp. 105-122.

Darbelnet, Jean (1983). "La norme lexicale et l'anglicisme au Québec." In Bédard *et al.*, eds., pp. 603-624.

---------- (1979). "Le maintien du français face à l'anglais au Québec." In Valdman, ed., pp. 61-74.

---------- (1970). "Dictionnaires bilingues et lexicologie différentielle." *Langages 19*, pp. 92-102.

Delisle, Jean (1993). *La traduction raisonnée: Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*. Ottawa, Canada: Presses Universitaires d'Ottawa. Collection "Pédagogie de la traduction."

---------- (1981). *L'Enseignement de l'interprétation et de la traduction: de la théorie à la pédagogie*. Ottawa: Éditions de l'Université d'Ottawa.

Delisle, Jean, Hannelore Lee-Jahnke and Monique Catherine Cormier, eds. (1999). *Terminologie de la traduction*. Amsterdam: John Benjamins.

Dodd, Bill, ed. (2000). *Working with German Corpora*. Birmingham: University of Birmingham Press.

Duff, Alan (1981). *The Third Language: Recurrent Problems of Translation into English*. Oxford: Pergamon Press.

Du-Nour, Miryam (1995). "Retranslation of Children's Books as Evidence of Changes of Norms." *Target 7:*2, pp. 327-346.

Dyvik, Helge (2004). "Translations as Semantic Mirrors: From Parallel Corpus to Wordnet." Project Report, University of Bergen, Norway. In collaboration with Martha Thunes and Gunn Inger Lyse. Accessed April 2004: <http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/ICAMEpaper.pdf>.

---------- (1998). "A Translational Basis for Semantics." In Johansson and Oksefjell, eds., pp. 51-86. Accessed April 2004: <http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/transem.html>.

Ebeling, Jarle (1998). "Contrastive Linguistics, Translations, and Parallel Corpora." In Laviosa, ed., pp. 602-615.

---------- (1998b). "Using Translations to Explore Construction Meaning in English and Norwegian." In Johansson and Oksefjell, eds., pp. 169-198.

Englund Dimitrova, Birgitta (forthcoming). *Expertise and Explicitation in Translation: A Study of Russian-Swedish Translation.*

---------- (2003). "Explicitation in Russian-Swedish Translation: Sociolinguistic and Pragmatic Aspects." In Englund Dimitrova and Pereswetoff-Morath, eds., pp. 21-31. Published on-line. Accessed April 2004: < http://www.tolk.su.se/index20.php>, link entitled "Article on Explicitation."

Englund Dimitrova, Birgitta and Alexander Pereswetoff-Morath, eds. (2003). *Swedish Contributions to the Thirteenth International Congress of Slavists, Ljubljana, 15-21 August 2003, Lund.*

Eskola, Sari (2004). "Untypical Frequencies in Translated Language: A Corpus-based Study on a Literary Corpus of Translated and Non-translated Finnish." In Mauranen and Kujamäki, eds., pp. 83-97.

Even-Zohar, Itamar (1990; 1978). "The Position of Translated Literature Within the Literary Polysystem." *Poetics Today 11*:1, pp. 45-51. Original version in Holmes *et al.,* eds., pp. 117-127. Revised 1990 version published on-line. Accessed 2002: <http://www.tau.ac.il/~itamarez/ps/pos_trli.htm>.

Fabricius-Hansen, Catherine (1998). "Informational Density and Translation, with Special Reference to German-Norwegian-English." In Johansson and Oksefjell, eds., pp. 197-234.

Faerch, Claus and Gabriele Kasper, eds. (1983). *Strategies in IL Communication.* London: Longman.

Fawcett, Peter (1997). *Translation and Language: Linguistic Theories Explained.* Manchester: St. Jerome Press.

Folkart, Barbara (forthcoming). *Second Finding: A Poetics of Translation.* Ottawa: University of Ottawa Press.

---------- (1991). *Le conflit des énonciations: Traduction et discours rapporté.* Candiac, Québec: Les Éditions Balzac.

---------- (1990). "La fonction heuristique de la traduction." *Meta 35*:1, pp. 37-44. Accessed September 2003: <http://www.erudit.org/revue/meta/1990/v35/n1/002754ar.pdf>.

---------- (1989). "Translation and the Arrow of Time." *TTR 2*:1, pp. 19-50.

---------- (1981) "L'enseignement de la traduction technique: une approche formelle du discours technique." *Revue de l'université d'Ottawa/University of Ottawa Quarterly 51*:3, pp. 505-521. Reprinted in Delisle (1981).

---------- (1988). "Cohesion and the Teaching of Translation." *Meta 33*:2, pp. 142-155. Accessed July 2002: <http://www.erudit.org/revue/meta/1988/v33/n2/002755ar.pdf>.

---------- (1984) "A Thing-bound Approach to the Practice and Teaching of Technical Translation." *Meta 29*: 3, pp. 229-246. Accessed September 2003 <http://www.erudit.org/revue/meta/1984/v29/n3/002751ar.pdf>.

France, Peter, ed. (2000). *The Oxford Guide to Literature in English Translation.* Oxford: Oxford University Press.

Frawley,William (1984). "Prolegomenon to a Theory of Translation." In Frawley, ed., pp. 159-175.

Frawley,William, ed. (1984). *Translation: Literary, Linguistic, and Philosophical Perspectives*. London: Associated University Presses.

Gellerstam, Martin (1986). "Translationese in Swedish Novels Translated from English." In Wollin and Lindquist, eds., pp. 88-95.

Gerloff, Peter (1986). "Second-Language Learners' Reports on the Interpretive Process: Talk-aloud Protocols of Translation." In House and Blum-Kulka, eds., pp. 243-262.

Gerzymisch-Arbogast, Heidrun and Klaus Mudersbach (1998). *Methoden des wissenschaftlichen Übersetzens*. Tübingen: Francke.

Gile, Daniel (1995). *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam: John Benjamins.

González Davies, Maria, Christopher Scott-Tennent, Fernanda Rodríguez Torras (2001). "Training in the Application of Translation Strategies for Undergraduate Scientific Translation Students." *Meta 46*: 4, pp. 737-744.

Government of Canada (1985). *Official Languages Act: An Act Respecting the Status and Use of the Official Languages of Canada*. Consolidated Statutes and Regulations. Accessed 2002: <http://laws.justice.gc.ca/en/o-3.01/text.html>

Granel, Gérard, éd. (1985). *Les tours de Babel*. Mauvezin: Trans-Europ-Repress.

Granger, Sylviane, Jacques Lerot and Stephanie Petch-Tyson, eds. (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi.

Gutt, Ernst-August (2000; 1991). "Translation as Interlingual Interpret ive Use." Chapter 5 in *Translation and Relevance: Cognition and Context*. Oxford: Blackwell, pp.100-122. Reprinted in Venuti, ed. (2000), pp. 376-396.

Halverson, Sandra (1998). "Translation Studies and Representative Corpora: Establishing Links Between Translation Corpora, Theoretical/Descriptive Categories and a Conception of the Object of Study." *Meta 43*:4. Accessed 2003: <http://www.erudit.org/revue/meta/1998/v43/n4/003425ar.html>.

Hansen, Silvia and Elke Teich (2001). "Multilayer Analysis of Translation Corpora: Methodological Issues and Practical Implications." *Proceedings: Eurolan (Romania) 2001, Summer Institute on "Creation and Exploitation of Annotated Language Resources"; Workshop: Multi-layer Corpus-based Analysis*. Accessed June 2003 <http://www.coli.uni-sb.de/~hansen/hansenteich.pdf> and <http://www.racai.ro/EUROLAN-2001/page/resources/workshops/corpus/>.

Harri Jantunen, Jarmo (2004). "Untypical Patterns in Translations: Issues on Corpus Methodology and Synonymity." In Mauranen and Kujamäki, eds., pp. 101-126.

Harris, Brian (1990). "Norms in Interpretation." *Target 2*:1, pp. 115-119.

Harvey, Keith, ed. (2002). *CTIS Occasional Papers Volume 2*. Manchester: UMIST.

Hasselgård, Hilde (1998). "Thematic Structure in Translation Between English and Norwegian." In Johansson and Oksefjell, eds., pp. 145-167.

Hermans, Theo (1999). "Translation and Normativity." In Schäffner, ed., pp. 50-71.

---------- (1996). "Norms and the Determination of Translation: A Theoretical Framework." In Álvarez and Vidal, eds., pp. 25-51.

Hermans, Theo, ed. (1985). *The Manipulation of Literature: Studies in Literary Translation*. London: Croom Helm.

Hervey, Sándor and Ian Higgens (1992). *Thinking Translation: A Course in Translation Method: French to English*. London: Routledge.

Holmes, James S. (1988). "The Name and Nature of Translation Studies." In Holmes, ed., pp. 67-80.

Holmes, James S., ed. (1988). *Translated!: Papers on Literary Translation and Translation Studies*. Amsterdam: Rodopi. Introduction by Raymond van den Broeck.

Holmes, James S., José Lambert, and Raymond van den Broeck, eds. (1978). *Literature and Translation*. Leuven: Acco.

House, Juliane (2001). In Steiner and Yallop, eds., pp. 127-160.

---------- (1997). *Translation Quality Assessment: A Model Revisited.* Tübingen: Gunter Narr.

---------- (1977). *A Model for Translation Quality Assessment.* Tübingen: Gunter Narr.

House, Juliane and Shoshana Blum-Kulka, eds. (1986). *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies.* Tübingen: Narr.

Hutchins, John (1979). "Linguistic Models in Machine Translation." *UEA Papers in Linguistics 9*, January 1979, pp. 29-52. PDF file accessed April 2004: <http://ourworld.compuserve.com/homepages/WJHutchins/UEAPIL-1979.pdf>.

Ignatieff, Michael (1993). *Blood and Belonging: Journeys into the New Nationalism.* Toronto: Penguin Books Canada.

Ivanova, Adelina (1998). "Educating the 'Language Élite'—Teaching Translation for Translator Training." In Malmkjaer, ed., pp. 91-109.

Jakobson, Roman (1966; 1959). "On Linguistic Aspects of Translation." In Brower, ed., 232-239. Reprinted in Venuti, ed., pp. 113-118.

Jaskanen, Susanna (1999). *On the Inside Track to Loserville, USA:Strategies Used in Translating Humour in Two Finnish Versions Of Reality Bites.* Pro gradu [Master's] thesis, University of Helsinki. On-line publication (HTML and PDF). Accessed May 2002 and April 2004: <http://ethesis.helsinki.fi/julkaisut/hum/engla/pg/jaskanen>.

Johansson, Mats (2002). *Clefts in English and Swedish: A Contrastive Study of IT-clefts and WH-clefts in Original Texts and Translations.* PhD dissertation. Department of English, Lund University. Accessed April 2004: <http://www.englund.lu.se/research/corpus/publications.phtml> (citation).

---------- (2002b). "English Translations as a Clue to the Structure of Swedish *Över.*" In *The Department of English in Lund Working Papers in Linguistics* 2. Accessed April 2004: <http://www.englund.lu.se/research/workingpapers/Volume-2-2002.phtml>

and <http://www.lingref.com/lwpd>. Edited by Satu Manninen and Carita Paradis, Department of English, University of Lund, Sweden.

---------- (2001). "Clefts in Contrast: A Contrastive Study of *-it* Clefts and *-wh* Clefts in English and Swedish Texts and Translations." *Linguistics 39*: 547-582.

---------- (1996). "Fronting in English and Swedish: A Text-based Contrastive Analysis." In Percy, *et al.,* eds., pp. 29-39.

Johansson, Stig (1997). "Using the English-Norwegian Parallel Corpus: A Corpus for Contrastive Analysis and Translation Studies." In Lewandowska-Tomaszczyk and Melia, eds., pp. 282-296.

Johansson, Stig and Knut Hofland (2000). "The English-Norwegian Parallel Corpus: Current Work and New Directions." In Botley, McEnery, and Wilson, eds., pp. 134-147.

Johansson, Stig and Signe Oksefjell, eds. (1998). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Amsterdam: Rodopi.

Jääskeläinen, Riita (2004). "The Fate of 'The Families of Medellín': Tampering with a Potential Translation Universal in the Translation Class." In Mauranen and Kujamäki, eds., pp. 205-214.

---------- (1993). "Investigating Translation Strategies." In Tirkkonen-Condit, ed., pp. 99-120.

Jääskeläinen, Riita and Sonja Tirkkonen-Condit (1991). "Automatised Processes in Professional vs. Non-professional Translation: A Think-aloud Protocol Study." In Tirkkonen-Condit, ed., 89-109.

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

---------- (2000). "Translators at Play: Exploitations of Collocational Norms in German-English Translation." In Dodd, ed., pp. 143-160.

---------- (2000b). "Lexical Hide-and-Seek: Looking for Creativity in a Parallel Corpus." In Olohan, ed., pp. 93-104.

---------- (1999). "The German-English Parallel Corpus of Literary Texts (GEPCOLT): A Resource for Translation Scholars." *Teanga 18*, pp. 25-42.

---------- (1999b). *Norms and Creativity: Lexis in Translated Text.* Manchester: Centre for Translation and Intercultural Studies UMIST. Ph.D Thesis.

---------- (1998). "Theme and Rheme in Irish and English: A Corpus-based Study." In *Working Papers in Language and Society, School of Applied Language and Intercultural Studies, Dublin City University*, pp. 1-25.

---------- (1998b). "Corpora in Translation Studies." In Baker, ed., pp. 50-53.

---------- (1998c). "Equivalence." In Baker, ed., pp. 76-80.

---------- (1998d). "Creatures of Habit? What Translators Usually Do with Words." *Meta 43*:4, pp. 515-523.

---------- (1997). "(Ab)normal Translations: A German-English Parallel Corpus for Investigating Normalization in Translation." In Lewandowska-Tomaszczyk and Melia, eds., pp. 387-392.

Klaudy, Kinga (1998). "Explicitation." In Baker, ed., 80-84.

Klaudy, Kinga and J. Kohn, eds. (1997). *Transferre Necesse Est. Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting, 5-7 September, 1996, Budapest, Hungary*. Budapest: Scholastica.

Kolers, P.A. and Michel Paradis (1980). "Psychological and Linguistic Studies in Bilingualism." *Canadian Journal of Psychology 34*: 287-303.

Kujamäki, Pekka (2004). "What Happens to 'Unique Items' in Learners' Translations? 'Theories' and 'Concepts' as a Challenge for Novices' Views on 'Good Translation'." In Mauranen and Kujamäki, eds., pp. 187-204.

Kussmaul, Paul (1995). *Training the Translator*. Amsterdam: John Benjamins.

Lambert, José (1994). "The Cultural Component Reconsidered." In Snell-Hornby, *et al.*, eds., pp. 17-26.

Lambert, José and Hendrik van Gorp (1985). "On Describing Translations." In Hermans, ed., pp. 42-52.

Larson, Mildred (1998; 1984). *Meaning-based Translation: A Guide to Cross-Language Equivalence*. Lanham: University Press of America. Second edition.

Larson, Mildred, editor (1992). *Translation: Theory and Practice, Tension and Interdependence. ATA (American Translators' Association) Series, Vol. 5.* New York: John Benjamins.

Laviosa, Sara (2003). "Corpora and Translation Studies." In Granger, *et al.*, eds., pp. 45-54.

---------- (2002). *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

---------- (2002b). "Europe in the Making in Translational and Non-translational English." In Harvey, ed., pp. 85-93.

---------- (2001). "Corpus and Simplification in Translation." In Petrilli, ed., pages 78-86.

---------- (2000). "*TEC*: A Resource for Studying What is 'In' and 'Of' Translational English." *Across Languages and Cultures 1*: 2, pp. 159-177.

---------- (1998). "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta 43*:4, pp. 557-570. Accessed 2003: <http://www.erudit.org/revue/meta/1998/v43/n4/003425ar.html>.

---------- (1998b). "The Corpus-based Approach: A New Paradigm in Translation Studies." *Meta 43*:4. Accessed 2003: <http://www.erudit.org/revue/meta/1998/v43/n4/003425ar.html>.

---------- (1998c). "The English Comparable Corpus: A Resource and a Methodology." In Bowker *et al.*, eds., pp. 101-112.

---------- (1997). "How Comparable Can 'Comparable Corpora' Be?" *Target 9*:2, pp. 289-319.

Laviosa, Sara, ed. (1998). *L'Approche basée sur le corpus/ The Corpus-based Approach*. Special Issue of *Meta 43*(4). Montréal: Les Presses de l'Université de Montréal.

Laviosa-Braithwaite, Sara (1998). "Universals of Translation." In Baker, ed., pp. 288-291.

---------- (1997). "Investigating Simplification in an English Comparable Corpus of Newspaper Articles." In Klaudy and Kohn, eds., pp. 531-540.

---------- (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation.* Manchester: University of Manchester Institute of Science and Technology. Unpublished doctoral dissertation.

---------- (1995). "Comparable Corpora: Towards a Corpus Linguistic Methodology for the Empirical Study of Translation." In Thelen and Lewandowska-Tomaszczyk, eds., pp. 153-163.

Lecercle, Jean-Jacques (1990). *The Violence of Language.* London: Routledge. French title : *La violence du langage.* Paris: PUF.

Lederer, Marianne (1994). *La traduction aujourd'hui: le modèle interprétatif.* Paris: Hachette.

Lefevere, André and Kennith David Jackson, eds. (1982). *The Art and Science of Translation, Disposition 7.* Special Issue.

Leuwen-Zvart, Kitty M. van (1990). "Translation and Original: Similarities and Dissimilarities, II." *Target 2*:1, pp. 69-95.

---------- (1989). "Translation and Original: Similarities and Dissimilarities, I." *Target 1*:2, pp. 151-181.

Leuwen-Zwart, Kitty M. van and Ton Naaijkens, eds. (1991). *Translation Studies, the State of the Art: Proceedings of the First James S. Holmes Symposium on Translation Studies.* Amsterdam: Rodopi.

Lewandowska-Tomaszczyk, Barbara and Patrick James Melia, eds. (1997). *Practical Applications in Language Corpora. PALC '97 Proceedings.* Lódz: Lódz University Press.

Lörscher, Wolfgang (1991). *Translation Performance, Translation Process, and Translation Strategies: A Psycholinguistic Investigation.* Tübingen: Günter Narr.

Long, Michael H. (1985). "Input and Second Language Acquisition Theory." In Gass and Madden, eds., pp. 377-393.

Mahoney, Anne (2002). "Review of Computer-aided Translation Technology: A Practical Introduction, L. Bowker." *Bryn Mawr Classical Review July 2002.* Accessed September 2002 <http://ccat.sas.upenn.edu/bmcr/2002/2002-07-28.html>.

Malmkjaer, Kirsten (1998). "Love Thy Neighbour: Will Parallel Corpora Endear Linguists to Translators?" *Meta 43*:4. Accessed June 2002 <http://www.erudit.org/revue/meta/1998/v43/n4/003545ar.html>.

Malmkjaer, Kirsten, (1998). "Andrew Chesterman, Memes of Translation: The Spread of Ideas in Translation Theory." *Target 10*:1.

---------- (1997). "Punctuation in Hans Christian Andersen's Stories and in Their Translations into English. In Poyatos, ed., pp. 151-162.

Malmkjaer, Kirsten, editor (1998). *Translation and Language Teaching.* Manchester: St. Jerome.

Mauranen, Anna (2004). "Corpora, Universals, and Interference." In Mauranen and Kujamäki, eds., pp. 65-82.

---------- (2000). "Strange Strings in Translated Language: A Study on Corpora." In Olohan, ed., pp. 119-141.

Mauranen, Anna and Pekka Kujamäki (2004). *Translation Universals: Do They Exist?* Amsterdam: John Benjamins.

McLuhan, Marshall (1962). *The Gutenberg Galaxy: The Making of Typographic Man.* Toronto: University of Toronto Press.

Millán-Varela, Carmen (1999). "Translation and the Business of Normalization: The Galician Case." In Vandaele, ed., pp. 339-356.

Milton, John (2000). "The Legacy of Descriptive Translation Studies: Introductory Paper to the Translation Section of the ICLA." *IRN Bulletin 20*, pp. 1-20.

Mossop, Brian (1998). "What is a Translating Translator Doing?" *Target 10*:2, pp. 231-266.

---------- (1990). "Translating Institutions and 'Idiomatic' Translation." *Meta 35*:2, pp. 343-355. Accessed July 2003: <http://www.erudit.org/revue/meta/1990/v35/n2/003675ar.pdf >. Revised on-

line version accessed April 2004: <http://www.geocities.com/brmossop/TranslatingInstitutionsRevised.htm >.

---------- (1983). "The Translator as Rapporteur: A Concept for Training and Self-Improvement." *Meta 28*:3, pp. 244-278.

Mounin, Georges (1963). *Les problèmes théoriques de la traduction*. Paris: Gallimard. Préface de Dominique Aury.

Munday, Jeremy (2001). *Introducing Translation Studies: Theories and Applications*. London: Routledge.

---------- (1998). "A Computer-assisted Approach to the Analysis of Translation Shifts." *Meta 43*:4. Accessed August 2002: <http://www.erudit.org/revue/meta/1998/v43/n4/003680ar.pdf>.

---------- (1997). "Systems in Translation: A Computer-assisted Systemic Analysis of the Translation of García Márquez." Ph.D Thesis, University of Bradford, U.K.

Nabokov, Vladimir (1992; 1955). "Problems of Translation: *Onegin* in English." *Partisan Review 22*:4 (1955), pp. 498-512. Reprinted in Schulte and Biguenet (1992): pp. 127-143.

Nida, Eugene (1964). *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: E. J. Brill.

Nida, Eugene and Charles R. Taber (1974; 1969). *The Theory and Practice of Translation*. Leiden: E. J. Brill.

Nilsson, Per-Ola (2004). "Translation-specific Lexicogrammar? Characteristic Lexical and Collocational Patterning in Swedish Texts Translated from English." In Mauranen and Kujamäki, eds., pp. 129-141.

---------- (2003). "Investigating Characteristic Lexical Distributions and Grammatical Patterning in Swedish Texts Translated from English." In Wilson, *et al.,* eds., pp. 99-107.

Nord, Christiane (1997). *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

---------- (1991). "Scopos, Loyalty, and Translational Conventions." *Target 3*:1, pp. 91-109.

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies.* Oxford: Routledge (Taylor and Francis).

---------- (2003). "How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus." *Target 15*:1, pp. 59 – 89.

---------- (2002). "Leave it out! Using a Comparable Corpus to Investigate Aspects of Explicitation in Translation." *Cadernos de Tradução VI*, Número IX, 2002/1, pp. 153-169. Número temático: Corpora e Tradução.

---------- (2002b). "Comparable Corpora in Translation Research: Overview of Recent Analyses Using the Translational English Corpus." *Proceedings of the LREC 2002 Workshop on Language Resources and Translation Work and Research Paris: ELRA*, pp. 5-9.

---------- (2001). "Spelling Out the Optionals in Translation: A Corpus Study." *UCREL Technical Papers Volume 13*, pp. 423-432. Special Issue: Proceedings of the Corpus Linguistics 2001 Conference. Lancaster, UCREL, University Centre for Computer Corpus Research on Language Technical Papers.

Olohan, Maeve, ed. (2000). *Intercultural Faultlines: Research Models in Translation Studies 1: Textual and Cognitive Aspects.* Manchester: St. Jerome.

Olohan, Maeve, and Mona Baker (2000). "Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation?" *Across Languages and Cultures 1*:2, pp. 141-158.

Ong, Walter J. (1982). *Orality and Literacy: The Technologizing of the Word.* London: Methuen.

Paradis, Michel (2001). "Bilingual and Polyglot Aphasia." *Handbook of Neuropsychology*, pp. 69-91. Oxford: Elsevier Science. Second Edition.

---------- (2001b). *Manifestations of Aphasia Symptoms in Different Languages.* Oxford: Pergamon Press.

---------- (2000). "Generalizable Outcomes of Bilingual Aphasia Research." *Folia Phoniatrica et Logopaedica 52*, pp. 54-64.

---------- (1995). *Aspects of Bilingual Aphasia*. Oxford: Pergamon Press.

---------- (1985). "On the Representation of Two Languages in One Brain." *Language Sciences 7*, pp. 1-39.

---------- (1984). "Aphasie et traduction." *Meta 29*:57-67.

Paradis, Michel, Goldblum, M.-C, and R. Abidi (1982). "Alternate Antagonism with Paradoxical Translation Behavior in Two Bilingual Aphasic Patients." *Brain and Language 15*:55-69.

Paradis, Michel and Lebrun, Y. (1983). "La neurolinguistique du bilinguisme." *Langages 72*: 1-13.

Pápai, Vilma (2004). "Explicitation: A Universal of Translated Text?" In Mauranen and Kujamäki, eds., pp. 143-164.

Pergnier, Maurice (1993). *Les fondements socio-linguistiques de la traduction*. Paris: Presses Universitaires de Lille.

Petrilli, Susan, ed. (2001). *Lo Stesso Altro XII*: 4. Athanor.

Poirier, Éric (2003). "Conséquences didactiques et théoriques du caractère conventionnel et arbitraire de la traduction des unités phraséologiques." *Meta 48*: 3, pp. 402-410.

Poyatos, Fernando (2002)."Punctuation as Nonverbal Communication." In Poyatos, ed., pp. 125-151.

Poyatos, Fernando, ed. (2002). *Nonverbal Communication Across Disciplines, Volume 3: Narrative Literature, Theater, Cinema, Translation*. Amsterdam: John Benjamins.

---------- (1997). *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*. Amsterdam: John Benjamins.

Puurtinen, Tiina (2004). "Explicitation of Clausal Relations: A Corpus-based Analysis of Clause Connectives in Translated and Non-translated Finnish Children's Literature. In Mauranen and Kujamäki, eds., pp. 165-176.

---------- (2003). "Nonfinite Constructions in Finnish Children's Literature: Features of Translationese Contradicting Translation Universals?" In Granger, *et al.*, eds., pp. 141-154.

---------- (1998). "Syntax, Readability and Ideology in Children's Literature." *Meta 43*:4, 557-570.

Pym, Anthony (forthcoming). "A Theory of Cross-Cultural Communication." Accessed December 2003: <http://www.fut.es/~apym/on-line/cross-cultural.pdf>.

---------- (2003). "What Localization Models Can Learn From Translation Theory." *The Lisa Newsletter XII*: 2, p. 4. Accessed October 2003: <http://www.lisa.org/archive_domain/newsletters/2003/2.4/pym.html>.

---------- (2000). "Historical Development." In France, ed., pp. 73-81.

---------- (1998). *Method in Translation History*. Manchester: St. Jerome Press.

---------- (1996). "Multilingual Intertextuality in Translation." In Penas Ibáñez, ed., pp. 207-218.

Pym, Anthony, ed. (1988). *L'internationalité littéraire*. Paris: Noesis.

Reiß, Katharina (1978). "Texttyp und Übersetzungsmethode: Der operative Text." *Germanistik 19*, pp. 630-631.

Rézeau, Pierre (1986). "Les régionalismes et les dictionnaires de français." In *La Lexicographie québécoise: Bilan et perspectives*. Québec: Presses de l'Université Laval.

Riccardi, Alessandra, ed. (2002). *Translation Studies: Perspectives on an Emerging Discipline.* Cambridge: Cambridge University Press.

Riffaterre, Michel (1992; 1985). "Transposing Presuppositions on the Semiotics of Literary Translation." *Texte: Revue de critique et de théorie 4* (1985), pp. 99-110. Reprinted in Schulte and Biguenet, eds. (1992), pp. 204-217.

Robinson, Douglas (1997). *Western Translation Theory: From Herodotus to Nietzsche*. Manchester: St. Jerome Publishing.

---------- (1997b). "Translation and the Repayment of Debt." *Delos 7*: 1-2, pp. 10-22.

---------- (1997c). *Becoming a Translator: An Accelerated Course*. London: Routledge.

Sager, Juan C. (1994). *Language Engineering and Translation: Consequences of Automation*. Amsterdam: John Benjamins.

---------- (1984). "Reflections on the Didactic Implications of an Extended Theory of Translation." In Wilss *et al.*, eds., pp. 333-343.

Schäffner, Christina, ed. (1999). *Translation and Norms*. Clevedon: Multilingual Matters. Reprint of *Current Issues in Language and Society 5*: 1-2.

Schmied, Josef and Hildegard Schäffler (1997). "Explicitness as a Universal Feature of Translation." In Ljung, ed., pp. 21-34.

Schulte, Rainer and John Biguenet (1992). *Theories of Translation: An Anthology of Essays from Dryden to Derrida*. Chicago: University of Chicago Press.

Scott, M. N. (1998). *Normalisation and Readers' Expectation: A Study of Literary Translation with Reference to Lispector's A Hora da Estrela*. Ph.D Thesis. Liverpool: AELSU University of Liverpool. Summarized in Laviosa 2002, pp. 68-69.

Séguinot, Candace (1991). "A Study of Student Translation Strategies." In Tirkkonen-Condit, ed., pp. 79-88.

Selinker, Larry (1972). "Interlanguage." *IRAL 10*, pp. 209-231.

---------- (1969). "Language Transfer." *General Linguistics 9*, pp. 67-92.

Shamaa, Najah (1978). *A Linguistic Analysis of Some Problems of Arabic to English Translation*. Oxford University: Oxford University Press. Ph.D Thesis.

Shlesinger, Miriam (1991). "Interpreter Latitude versus Due Process: Simultaneous and Consecutive Interpretation in Multilingual Trials." in Tirkkonen-Condit, ed., pp. 147-155.

---------- (1989). *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-literate Continuum*. M.A. Thesis, Tel Aviv University.

Shuttleworth, Mark and Moira Cowie (1997). *Dictionary of Translation Studies*. Manchester: St. Jerome Publishing.

Somers, H.L., ed. (1996). *LSP, Terminology and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: John Benjamins.

Snell-Hornby, Mary, Franz Pöchhacker and Klaus Kaindl, eds. (1994). *Translation Studies: An Interdiscipline.* Amsterdam: John Benjamins.

Snel Trampus, Rita D. (2002). "Aspects of a Theory of Norms and Some Issues in Teaching Translation." In Riccardi, ed., pp. 38-55.

Steiner, Erich and Colin Yallop, eds. (2001). *Exploring Translation and Multilingual Text Production: Beyond Content.* Berlin: Mouton de Gruyter.

Stewart, Dominic (2000). "Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation." In Olohan, ed., pp. 73-91.

Teubert, Wolfgang (2002). "The Role of Parallel Corpora in Translation and Multilingual Lexicography." In Altenberg and Granger, eds., pp. 189-214.

---------- (1999). "Corpus Linguistics: A Partisan View." *TELRI Newsletter 8:*99, pp. 4-19. Cited in Stubbs (2001: 221). Accessed April 2004: <http://tractor.bham.ac.uk/ijcl/teubert_cl.html>.

---------- (1999b) "Korpslinguistik und Lexikographie." *Deutsche sprache 27*:4, pp. 292-313.

Thelen, Marcel and Barbara Lewandowska-Tomaszczyk, eds. (1995). *Translation and Meaning: Part 3. Proceedings of the Maastricht Session of the 2nd International Maastricht-Lodz Duo Colloquium on "Translation and Meaning," Held in Maastricht, Netherlands, 19-22 April 1995.* Maastricht: Hogeschool Maastrict UPM.

Thunes, Martha (1998). "Classifying Translational Correspondences." In Johansson and Oksefjell, eds., pp. 25-50.

Tirkkonen-Condit, Sonja (2004). "Unique Items—Over- or Under- Represented in Translation Studies?" In Mauranen and Kujamäki, eds., pp. 177-184.

---------- (2002). "Translationese: A Myth or an Empirical Fact? A Study into the Linguistic Identifiability of Translated Language." *Target 14*:2, pp. 207-220.

---------- (2000). "In Search of Translation Universals: Non-equivalence or 'Unique' Items in a Corpus Text." Paper presented at Research Models in Translation Studies. UMIST and UCL Manchester 28-30 April 2000. Summarized in Laviosa 2002, p. 73.

Tirkkonen-Condit, Sonja, ed. (1991). *Empirical Research in Translation and Intercultural Studies: Selected Papers of the TRANSIF Seminar, Savonlinna 1988*. Tübingen: Gunter Narr.

Toury, Gideon (2004). "Probabilistic Explanations in Translation Studies: Welcome as They Are, Would They Still Qualify as Universals?" In Mauranen and Kujamäki, eds., pp. 15-31.

---------- (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

---------- (1993). "Probabilistic Explanations in Translation Studies: Plenary Lecture in the Postgraduate Seminar Translation Theory and Research." University of Joensuu: Savonlinna School of Translation Studies.

---------- (1991). "What are Descriptive Studies into Translation Likely to Yield Apart from Isolated Descriptions?" In Leuwen-Zwart, *et al.*, eds., pp. 179-192.

---------- (1981). "Translated Literature: System, Norm, Performance: Toward a TT-oriented Approach to Literary Translation." *Poetics Today 2*:4, pp. 9-27. Republished in Toury 1980, pp. 35-50.

---------- (1980). *In Search of a Theory of Translation*. Tel Aviv: The Porter Institute for Poetics and Semiotics.

---------- (1978). "The Nature and Role of Norms in Literary Translation." In Holmes *et al.*, eds., pp. 83-100.

---------- (1977). *Normot shel tirgum ve-ha-tirgum ha-sifruti le-Ivrit ba-shanim 1930-1945/Translational Norms and Literary Translation into Hebrew, 1930-1945*. Tel Aviv: The Porter Institute for Poetics and Semiotics.

TTR 12:2. *Poésie, cognition, traduction II: Autour d'un poème de W. H. Auden/ Cognition, Translation II: On a Poem by W. H. Auden.*

Tymoczko Maria (1998). "Computerized Corpora and the Future of Translation Studies." *Meta 43*: 4, pp. 652-660.

Ulrych, Margherita (2002). "An Evidence-based Approach to Applied Translation Studies." In Riccardi, ed., pp. 198-213.

---------- (1999). *Focus on the Translator in a Multidisciplinary Perspective*. Padua: Unipress.

Vandaele, Jeroen, ed. (1999). *Translation and the (Re)Location of Meaning: Selected Papers of the CETRA Research Seminars in Translation Studies, 1994-1996.* Leuven : CETRA.

Vanderauwera, Ria (1985). *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature*. Amsterdam: Rodopi.

Veisbergs, Andrejs and Ieva Zauberga, eds. (2000). *The Second Riga Symposium on Pragmatic Aspects of Translation*. Riga: University of Latvia.

Venuti, Lawrence (2002). "The Difference that Translation Makes: The Translator's Unconscious." In Riccardi, ed., pp. 214-241.

---------- (1998). *The Scandals of Translation: Towards and Ethics of Difference.* London: Routledge.

---------- (1995). *The Translator's Invisibility: A History of Translation.* London: Routledge.

Venuti, Lawrence, ed. (2000). *The Translation Studies Reader*. London: Routledge.

Vinay, J.-P. and J. Darbelnet (1977; 1958). *Stylistique comparée du français et de l'anglais*. Montréal: Beauchemin.

Vinay, J.-P. and J. Darbelnet (1995). *Comparative Stylistics of French and English: A Methodology for Translation.* Amsterdam: John Benjamins. Translated and edited by J.C. Sager and M.-J. Hamel.

Weaver, William (1989). "The Process of Translation." In Biguenet and Schulte, eds., pp. 117-124.

Weinreich, Uriel (1953). *Languages in Contact: Findings and Problems*. The Hague: Mouton.

Weissbrod, Rachel (1992). "Explicitation in Translations of Prose-fiction from English to Hebrew as a Function of Norms." *Multilingua 11*: 2, pp. 153-171.

----------(1992b). "Translation of Prose-Fiction from English to Hebrew in the 1950s and 1960s: A Function of Norms." In Larson, ed., pp. 206-223.

---------- (1990). "Linguistic Interference in Literary Translations from English to Hebrew of the 1960s and 1970s." *Target 2*:2, pp. 165-181.

Wilson, Andrew, Paul Rayson, and Tony McEnery, eds. (2003). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World.* München: Lincom-Europa.

Wilss, Wolfram and Gisela Thome, eds. (1984). *Die Theorie des Übersetzens und ihr Aufschlußwert für die Übersetzungs- und Dolmetschdidaktik / Translation Theory and Its Implementation in the Teaching of Translating and Interpreting*. Tübingen: Gunter Narr.

Wollin, Lars and Hans Lindquist, eds. (1986). *Translation Studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II, Lund, 14-15 June, 1985*. Malmö: CWK Gleerup.

Zanettin, Frederico (2001). "Swimming in Words: Corpora, Translation and Language Learning." In Aston, ed., pp. 177-197.

---------- (2000). "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis." In Olohan, ed., pp. 105-118.

Zanettin, F., S. Bernardini, and D. Stewart, eds. (2003). *Corpora in Translator Education*. Manchester: St. Jerome.

Øverås, Linn (1998). "In Search of the Third Code: An Investigation of Norms in Literary Translation." *Meta 43*:4, pp. 571-588.

Glossary and Legend

Please note: These glossary entries are provided for the reader who may not be familiar with all of the terminology used in this text. They are meant to be understood in the context of the present study and not as formal definitions.

Average: 1. The middle value in a pre-defined continuum of qualities, such as readability in texts, or height in human beings. 2. The middle value in an array of numbers, which can be calculated in three different ways, as the arithmetic **Mean**, as the **Median**, or as the **Mode** or modal value. See **Centrality**.

ASL: Abbreviation for "Average Sentence Length" (i.e. [Arithmetic] Mean Sentence Length). In the field of readability instruments and in corpus linguistics, the mean length (in number of words) of a set of texts. This is a prime measure for comparing lengths of T-units in two sets of texts, as is done in the present study. See **T-unit**.

BNC: The abbreviation commonly used for "British National Corpus." An electronic corpus of contemporary British English, consisting of spoken and written texts taken from a large number of sources and "designed to represent as wide a range of Modern British English as possible" (British National Corpus 2002). Compiled by a consortium of academic and industrial institutions, the BNC has an overall size of over 100 million words (*ibid.*).

Centrality: The tendency of the values in a set of data to occur in bunches around a central value. Can be measured by calculating the Mean, the Median, and the Mode. See **Clustering** and **Dispersion**.

Clustering: The tendency of the values in a set of data to occur in bunches instead of being evenly dispersed over the range of values. See **Centrality** and **Dispersion**.

Coinage, transient or attempted: An unattested word or short phrase that did not appear be part of the Canadian lexicon at the time the corpora were gathered for our research. In either of the corpora used in this study, a word or short phrase that is not attested in any of a range of dictionaries and reference works consulted, and that is not found in TEXTUM or among site with Canadian domain names on the World Wide Web. Such coinages are assumed to have been attempted by a given writer or translator without their achieving permanent widespread use in the language. In the present synchronic study, we therefore take the precaution of acknowledging that the coinages retained and discussed are quite possibly "transient," because it cannot be known whether they will survive. Barring future evidence to the contrary, they are assumed to be ephemeral.

Comparable corpora: Electronic corpora consisting of two sets of texts in one and the same language; in the present study, a set of translated texts and a set of non-translated texts in the same language, of approximately the same word length, and gathered from similar sources.

Concordancer: Software that retrieves from a corpus all occurrences of a search term, displaying them in an easily read format. See Bowker (2002: 53).

Content words: A class of words whose role is to convey the meaning, or semantic content, of a text. Also called "lexical" words (Baker 1995: 237; Crystal 1997: 222; Quirk *et al.* 1985: 68-69). The class of content words has many members, and is open to new words (for example, new nouns and verbs). It consists generally of nouns, adjectives, adverbs, and main verbs. Related term: **function word.**

Corpus (Plural *Corpora*): "A large collection of texts in electronic form selected according to explicit criteria" (Bowker 2003: 169; 174). In the present study, a selected sample of Canadian texts gathered to represent and compare translated and non-translated English and French.

Corpus-based Translation Studies (CTS): The branch of translation studies that uses corpora of translated and non-translated texts "for the empirical study of the product and process of translation" (as in the present study) and for the training of translators. CTS uses a "rigorous and flexible" methodology, theoretical principles that are based on empirical observations, and a combination of inductive and deductive approaches (Laviosa 2003: 45).

Corpus linguistics: The branch of linguistics that studies collections of electronic texts that are samples of naturally occurring language.

Descriptive Translation Studies (DTS): The field of study devoted to examination of the product, the function, and the process of translation, following Holmes (1988) and Toury (1995:10).

Dispersion: In basic statistics, the degree to which a set of data points are spread out from their central value, above and below their mean. Determined, for the purposes of the present study, by calculating the standard deviation of a given set of data values. See **Centrality**, **Clustering**, and **Standard deviation**.

Dispersion Plot: In the WordSmith Tools'Concordancer, a display that shows where the search word occurs in the file to which the current entry of the search term belongs, allowing the researcher to see at a glance whether the use of the word is concentrated in a single area of the corpus.The Dispersion Plot function (Version 4) displays the source text file name, the number of words in the source

text, the number of occurrences of the search-word per 1,000 words, how many occurrences there are per 1,000 words, the "plot dispersion value," and a visual display consisting of a line marking each occurrence within the corpus.

EST: In the present study, abbreviation for "English Source Text": the designation used for the Non-translated English Corpus.

ETT: In the present study, abbreviation for "English Target Text": the designation used for the Translated English Corpus.

Equalizing Effect: A recurrent feature of translation hypothesized by Shlesinger (1989: 170-171), in which texts that are interpreted (i.e. translated) will tend to be positioned toward the centre of the oral-literate continuum, compared to non-translated texts. The prototype of the hypothesis of **Levelling-out**.

Explicitation: The hypothesis according to which "a range of textual phenomena" will provide consistent evidence that there is "an overall tendency to spell things out rather than leave them implicit" in a corpus of translated texts (Baker 1996: 176; 180-181). In the present study, explicitation is assumed to be comparative: the "textual phenomena" selected as measures are predicted to be more frequent in a corpus of translated texts than in a comparable corpus of non-translated texts.

Frequency list: A list of words in a corpus, given in order of the number of times each word appears, from the words occurring most often to words occurring only once.

FST: In the present study, abbreviation for "French Source Text": the designation used for the French Non-translated Corpus.

FTT: In the present study, abbreviation for "French Target Text": the designation used for the French Translated Corpus.

Function words: A class of words whose role is mainly grammatical, e.g. articles, pronouns, conjunctions (Crystal 1997: 162). Also called "form word," "grammatical word" (Quirk *et al.* 1985: 68-69; Baker 1995: 237), "functor," or "empty word" (*ibid.*). Serves to link **content words** syntactically, in a sentence. A class of words that has few members and few or restricted inflections, and that is rarely open to new words. There is no set semantic boundary between content and function words. For example, pronouns may function deictically, by referring to something outside the text; modal verbs (*must, can, should* etc.) may convey content by expressing ability, permission, or obligation. Related term: **content word.**

GOL: The official abbreviation for the Government On-Line initiative, the clearinghouse project for making Government of Canada services and information available to the general public via the Internet. In the present study, the contact list for government departments and agencies listed (in 2002) as participants in the initiative was used to compile a list of email addresses to which the request for material (corpus texts) was sent. See Bibliography, Corpora: Government of Canada (2002).

Hapax legomena (Singular *-on*): Word forms (types) that occur only once in a corpus. From the ancient Greek, with the literal meaning of "things that happen only once."

ICA: The official abbreviation for the Institut Canadien des Actuaires.

Keyword List: A program included in the WordSmith Tools suite, and which compares the word list of a single text with the word list of a larger set of reference texts. The "key words" displayed by the program are those in the shorter text that have an unusually high frequency compared to the overall word frequencies of the reference texts. See **WordList**.

KWIC: Abbreviation for "Key Word In Context." A common display format of concordancers, in which all occurrences of a search term in a corpus are lined up in the centre of the screen, and in which the context on either side of the search term occurrence is displayed in an order specified by the user (Bowker 2002: 53-54).

Levelling out: The hypothesis according to which there is an observable tendency to "gravitate towards the centre of a continuum" in a corpus of translated texts (1996: 184-185). This tendency, according to Baker, is "neither target-language nor source-language dependent" (*ibid.*). The "notions of centre and periphery" of this continuum can be internally generated (e.g. from the minimum to the maximum of a group's scores), and therefore "defined from within the translation corpus itself" as in Laviosa-Braithwaite (1996; cited in Baker 1996: 184). The extremes of the continuum may also be independently established (as in the oral-literate continuum used in Shlesinger 1989; cited in Baker 1996: 184-185). In either case, the hypothesis of levelling-out predicts that translated texts will tend to be positioned toward the centre of the continuum, compared to non-translated texts. See **Equalizing effect**.

Lexical density ratio: The proportion of the total number of content words (also called "lexical" words; generally nouns, adjectives, adverbs, and main verbs) to the total number of running words in a given text. This ratio is usually expressed as a percentage. The number of content words is often considered to be equal to the number of running words minus the number of function words, the latter of which may be counted by a concordancer using a stoplist.

List head: The 108 most frequent words in a corpus, obtained from the word frequency list of a concordancer.

Mean, arithmetic: A value that is calculated by dividing the sum of a set of terms by the number of terms. See **Centrality** and **Clustering**.

Median: In statistics, the value below which 50% of the cases in a sample fall. The middle value of an ordered set of values. See **Centrality** and **Clustering**.

Mode: In statistics, the most common value in a data set. In the set 1, 2, 2, 3, 4, 7, the value 2 occurs most often and is therefore the mode, or modal value. The mode is used as a measure of centrality in cases where the mean is distorted by a small number of extreme values. See **Centrality** and **Clustering**.

Normalization: The hypothesis according to which there is an observable tendency to conform to the "typical patterns" of the target language in a corpus of translated texts (Baker 1996: 183-184). Baker also calls this hypothetical recurrent feature of translation "conservatism," and notes that non-standard forms, whether "experimental" or unintentional, will tend to be normalized in translation (*ibid.*). In the present study, normalization is assumed to be comparative: the selected measures are predicted to be more frequent in a corpus of translated texts than in a comparable corpus of non-translated texts.

Norm, sociolinguistic: A notion of correctness in language use that is believed by individual speakers to be widely held in a community of speakers. See Bartsch (1987:4).

Norm, translation: A notion of correctness in the process and product of translation that is believed by individual translators to be widely held in a community of translators. See Gideon Toury's *oeuvre* (particularly Toury 1980; 1995).

ODC: In the present study, the abbreviation used for "Object of a Defining Clause"; optional *that* or *which* in clauses such as *Shells my sister gathered* or *Lies my father taught me*. Defining clauses are also known as "restrictive" clauses.

Parallel corpus: An electronic corpus consisting of source text (language A: original) and target text (language B: translation) pairs.

Readability: The difficulty or ease with which a text is read by a given population. Readability is distinguished from legibility (how easily characters are recognized) and interest.

Readability index (pl. indices): An instrument for measuring the degree of complexity of a written text's form and of correlating the results with a given reading population's skill level.

Recurrent features of translation: In the present study, characteristics that are observed to be more frequent in sets of translated texts than in similar sets of non-translated texts in the same language. See **Translation universals**.

Register: In stylistics and sociolinguistics, the range of varieties of language and attitude appropriate in different social situations and fields of activity (Crystal 1997: 327; Quirk *et al.*1985: 23-26). For the purposes of the present study, this range is assumed to form a continuum from "very informal" to "very formal" writing.

Running words: The total number of words in a corpus. See **Token**.

Simplification: The hypothesis according to which there is an observable "tendency to simplify the language used" in a corpus of translated texts (Baker 1996: 181-183). In the present study, simplification is assumed to be comparative: the selected measures are predicted to be more frequent in a corpus of translated texts than in a comparable corpus of non-translated texts.

SL: In Translation Studies, the abbreviation commonly used for "source language," the language in which the source text is written (or spoken). See **ST**.

ST: In Translation Studies, the abbreviation commonly used for "source text," the original source from which a translation is made. See **TT**.

Standard deviation: The most common statistical means of representing dispersion, calculated as the square root of the value by which a score deviates from the mean score. See **Dispersion**.

Standardization, statistical: In corpus linguistics, a technique for comparing two corpora of different sizes, where a statistic such as type/token ratio is calculated automatically every ***n*** words, and the results are averaged. For large compiled corpora, ***n*** = 1000 is a conventional value for statistical standardization. With corpora made up of a long series of short individual texts, the value for ***n*** should be close to the number of words of the shortest texts.

Stoplist: A list of words that may be used as an optional setting for concordancing software, and that excludes the words on the list from the count made by the concordancer.

T-unit: In readability instruments, and for the purposes of the present study, a key measure of syntactic complexity denoting paratactic or hypotactic independent clauses. "A single main clause (defined here as a subject, or coordinated subjects, with a finite verb or coordinated finite verbs) or independent clause…plus whatever other subordinate clauses or nonclauses are attached to, or embedded within, that

one main clause" (Hunt 1977: 92-93). An abbreviated form coined by Hunt to stand for "terminable unit."

TEC: The abbreviation commonly used for "Translational English Corpus." An electronic corpus of "contemporary translational" British English, consisting of written texts translated from a variety of European and non-European source languages. Compiled at the Centre for Translation and Intercultural Studies at the University of Manchester's School of Modern Languages, the TEC has an overall size of at least 10 million words (Baker 2003).

Text type: In the present study, the division of texts into one of two types, "non-specialized" or "specialized." This categorization is made introspectively, from the point of view of a professional translator assessing a text for the purposes of translation.

TL: In Translation Studies, the abbreviation commonly used for "target language," the language in which the target text is written (or spoken). See **TT**.

Token: A single occurrence of a word in a corpus (Bowker 2002: 47-48). See **Type/token ratio**.

Translation universals: In Translation Studies, characteristics that theoretically distinguish translation from all other types of speaking and writing, and that are observable in all products of the act of translation. See **Recurrent features of translation**.

TT: In Translation Studies, the abbreviation commonly used for "target text," the translated text. See **ST**.

Type: In corpus analysis, a single word form. See **Type/token ratio**.

Type/token/*n* ratio: The ratio of word forms (types) to running words (tokens) in a corpus of texts, which reflects the range of vocabulary in a corpus. Type/token ratio is usually standardized by *n*, the number of words in the shortest text in a corpus. For instance, if a corpus ranges in size from 50 to 1,000 words, the researcher may set the concordancer to calculate a type/token ratio every 50 words. The concordancer then calculates a mean or "standardized" type/token ratio, which may also be called a "type/token/*n* ratio." See **Standardization, statistical**.

WordList: One of the three main features of the WordSmith Tools suite. Software which produces an ordinally-ranked list of all words in a corpus. WordSmith's WordList feature produces two word lists, one in which words are ranked in descending order by their raw frequency, and another in which the words are listed in alphabetical order. WordSmith's WordList feature also produces a third list, a set of compiled statistics that includes type/token[*n*] ratios and mean sentence lengths.

Appendix I: Corpus Sources and Designations

Non-translated English (EST) Text Sources

Acronym	Source	# texts	# words ST
AG	Agriculture and Agri-Food Canada	1	7,020
CB	CPP Investment Board	5	4,062
CC	Canada Council	18	13,290
CD	CIDA	2	1,207
CP	Canada Post	21	18,178
NR	Natural Resources Canada	4	12,372
PW	Public Works Canada[220]	11	3,347
SC	Solicitor General of Canada[221]	1	2,223
		63	61,699

Translated English (ETT) Text Sources

Acronym	Source	# texts	# words TT
CC	Canada Council	17	8,674
CS	Canadian Wildlife Service	3	3,120
CT	CAPPRT	11	32,935
DT	DFAIT	17	10,196
NR	Natural Resources Canada	2	4,635
PW	Public Works Canada	3	875
		53	60,435

Non-translated French (FST) Text Sources

Acronym	Source	# texts	# words ST
CC	Canada Council	11	5,082
CS	Canadian Wildlife Service	3	5,326
CT	CAPPRT	11	33,465
DT	DFAIT	17	10,708
NR	Natural Resources Canada	2	5,141
PW	Public Works Canada	3	1,003
		47	60,725

Translated French (FTT) Text Sources

Acronym	Source	# texts	# words ST
AG	Agriculture and Agri-Food Canada	1	7,909
CB	CPP Investment Board	2	357
CC	Canada Council	8	7,497
CD	CIDA	2	1,581
CP	Canada Post	11	11,042
NR	Natural Resources Canada	3	30,908
PW	Public Works Canada	5	1,504
SC	Solicitor General of Canada	1	2,706
		33	60,798

[220] Full title: Public Works and Government Services Canada.
[221] Now Public Safety and Emergency Preparedness Canada.

Appendix II: Sample Correspondence with Sources

Original letter of inquiry, sent to all GOL participants:[222]

June 19, 2002
Subject: Doctoral research question for your translation department

Hello,
This message is addressed to your translation department. I'm doing research toward my doctoral dissertation at the University of Ottawa's School of Translation and Interpretation, and I'm hoping someone there can answer this question: it would be extremely helpful if you could tell me the direction of translation for the pages on your Web site. That is,
which of the pages were originally written in French, and which in English?
Does anyone there have a record? I don't need to know the names of the authors, just the language of the source text for as many pages as possible.
(I would be happy to come in and consult records rather than putting you to the trouble of copying titles over and over.)
If you would like to verify my credentials, please feel free to contact Dr. Jean Delisle, [...] who is chairman of the department [...], or Madame Jeanne d'Arc Turpin [...], who is the department's secretary.
I may also be reached by telephone at any hour convenient to you [...]]

Thank you very much,
--Donna A. Williams

First reply:

June 20, 2002
Dear Ms. Williams:

I am informed by our translation department that everything on our Website was originally written in English and translated into French. The only exceptions are the following documents, which were written in French and translated into English:
-Monsieur Jean-Louis Roux's speeches
-certain articles from the "For the Arts" Bulletin
-certain press releases

I hope this information is useful to you.

Sincerely,
D. Sarault
Agente d'information / Information Officer
Services aux arts / Arts Services Unit
Le Conseil des Arts du Canada / The Canada Council for the Arts

August 13, 2002

Hello Donna:
We recently made the Rural Team Quebec pages live. They were originally written in French and then translated into English.
http://www.rural.gc.ca/team/qc/quebec_e.phtml
http://www.rural.gc.ca/team/qc/quebec_f.phtml

Good-luck with your project.

---

[222] Personal information has been removed.

L. Thacker
Canadian Rural Information Service

August 8, 2002
Donna:

This is in response to your enquiry regarding language usage etc. on the CIC main Web site http://www.cic.gc.ca/. I have just recently returned from holiday, so please forgive the lateness of this reply.
1. Content on the CIC main site is generally prepared in English first and then translated to French. This is not a "policy", as content providers are free to prepare content in the official language of their choice, however here at CIC NHQ it appears to be more common to work in English first.
2. There are no exceptions to this rule that I am aware of.
3. Translations of content are generally prepared by the Translation Bureau using their policies and directives.

I hope you find this information helpful, Donna. Please don't hesitate to contact me should you need any further information. Good luck with your research!

S. Northrup
Adviser/Conseillère
New Media and Marketing/Nouveaux média et marketing
Communications Branch/Direction génerale des communications
Citizenship and Immigration Canada/Citoyenneté et Immigration Canada

Appendix III: Compiled List of English Reporting Verbs

accept
acknowledge
add
admit
advise
agree
allege
announce
answer
argue
arrange
ask
assert
assume
assure
authorize
believe
boast
call
check
chorus
claim
comment
complain
concede
conclude
confess
confirm
consider
consent
contend
contest
continue
convince
cry
decide
declare
decree
demand
demonstrate
deny
desire
determine
dictate
direct
discover
dispute
doubt
dream
elicit
emerge (it)
ensure
estimate
expect
explain
fear
feel
figure
find
foresee
forget
gather
grant
grumble
guarantee
guess
happen (it)
hear
hint
hold
hope
ignore
imagine
imply
indicate
infer
inform
inquire
insist
intend
judge
know
learn
maintain
mean
mention
move
mumble
murmur
muse
mutter
note
notice
notify
object
observe
order
overlook
perceive
permit
persuade
plead
pledge
ponder
pray
predict
prefer
pretend
proclaim
promise
prophesy
propose
prove
read
realize
reason
reassure
recall
recite
reckon
recommend
record
reflect
regret
remark
remember
remind
repeat
reply
report
request
require
resolve
respond
reveal
rule
rumoured (be)
say
see
sense
scream
shout
show
shriek
specify
state
stipulate
storm
suggest
suppose
swear
teach
tell
think
threaten
thunder
transpire (it)
understand
urge
vow
wail
warn
whisper
wish
worry
write
yell

The above is a list of all the reporting verbs which may be followed by optional *that* according to Quirk *et al.* (1973) *Grammar of Contemporary English* and COBUILD (2000) *English Grammar*. A brief overview of the explanations given by their authors is in order. Quirk *et al.* (1973:832) note that "the conjunction in *that*-clauses may be zero except when it has initial position in passive clauses (and thus obeys the same rules as other nominal clauses as subject)." Partial lists of verbs followed by several types of optional-*that* clauses are given by Quirk *et al.* (1973: 832-4) to show examples of the kind of verb that can appear in this optional-*that* structure.[223] The optional use of the conjunction *(that)* to begin reporting clauses is emphasized by COBUILD *English Grammar* (2000):

> "In informal speech and writing, the conjunction *that* is commonly omitted. [...] *That* is often omitted when the reporting verb refers simply to the act of saying or thinking. [...] This kind of reported clause is often called a *"that*-clause," even though many occur without *that*." (320-321)

Reporting *(that)* clauses occur after reporting verbs, after nouns derived from reporting verbs, and after adjective complements of a linking verb (COBUILD *English Grammar* 2000: 314-339). These reporting *(that)* clauses are used in the following ways:

1. To report a statement or "someone's thoughts" (sections 7.7, 7.9, 7.26-7.28). Eg "She announced (that) the lecture would now begin."
2. To make a suggestion about what someone should do (sections 7.40-7.41). Eg "He proposes (that) the government should hold an inquiry."
3. To leave the speaker unspecified by using impersonal-*it* as subject (sections 7.64-7.66). Eg "It was said (that) he could speak their language."
4. To refer to the hearer as a direct object (section 7.71). Eg "I told them (that) you were at the dentist."
5. To refer to fact-related actions, such as checking or proving facts (section 7.82) or learning and perceiving facts (section 7.10). Eg "Research shows (that) males will mother an infant as well as any female." Eg "Then she saw (that) he was sleeping."
6. To say that something happens, is the case, or becomes known, using impersonal-*it* (section 7.83: *happen, transpire, emerge*; see list of reporting verbs below). Eg "It happened (that) he had a client interested in that sort of thing."
7. To refer to what someone says or thinks, and to refer to or relate facts or beliefs, using nouns derived from reporting verbs (section 7.84). Eg "There was little hope (that) he would survive"; *(to) hope → (the) hope.*
8. To ascribe a cause to a person's state of mind, using an adjective that is the complement of a linking verb, usually *be* (sections 7.85-7.87). The adjective may derive from a reporting verb. Eg "I was worried (that) she'd say no"; *(to) worry → (be) worried.*
9. To comment on a fact, using an adjective complement of a linking verb that has impersonal-*it* as its subject (section 7.88). Eg "It is true (that) the authority of parliament has declined."

---

[223] The word "et cetera" appears in abbreviated form at the end of each list given by Quirk *et al* .(see 833-834), indicating that each list is partial. The reason for this is that Quirk *et al.* identify several types of verb phrase that may be followed by optional-*that* clauses: phrases with verbs in the indicative (*I suppose [that]) he is coming)*, "putative" verb phrases formed with *should* or other modal + verb in the subjunctive (*I regret [that]he should be so stubborn*), and subjunctive verb phrases (*I request [that] she go alone*). Quirk *et al.* then list two sets of of verbs that fit such structures: a list of verbs followed by a *that*-clause with the verb in the indicative, and a list of verbs followed by a *that*-clause with a subjunctive verb. Since in our study we are using a concordancer to gather all instances of optionally-included *that* clauses in our corpus, the type of verb phrase is immaterial to our search, and we have integrated the two lists into a single list of verbs that according to Quirk *et al.* may be followed by optional –*that*.

COBUILD *English Grammar* (2000: 314-340) provides thirty-seven lists of reporting verbs, "reporting" nouns, and "reporting" adjective complements of linking verbs. All three types (verbs, nouns, and adjectives) are followed by optional *that*. Our compiled list includes only those verbs given by COBUILD as "reporting."

Appendix IV: Reporting Verbs with "Zero" and *That* in the English Corpora

	EST "zero"	EST + *that*	EST all	ETT "zero"	ETT + *that*	ETT all
accept	3	1	4	9	3	12
acknowledge	4	0	4	3	3	6
add	19	0	19	5	0	5
admit	0	1	1	0	0	0
advise	1	0	1	1	1	2
agree	6	3	9	8	4	12
allege	0	0	0	0	1	1
announce	7	6	13	5	1	6
answer	1	0	1	0	0	0
argue	0	0	0	1	9	10
arrange	0	0	0	0	0	0
ask	31	0	31	13	0	13
assert	0	0	0	4	2	6
assume	3	1	4	5	1	6
assure	0	0	0	1	0	1
authorize	0	0	0	0	0	0
believe	5	9	14	7	8	15
be rumoured*	0	0	0	0	0	0
boast	0	0	0	0	0	0
call	31	0	31	11	0	11
check	4	0	4	0	0	0
chorus	0	0	0	0	0	0
claim	0	0	0	3	2	5
comment	2	1	3	0	0	0
complain	0	0	0	0	0	0
concede	0	0	0	0	0	0
conclude	0	1	1	2	9	11
confess	0	0	0	0	0	0

confirm	4	1	5	1	3	4
consider	8	0	8	0	0	0
consent	0	0	0	0	0	0
contend	0	0	0	0	0	0
contest	0	0	0	0	0	0
continue	1	0	1	4	0	4
convince	1	0	1	2	4	6
cry	0	0	0	0	0	0
decide	7	2	9	15	5	20
declare	0	0	0	2	11	13
decree	0	0	0	0	0	0
demand	0	0	0	0	0	0
demonstrate	5	1	6	1	2	3
deny	0	0	0	1	0	1
desire	0	0	0	13	0	13
determine	0	0	0	3	4	7
dictate	0	0	0	0	0	0
direct	0	0	0	5	0	5
discover	5	0	5	1	0	1
dispute	0	0	0	1	0	1
doubt	0	1	1	0	1	1
dream	0	0	0	0	0	0
elicit	0	0	0	0	0	0
emerge (it)	2	0	2	4	0	4
ensure	22	26	48	6	20	26
estimate	1	1	2	7	3	10
	EST "zero"	EST + *that*	EST all	ETT "zero"	ETT + *that*	ETT all
expect	14	1	15	0	0	0
explain	15	0	15	19	5	24
fear	0	0	0	0	0	0
FEEL[224]	11	42	53	1	3	4
figure	0	0	0	0	0	0

[224] feel (EST 14: 5); (ETT 1:0); felt (EST reporting + past tense: 39: 37);(ETT 3:3)

find[225]	2	2	4	2	8	10
foresee	0	0	0	0	0	0
forget	0	0	0	0	0	0
gather	0	0	0	0	0	0
grant	0	0	0	0	0	0
grumble	0	0	0	0	0	0
guarantee	0	0	0	0	0	0
guess	0	0	0	1	0	1
happen (it)	0	0	0	0	1	1
hear	7	0	7	11	0	11
hint	0	0	0	0	0	0
hold	1	0	1	1	0	1
hope	6	1	7	4	2	6
ignore	0	0	0	0	0	0
imagine	1	1	2	0	0	0
imply	1	0	1	1	1	2
indicate	0	4	4	0	10	10
infer	0	0	0	0	0	0
inform	2	0	2	3	7	10
inquire	0	0	0	0	0	0
insist	0	0	0	0	0	0
intend	0	0	0	0	0	0

	EST "zero"	EST + *that*	EST all	ETT "zero"	ETT + *that*	ETT all
judge	0	0	0	0	0	0
KNOW[226]	30	7	37	10	0	10
learn	26	2	28	1	1	2
maintain	1	0	1	0	4	4
mean	9	7	16	5	3	8
mention	6	1	7	3	0	3
move	0	0	0	0	0	0
mumble	0	0	0	0	0	0
murmur	0	0	0	0	0	0

[225] find (EST: reporting verb occurs only in past tense); (ETT: most rv occur in present tense)
[226] Know (EST 19: 4; (ETT 10:0); Knew (EST 18: 3; (ETT 0:0)

muse	0	0	0	0	0	0
mutter	0	0	0	0	0	0
note	0	3	3	0	13	13
notice	1	0	1	0	0	0
notify	0	0	0	0	3	3
object	0	0	0	1	0	1
observe	1	0	1	0	0	0
order	3	0	3	0	2	2
overlook	0	0	0	0	0	0
perceive	0	0	0	0	0	0
permit	0	0	0	0	0	0
persuade	0	0	0	2	0	2
plead	2	0	2	0	0	0
pledge	0	0	0	0	0	0
ponder	0	0	0	0	0	0
pray	0	0	0	0	0	0
predict	3	1	4	0	0	0
prefer	0	0	0	0	0	0
pretend	0	0	0	0	0	0
	EST "zero"	EST + *that*	EST all	ETT "zero"	ETT + *that*	ETT all
proclaim	0	0	0	0	0	0
promise	0	0	0	0	0	0
prophesy	0	0	0	0	0	0
propose	0	1	1	0	2	2
prove	2	0	2	1	0	1
read	2	0	2	8	0	8
realize	1	1	2	0	0	0
reason	0	0	0	0	0	0
reassure	0	0	0	0	0	0
recall	1	0	1	0	0	0
recite	0	0	0	0	0	0
reckon	0	0	0	0	0	0
recommend	0	2	2	2	0	2
record	0	0	0	0	0	0

reflect	0	0	0	0	0	0
regret	0	0	0	0	0	0
remark	0	0	0	0	0	0
remember	4	1	5	2	0	2
remind	0	0	0	0	1	1
repeat	0	0	0	3	0	3
reply	0	0	0	1	1	2
report	11	0	11	2	1	3
request	2	0	2	10	7	17
require	48	0	48	35	3	38
resolve	6	0	6	0	0	0
respond	16	0	16	5	0	5
reveal	6	2	8	1	0	1
	EST "zero"	EST + *that*	EST all	ETT "zero"	ETT + *that*	ETT all
rule	0	0	0	0	0	0
SAY[227]	50	4	54	7	3	10
SEE	5	1	6	0	0	0
sense	0	1	1	0	1	1
scream	0	0	0	0	0	0
shout	0	0	0	0	0	0
show	2	2	4	0	2	2
shriek	0	0	0	0	0	0
specify	7	0	7	3	2	5
state	2	2	4	9	14	23
stipulate	0	0	0	0	0	0
storm	0	0	0	0	0	0
suggest	6	35	41	4	5	9
suppose	0	0	0	0	0	0
swear	0	0	0	0	0	0
teach	0	0	0	1	1	2
TELL [228]	6	0	6	3	0	3
THINK [229]	13	2	15	0	1	1

[227] say (EST 23: 1), (ETT 3:2); said (EST 31:3), (ETT 7:1)
[228] tell (EST 5: 0; ETT 3:0); told (EST 1: 0; ETT 0:0)
[229] think (EST 12:2; ETT 0:0); thought (EST 3: 0; ETT 1:1)

threaten	2	0	2	4	0	4
thunder	0	0	0	0	0	0
transpires (it)	0	0	0	0	0	0
understand	5	1	6	0	0	0
urge	0	0	0	3	0	3
vow	0	0	0	0	0	0
wail	0	0	0	0	0	0
warn	0	1	1	0	0	0
	EST "zero"	EST *+ that*	EST all	ETT "zero"	ETT *+ that*	ETT all
whisper	0	0	0	0	0	0
wish	1	0	1	17	0	17
worry	1	0	1	0	0	0
write	4	0	4	1	0	1
yell	0	0	0	0	0	0
	518	**183 (26.1% of total)**	**701**	**331**	**204 (38.1% of total) +12.0%**	**535**

Appendix V: *That* and *Which* (Total Instances vs. ODC)

	EST	ETT
That ODC	11 (2.0% of total)	21 (2.88% of total) **TT +.8%**
That (all instances)	532	727
Which ODC	1 (.98% of total)	17 (11.33% of total) **TT + 10.4%**
Which (all instances)	102	150

Appendix VI: Lix Scores, English and French

Please note that these scores have been calculated using an adaptation of the Lix formula as follows:

Adapted Lix = (ASL) + (Number of long words/Number of words)

1. Non-translated English Non-specialized Sub-corpus

	ASL	6+ letters (%)	# words	Lix
1. CBEST1	27	36	114	**58.57**

2.	CBEST2	24.88	363	874	**66.41**
3.	CBEST3	28.86	76	202	**66.48**
4.	CDEST2	22.48	319	628	**73.27**
5.	CPEST7	20.76	1118	1994	**76.82**
6.	CPEST8	15.94	313	698	**60.78**
7.	CPEST11	21.33	185	426	**64.75**
8.	CPEST12	21.37	208	643	**53.71**
9.	CPEST13	19.68	214	610	**54.76**
10.	CPEST15	25.83	526	1459	**61.88**
11.	CPEST16	25.33	247	683	**61.49**
12.	CPEST18	28.88	472	1174	**69.08**
13.	CPEST19	18.52	264	696	**56.45**
14.	CPEST20	27.29	243	609	**67.19**
15.	PWEST1	23.00	104	212	**72.05**
16.	PWEST3	24.17	250	429	**82.44**
17.	PWEST4	23.18	174	321	**76.45**
18.	PWEST5	29.78	356	645	**84.97**
19.	PWEST6	23.20	155	356	**66.73**
20.	PWEST8	11.38	60	146	**52.47**
21.	PWEST9	13.18	81	237	**47.35**
22.	PWEST10	21.14	67	162	**62.49**
23.	PWEST11	20.33	45	138	**52.93**
24.	SCEST1	22.96	753	2196	**57.24**
				15,652	**64.4 (mean)**

2. Translated English Non-specialized Sub-corpus

		ASL	6+ letters (%)	# words	**Lix**
1.	CCETT2	22.13	153	354	**65.35**
2.	CCETT3	25.23	518	1451	**60.92**
3.	CCETT5	22.25	269	723	**59.45**
4.	CCETT6	21.06	205	514	**60.94**
5.	CCETT7	22.64	247	643	**61.05**
6.	CCETT8	25.67	155	349	**70.08**
7.	CCETT9	30.00	97	262	**67.02**
8.	CCETT10	18.50	117	378	**49.45**
9.	CCETT11	27.67	87	210	**69.09**
10.	CCETT12	42.67	93	231	**82.92**
11.	CCETT13	27.40	106	243	**71.02**
12.	CCETT14	29.00	125	301	**70.52**
13.	CCETT15	22.45	164	300	**77.11**
14.	CCETT17	25.97	547	1389	**65.35**
15.	CCETT18	31.31	305	731	**73.03**
16.	CSETT3	23.62	129	355	**59.95**

17.	CTETT1	27.20	79	183	**70.36**
18.	DTETT1	22.95	383	872	**66.87**
19.	DTETT2	20.51	343	790	**63.92**
20.	DTETT3	22.35	255	581	**66.23**
21.	DTETT4	24.11	381	928	**65.16**
22.	DTETT5	25.71	458	976	**72.63**
23.	DTETT6	20.00	297	578	**71.38**
24.	DTETT7	24.67	140	287	**73.45**
25.	DTETT8	25.23	423	913	**71.56**
26.	DTETT11	19.62	128	263	**68.28**
27.	DTETT12	10.31	189	450	**52.31**
28.	DTETT13	18.21	124	253	**67.22**
29.	DTETT16	27.00	181	385	**74.01**
30.	DTETT17	26.71	279	634	**70.71**
				15,689	**67.2 (mean)**
					ETT + 2.8

3. Non-translated French Non-specialized Sub-corpus

		ASL	6+ letters	# words	Lix
1.	CCFST2	38.00	196	420	84.66
2.	CCFST7	30.48	342	910	68.06
3.	CCFST9	36.29	124	303	77.21
4.	CCFST11	36.00	111	276	76.21
5.	CCFST12	51.75	121	290	93.47
6.	CCFST13	23.50	116	276	65.52
7.	CCFST14	25.17	154	331	71.69
8.	CCFST15	26.08	173	358	74.40
9.	CCFST18	38.58	345	909	76.53
10.	CSFST3	27.08	157	388	67.54
11.	CTFST1	24.00	87	210	65.42
12.	DTFST1	29.28	380	1002	67.20
13.	DTFST2	20.32	313	751	61.99
14.	DTFST3	24.62	252	624	65.00
15.	DTFST4	26.31	421	984	69.09
16.	DTFST5	29.64	425	1020	71.30
17.	DTFST6	28.58	312	698	73.27
18.	DTFST7	24.75	125	300	66.16
19.	DTFST8	28.18	417	1005	69.67
20.	DTFST9	34.00	105	265	73.62
21.	DTFST10	24.46	391	986	64.11
22.	DTFST11	22.00	117	297	61.39
23.	DTFST12	13.80	191	475	54.01
24.	DTFST13	19.33	120	271	63.61
25.	DTFST14	14.95	347	758	60.72
26.	DTFST16	35.54	201	482	77.24
27.	DTFST17	34.19	300	712	76.32
28.	PWFST2	12.00	73	141	63.77
29.	PWFST3	33.71	131	291	78.72
				15,733	**70.3** (mean)

4. Translated French Non-specialized Sub-corpus

	ASL	6+ letters	#words	Lix
1. CBFTT1	37.00	49	133	73.84
2. CBFTT3	32.67	86	224	71.06
3. CCFTT1	21.90	624	784	101.49
4. CCFTT3	18.87	227	633	54.73
5. CCFTT5	24.90	92	262	60.01
6. CCFTT6	32.16	396	633	94.71
7. CCFTT7	37.73	236	261	128.15
8. CCFTT9	33.45	134	401	66.86
9. CCFTT10	35.26	332	762	78.82
10. CDFTT1	44.65	337	814	86.05
11. CDFTT2	28.75	349	767	74.25
12. CPFTT1	34.80	304	749	75.38
13. CPFTT3	32.60	116	249	79.18
14. CPFTT4	27.38	107	223	75.36
15. CPFTT5	24.33	129	312	65.67
16. CPFTT6	12.00	84	164	63.21
17. CPFTT8	24.96	398	943	67.16
18. CPFTT12	21.54	289	680	64.04
19. CPFTT14	21.71	351	855	62.76
20. CPFTT15	26.87	226	1574	41.22
21. CPFTT16	51.75	319	756	93.94
22. PWFTT1	24.80	120	268	69.57
23. PWFTT9	16.06	107	273	55.25
24. PWFTT11	19.00	63	159	58.62
25. SGFTT1	29.52	1008	2706	66.77
			15,670	**73.1** (mean) **TT +2.9**

Appendix VII: Henry-de Landsheere Scores, Calculation

From each sub-corpus, six samples of 250 words each (a total 1,500 words or 10% of the non-specialized sub-corpus) were taken.

A. Non-translated French sub-corpus

Texte 1

Du 10 au 14 juillet dernier, Jean-Louis Roux, **président** du Conseil des Arts du Canada, **accompagné** de Jeannita Thériault, **membre** du **conseil** d'**administration** du Conseil et **néo-brunswickoise**, ont **effectué** une **tournée** qui les a **menés** à Moncton, Dalhousie, Bouctouche et Caraquet.

Jean-Louis Roux, **président** du Conseil des Arts, et Bill Wells, **conseiller municipal** de Regina, **procédant** au lancement du **projet** « Moving Write Along ».
Déterminé à **informer** la communauté sur l'importance du **financement** et du **développement** régional des **arts**, Jean-Louis Roux a **accordé** des **entrevues** au Telegraph Journal, à la **station** CHOIX-FM, à l'Acadie Nouvelle, au Moncton Times Transcript et aux radios anglaise et française des différentes **chaînes** et **stations locales** de Radio-Canada. Le **président** et Mme Thériault ont rencontré, entre autres, des **membres** et des représentants des différentes communautés **artistiques**, du nouveau **comité** des **arts** de Moncton, du Centre culturel d'Aberdeen, de l'Association acadienne des **artistes professionnelles**, de **l'organisme** DansEncorps, du Conseil des **arts** du Nouveau-Brunswick et du Théâtre populaire d'Acadie.
Cette **tournée** a aussi été l'occasion, pour le **président** du Conseil, d'**assister** aux pièces Laurie ou la vie de galerie d'Herménégilde Chiasson et Les Troisses d'Antonine Maillet, de visiter le Musée Clément-Cormier et le **monument** Lefebvre, de **participer** à l'**ouverture** du 5e Festival de musique de chambre de Baie des Chaleurs et de **constater** le **dynamisme** des **artistes** et des **organismes artistiques** de la **province**.
Plus **tôt** cet été, **grâce** à l'initiative du Festival **international** de **poésie** de Trois-Rivières, le **projet** [...].

nombre de phrases:	6
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	41.6
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	51 x .4 = 20.4
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	1 x .4 = .04
Text 1 FST SCORE	Niveau 11-12 (secondaire supérieur) : 39

Texte 2 (Non-translated French)

Ce programme est **structuré autour** de trois **éléments** principaux : l'éducation de **base**, la **formation professionnelle** et technique et l'**appui** à l'**édition scolaire**. C'est l'éducation de base qui est la première des **priorités**. Les **activités** de ce **secteur** visent l'éducation **formelle** et **informelle** et comprennent des **projet** d'**alphabétisation**. L'**appui** à l'édition **scolaire** vise à

renforcer la **capacité éditoriale** des pays du Sud. La Francophonie est un **débouché** naturel pour l'**expertise** canadienne en **matière** d'éducation et de **formation professionnelle**. Notre **expertise** en **formation** à **distance** est très **appréciée**.

L'**enseigne**ment supérieur et la **recherche**, y compris les mesures **favorisant** la **coopération inter-universitaire** et la **mobilité** des étudiants et des chercheurs, font **également** partie des principaux **axes** d'**intervention** du programme. C'est dans le **cadre** de ce programme que les **activité**s de l'AUF sont **menées**.

Francophonie, **économie** et **développ**ement.

Ce programme s'intéresse à l'**économie**, aux **entreprises** et au **développement durable**. Les principaux **objectif**s de la **coopération multilatérale** dans ce **domaine** sont la **lutte** contre la pauvreté, le **rééquilibrage** du **commerce international**, **notamment** au moyen de mesures d'**incitation** à l'initiative **privée**, et la **mise** en place de mesures offrant de meilleures **garanties**, en particulier **juridiques**, pour la **croissance économique**. Dans ce **contexte**, le Canada a **annoncé** la **création** d'un **fonds** spécial de 700 000 dollars **destiné** à **appuyer** l'**intégration** et la pleine **participation** des pays les moins avancés au **système économique mondial**. Le Canada est aussi très **actif** au **niveau** des **projets** en **agriculture** et en **environnement** : **appui** aux PME **agroalimentaires**, **gestion** d'écos**ystèmes fluviaux**.

nombre de phrases:	13
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	19.2
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	86 x .4 = 34.4
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	0
Text 2 FST SCORE	37

Texte 3 (Non-translated French)

À **titre** de directeur général régional, M Normand Couture représente le **ministre** et la sous-**ministre**, et il est le principal **porte-parole** de Travaux publics et Services gouvernementaux Canada (TPSGC) pour la région du Québec. Il est aussi **chargé** de **régler** toutes les questions **relatives** aux services à la **clientèle** sur ce **territoire**.

Pendant 21 ans, M Couture a **oeuvré** à l'ancien **ministère** des Approvisionnements et des Services, d'abord, à Ottawa et Hull, puis à Québec et Montréal. De 1972 à 1993, il y a occupé **diverses fonctions**, **notamment** comme **agent négociateur** de **contrats**, chef des opérations régionales, **gestion**naire des services d'**imprimerie**, **gestion**naire des **acquisitions**, **conseil**ler principal pour les programmes et directeur de l'Ouest du Québec. Il était **affecté** à ce dernier poste lors de la **création** de TPSGC en juin 1993. Deux ans plus tard, il a été **nommé** directeur régional du Centre d'**expertise** - Services de **gestion** des **locaux** à bureaux et Services des biens **immobiliers**, **fonction** qu'il a **exercée** jusqu'au 18 décembre 1998. Il a ensuite agi à **titre** de directeur général régional par **intérim** de TPSGC - région du Québec jusqu'à sa **nomination officielle** le 17 mars 2000.

Au cours de sa **carrière** dans la **fonction** publique **fédérale**, M Couture a **planifié** et mis en **oeuvre** de nombreux **projets** axés, entre autres, sur la qualité de vie au travail, l'**intégration** des services d'**imprimer**ie, le **partenariat interministériel** pour la **gestion** d'un magasin libre-service, la **mise** sur pied d'un **secteur chargé** des opérations régionales, la **négociation** de plusieurs **dossiers**.

nombre de phrases:	8
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	31.2
nombre de mots différents absents du vocabulaire fondamental (les 1200 mots les plus fréquents en corpus de 15,000 mots) AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	52 x .4 = 20.8
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	0
Text 3 FST SCORE	43

Texte 4 (Non-translated French)

S'inspirant des Jeunesses musicales de France, il a mis sur pied, **outre** les Jeunesses musicales du Canada, l'Orchestre **mondial** et le Concours national. La **passion** de Gilles Lefebvre pour la culture a **certes** été **marquée** par une **volonté pédagogique**, mais aussi par un engagement **indéniable envers** les **organisme**s de **soutien** aux **arts**. En plus des trois **mandats** à la **présidence** de la Fédération internationale des JM, il a agi à **titre** de directeur du Centre culturel canadien à Paris, de directeur associé du Conseil des Arts du Canada, de secrétaire général intérimaire de la Commission canadienne pour l'UNESCO et de **président** du Conseil des arts de la Communauté urbaine de Montréal.

Jean-Louis Roux, ami de longue **date** de Gilles Lefebvre, **décrivait** en ces **termes** l'ami, le grand homme et son **œuvre** : « Gilles Lefebvre fut un homme d'**action** et d'engagement par excellence, et il laisse un **vigoureux héritage** : **grâce** à lui, des **générations** de **citoyens** et de **citoyennes** prendront un plaisir **accru** à la **fréquentation** des arts. »

Au cours de sa **carrière**, Gilles Lefebvre a reçu, à juste **titre**, plusieurs **mentions** d'**honneur**. Nombreux sont les jeunes musiciens pour qui l'**oeuvre** de Gilles Lefebvre a été un **tremplin** vers une **carrière professionnelle**; nombreux encore seront les jeunes musiciens pour qui l'**oeuvre** de Gilles Lefebvre sera un **tremplin** vers une **carrière professionnelle**.

L'État d'urgence, installation publique **montée** au centre-ville de Montréal par le groupe **multidisciplinaire** ATSA (Action terroriste socialement acceptable), en **collaboration** avec le Musée d'art contemporain de Montréal.

nombre de phrases:	7
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	35.71
nombre de mots différents absents du vocabulaire fondamental (les 1200 mots les plus fréquents en corpus de 15,000 mots) AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	45 x .4 = 18
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de	1 x .4 = .04

dialogue) x (100 ÷ 250 mots) = 7,7	
Text 4 FST SCORE	42

Texte 5 (Non-translated French)

Le Canada a toujours joué un rôle très **actif** en Francophonie. Dès les débuts de la **création** d'une Francophonie **structurée**, le Canada s'est employé à **construire** les bases d'une **coopération** vraiment **multilatérale**.

Cet **intérêt** du Canada répondait à des **objectifs** de politique étrangère autant que de politique **intérieure**. D'une part, le Canada avait et a toujours comme **pilier** de sa politique étrangère le **multilatéra**lisme. Le texte explique pourquoi la Francophonie en est un bel exemple et quels en sont les **bénéfices**. En **termes** de politique **intérieure**, une **participation active** du Canada à la Francophonie permettait non seulement la **promotion** de la **dualité linguistique** au pays, mais elle permettait aussi au Canada de **négocier** un rôle **approprié** pour les **provinces désirant** s'**affirmer** sur la scène **internationale**, **notamment** le Québec et le Nouveau-Brunswick.

Les **structures** de la Francophonie sont **multiples** et **complexes**. La Francophonie, d'abord un **véhicule** de **coopération**, est devenue plus politique par la suite. Le Canada a joué un rôle important tant au **niveau** de la **coopération** que sur le plan politique. La **présence** de deux Canadiens à la tête de l'Agence de **coopération** culturelle et technique en **témoigne**. Encore aujourd'hui, le Canada est très **influent** au **sein** de la Francophonie, autant pour les **budgets** qu'il y **consacre** que par ses **interventions** et ses **projets** de **coopération** au **niveau** des nouvelles **priorités**.

La Francophonie fait tout de même face à des **défis**, la langue française perdant du **terrain** à l'**échelle mondiale**. Certains mécanismes, comme la **création** d'un poste de Secrétaire....

nombre de phrases:	13
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	19.23
nombre de mots différents absents du vocabulaire fondamental (les 1200 mots les plus fréquents en corpus de 15,000 mots) AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	50 x .4 = 20
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	0
Text 5 FST SCORE	44

Texte 6 (Non-translated French)

La Francophonie ne **date** pas d'hier! En fait, le **terme** « **francophonie** » a été inventé en 1880 par un **géographe** français, pour **définir** l'ensemble des personnes et des pays **utilisant** le français à des **titres divers**.

Comme c'était le cas avec les grandes **puissances** d'autrefois, le **passé colonial** de la France a servi de base pour tisser des **liens** surtout **économiques**, mais aussi sociaux et culturels entre la France et ses nombreuses colonies au cours des derniers **siècles**. La langue française est devenu au fil des ans un **ciment** qui **unifiait** non seulement la France et ses colonies mais aussi

plusieurs colonies entre-elles. Il est vrai qu'au début de la **colonisation**, le français était surtout parlé par les **colons** Français et les **élites locale**s mais son **utilisation** s'est **démocratisée** avec le temps.

Le 20ième **siècle** a été **marqué** par la **création** de nombreux pays au **fur** et à mesure que ces colonies **proclamaient** leur **indépendance**. La langue française continuait **cependant** d'agir comme **trait-d'union** entre ces nouveaux pays et **facilitait** les **échanges commerciaux**, sociaux et culturels.

Les anciennes **puissances colon**iales **francophon**es, la France et la Belgique, ont alors mis sur pied des programmes d'aide **bilatérale** afin d'aider ces nouveaux pays **francophones** à se **structurer** tant sur le plan politique, qu'**économique** ou social. D'autres pays **nantis** et parlant français, comme le Canada et la Suisse, les ont imités et ont mis sur pied d'importants programmes d'aide **bilatéra**le au cours de la seconde moitié du 20ième **siècle**.
Cependant, les pays moins **développés**, ...

nombre de phrases:	9
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	27.77
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	43 x .4 = 17.2
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	2 x .4 = .8
Text 6 FST SCORE	43

B. Translated French sub-corpus

Texte 1

Les Canadiens ont le droit de savoir pourquoi, comment et où nous **investissons** leurs **cotisations** au Régime de pensions du Canada, qui prend les décisions de placement, quels placements sont **détenus** en leur nom et quel est leur **rendement**.

Qui nous sommes et notre raison d'être. Notre **organisme** est une **société** de placement **gérée**, **indépendamment** du Régime de pensions du Canada, par des **spécialistes** du placement **expérimentés provenant** du **secteur privé**. Notre rôle **consiste** à **investir** les **cotisations** au RPC dont celui-ci n'a pas besoin pour payer les **prestations courantes** dans le **but** d'**obtenir** un **rendement maximal** tout en **assumant** un **minimum** de **risques**. Les **fonds** que nous **investissons** aujourd'hui permettront au Régime de pensions du Canada de payer les **pensions** des travailleurs canadiens qui commenceront à prendre leur **retraite** dans 20 ans.

L'Office d'investissement du RPC a été **conçu** par suite d'un **examen** public du Régime de pensions du Canada (RPC) en 1996. Cet **examen** a été demandé par les **ministres** des Finances **fédéral** et provinciaux, **soucieux** de **préserver** le Régime et de faire en sorte qu'il **demeure abordable** pour les Canadiens de la **génération actuelle** comme des **générations futures**.

En 1996, le Régime de pensions du Canada ne **disposait** pas d'un **actif** suffisant pour **honorer** ses **obligations** à long **terme**. À **l'issue** de **l'examen public**, les **ministres** des Finances ont **convenu**, en 1997, d'**augmenter** les **taux** de **cotisation** de **manière** à **générer** des **fonds** en **excédent** de ceux qui sont nécessaires pour payer les **pensions**, au moins jusqu'en 2021.

nombre de phrases:	9
nombre de mots	250

nombre de mots par phrase MP = Mots (250) ÷ Phrases	27.77
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	60 x .4 = 24
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	3 3 x .04 = .12
Text 1 FTT SCORE	**36**

Texte 2 (Translated French)

Pour des enfants qui ont l'habitude d'entendre des coups de **feu** dans la **baie** de Miramichi ou qui **ressentent** de la peur lorsqu'ils voient des **hélicoptères voler** dans le **ciel**, travailler côte à côte avec les militaires et parler avec un **agent** de la GRC de ce qui l'a amené à devenir **agent** peuvent aider à **dissiper** ces **craintes** dans une large mesure.

Jouer dans l'eau avec Barbara Hall, **présidente** de la Stratégie nationale sur la sécurité communautaire et la **prévention** du **crime**, aide à comprendre qu'une dame **adulte** de la grande ville est simplement une autre personne qui aime s'amuser tout en essayant de se **rafraîchir** un jour où il fait chaud.

Jeannie nous l'explique ainsi : "La Roue de la médecine nous parle des quatre **couleurs** de l'homme et comment nous pouvons apprendre les uns des autres. Nous avons chacun un **don** qui nous a été **attribué** par le Créateur. Une fois que nous apprenons que nous avons quelque chose à offrir et que les autres peuvent nous **enseigner** des choses, nous devenons forts. Lorsque nous commençons à **respecter** et à **honorer** les autres, nous devenons une nation plus **puissante**."

Le **rêve** qui est né **grâce** aux camps culturels pour jeunes de Burnt Church continuera d'exister dans le coeur de chacune des personnes qui a été touchée par l'expérience, qu'elle **soit** jeune ou âgée, **autochtone** ou non.
Melony McCarthy, dont les bureaux sont **situés** au Nouveau-Brunswick, est **agente** régionale de **communications** du Centre national de prévention du crime.

nombre de phrases:	8
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	31.2
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	29 x .4 = 11.6
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	12 12 x .4 = 4.8
Text 2 FTT SCORE	**47**

Texte 3 (Translated French)

De quoi s'agit-il? Les Services aux régions du Nord ont créé un **réseau** de **livraison** pour l'ensemble du Nord qui offre des services **postaux reliant** les trois **océans** et qui de plus en plus se **conforme** à la **règle** de la plus **courte distance** qu'**appliquent** d'autres **réseaux**. Les plus âgés se rappelleront sans doute des sacs **postaux déposés** juste à côté du Twin Otter et qui étaient **transportés** dans le **qamutiik** tiré par un **tracteur** de la **coopérative**, pour **ensuite** être **empilés** devant la **cabane** qui servait de bureau de poste **local**. Pendant des années, ces sacs étaient bleus. Et pour les habitants du Nord, cette **couleur** était devenue **symbole** de **promesse**. Ces sacs leur apportaient les articles nécessaires à leur **survie**, c'est-à-dire les **médicaments**, la **nourriture** et les **chèques**, les **vêtements** et le nécessaire de chasse ainsi que les pièces de **rechange**.

La **rapidité**, la **fréquence** et la **fiabilité** des services offerts par Postes Canada aux régions du Nord et sa **capacité** de **desservir** chaque personne même dans les plus petites **collectivités** sont autant de questions qui ont depuis longtemps **revêtu** une importance première pour les habitants du Nord.

Voici un exemple de la **contribution** que nous avons apportée au cours des dix dernières années, depuis que le service **postal** est un **enjeu essentiel** pour Postes Canada.

Desservir le plus **vaste territoire** au Canada. Une région **géographique** qui représente environ 72 % de la terre **émergée** du Canada. **Territoire** pour lequel le **courrier** est levé et **livré** principalement par **voie aérienne**.

nombre de phrases:	11
nombre de mots	253
nombre de mots par phrase MP = Mots (250) ÷ Phrases	22.72
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	51 x .4 = 20.4
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	2 2 x .4 = .8
Text 3 FTT SCORE	**41**

Texte 4 (Translated French)

Le Centre canadien d'architecture (CCA), de **concert** avec l'Institut royal d'architecture du Canada et le Conseil canadien des écoles universitaires d'architecture, a choisi Melvin Charney pour représenter le Canada. Le Conseil des Arts du Canada et le **ministère** des Affaires étrangères et du Commerce international avaient **chargé** le CCA d'organiser le **processus** de **sélection**. La **participation** du Conseil à la Biennale de Venise **témoigne** de la **volonté** du Conseil de **renforcer** ses **relations** avec le milieu canadien de l'**architecture**. En **collaboration** avec un **comité consultatif composé** de **membres** du **domaine** de l'**architecture**, le Service des arts visuels du Conseil a commencé la **révision** des **programmes** et des prix qu'il **réserve** à l'**architecture**. L'**architecte** canadien Melvin Charney **représentera** le Canada lors de la Septième. **Diffusion** du message de la Commission du droit de **prêt** public à l'étranger. Pour donner suite à la Conférence internationale sur le droit de **prêt** public dont elle a été l'**hôte** l'**automne** dernier, la Commission du droit de **prêt** public organise un **séminaire** sur le droit de **prêt** public lors de la **réunion** de la Fédération internationale des associations de bibliothécaires (FIAB), qui a lieu à Jérusalem en août 2000. Paul Whitney, directeur de la bibliothèque publique de Burnaby et nouveau représentant de l'Association canadienne des bibliothèques au **sein** de la Commission, représentera le Canada **lors** de ce

séminaire. Cet **événement** a pour **objectif** de **fortifier** les **relations** entre les **écrivains** et les **bibliothèques**, qui **constituent** un **élément indispensable** à la **survie** du droit de **prêt** public.

nombre de phrases:	9
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	27.7
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	147 x .4 = 18.8
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	0
Text 4 FTT SCORE	**47**

Texte 5 (Translated French)

Un **trésor** à **découvrir** afin de permettre aux visiteurs de s'**informer** des travaux de **rénovation** en cours, afin de pouvoir faire une visite guidée en trois **dimensions** des **édifices** du Parlement. **Grâce** à l'**annuaire électronique** des programmes et des services, le public pourrait **accéder** en ligne aux **coordonnées** exactes et à jour sur les personnes-**ressources responsables** des différents programmes et services offerts par le gouvernement du Canada. Pour l'**instant**, la **population** canadienne peut trouver ces **renseignements** uniquement dans les **pages** bleues des **annuaires téléphoniques** publics.
Vous avez peut-être vu notre **logo** lorsque vous vous êtes arrêté pour **examiner** un nouvel **immeuble** en construction ou vous l'avez peut-être aperçu sur une **annonce** de **biens excédentaires** du **gouvernement publiée** dans les journaux. Vous l'avez aussi peut-être aperçu lors de votre visite sur la Colline du Parlement.
Qui sommes-nous ? Nous sommes Travaux publics et Services gouvernementaux Canada et nous sommes là pour vous. Le **Ministère** assure la **prestation** de services communs au gouvernement du Canada et emploie près de 12 000 personnes. Nous offrons des services et des programmes par le **biais** de bureaux **situés** dans tout le pays, sans oublier aux États-Unis et en Europe.
TPSGC joue de nombreux rôles. Dans cette partie de notre **site**, vous en **découvrirez davantage** sur notre travail **visant** à **restaurer** les **édifices** du Parlement. Vous pouvez **obtenir** de l'**information** sur ce que nous avons à vendre dans le **site** Biens de la Couronne, que ce soient des **chaises** des bureaux, des **outils** électriques ou des **véhicules**.

nombre de phrases:	11
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	22.7
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	43 x .4 = 17.2
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) =	14 14 x .4 = 5.6

7,7	
Text 5 FTT SCORE	**43**

Texte 6 (Translated French)

S'il est vrai que notre **succès** a fait notre **réputation**, c'est sans **conteste** à notre **maîtrise** de **l'innovation** que nous le devons. PCIL offre une **panoplie exceptionnelle** de services de **consultants**, de **solutions postales axées** sur la **technologie**, **notamment** en **matière** de **transformation postale**, ainsi que des **possibilités** de **commerce électronique** basées sur ces mêmes **systèmes** et ces mêmes services qui ont **propulsé** la Société canadienne des postes à l'avant-garde dans son **domaine**. **Grâce** à notre **savoir-faire** et à la **collaboration incessante** que nous ne manquerons pas d'entretenir avec votre **personnel** et vos autres **partenaires locaux**, nous nous faisons fort **d'accroître l'efficacité** de votre **exploitation postale au-delà** de ce que vous auriez cru possible. Un **portefeuille** à la mesure de vos besoins particuliers! Aujourd'hui, la **vaste panoplie** des **prestations** de PCIL **renferme** des **solutions technologiques**, des **méthodes** de **gestion**, des **formules** de produits et de services et un **savoir-faire postal adaptables** à toute situation et **entièrement aménageables**. **Profitez** de **l'appui** des meilleurs dans leur **domaine**. Son **savoir-faire**, PCIL le doit au **succès** de la **société**-mère, la Société canadienne des postes, dans le **réaménagement** et l'**amélioration** de ses **propres systèmes** et services. **Projets** Complété. Voici quelques-uns des **projets parmi** la **liste impressionnante** de ceux **couronnés** de **succès** auxquels PCIL a déjà eu l'**honneur** d'apporter son concours. **Expérience** reconnue et **innovation** d'**avant-garde**. Au cours des dix dernières années, Postes Canada internationale ltée (PCIL) a **acquis** une **vaste expérience**, et sa **réputation** d'excellence est aujourd'hui sans **rivale** dans l'**industrie** des services **postaux**.

nombre de phrases:	10
nombre de mots	250
nombre de mots par phrase MP = Mots (250) ÷ Phrases	25
nombre de mots différents absents du vocabulaire fondamental AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)	77 x .4 = 30.8
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	5 5 x .4 = 2
Text 6 FTT SCORE	**31**

Instructions, Formule de Henry

1. Compter le nombre de phrases.
2. Compter le nombre de mots (250 par échantillon)
3. Calculer le nombre de mots par phrase
   Exemple: MP = Mots (250) ÷ Phrases
4. Compter le nombre de mots différents absents du vocabulaire fondamental de Gougenheim.
   Échantillon de longueur constante (100 mots).
   AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250)
5. Compter le nombre d'indicateurs de dialogue (DEXGU)
   # de ! + # de « + # de prènoms employés seuls de premier et de deuxième personne (je, j', tu t', nous, vous)
6. Ramener à une longueur standard de 100 mots:
   DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7

Vandendooren (1999, Ch. 4, page 2):

Selon les recommandations de Henry,
Les différentes fiches sont regroupées en un seul article.
Les tests se font avec des blocs de 250 mots environ, parce que dans des blocs de moins de 250 mots, une phrase longue pourrait fausser tout le résultat.
Autres modifications :
   Les noms propres et les acronyms ne sont pas comptés parmi les « absents » de la liste du vocabulaire fondamental de Gougenheim.
   Les mots suivants ne sont pas comptés : canadien(ne), fédéral(e)

Appendix VIII : Henry-de Landsheere Scores, "Automated" Calculation

A. Non-translated French sub-corpus

15,545 words

nombre de phrases:	595
nombre de mots	15545
nombre de mots par phrase MP = Mots (250) ÷ Phrases	MP = 26.1
nombre de mots différents absents du vocabulaire fondamental (les 1200 mots les plus fréquents en corpus de 15,000 mots) AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250) OU AG = (# mots de fréquence supérieur à .01%)	WL(F) : 3231 types – 884 (mots de fréquence supérieur à .01% dans le corpus) = 2347 2,347 x .006 (c'est-à-dire, 100 ÷ 15545) = AG = 14,1
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	24 (! = 5; « = 18; je = 1; les autres = 0) 24 x .006 DEXGU = .14
SCORE	Niveau 11-12 (secondaire supérieur) : 46

B. Translated French Sub-corpus

15,433 words

nombre de phrases:	556
nombre de mots	15,433
nombre de mots par phrase MP = Mots (250) ÷ Phrases	MP = 27.7
nombre de mots différents absents du vocabulaire fondamental (les 1200 mots les plus fréquents en corpus de 15,000 mots) AG = (# AG en 250 mots) x (100 ÷ mots échantillon de 250) OU AG = (# mots de fréquence inférieur à .01%, ramenés à un échantillon de la	AG = types – mots fréquents Types = comptés par WL Mots fréquents = (mots de fréquence supérieure à .01% du corpus) 3686 types – 987 = 2699 AG = 2699 x (100 ÷ 15433) = 16.1 100 ÷ 15433 = .006

longueur constante de 100 mots)	
nombre d'indicateurs de dialogue (!, «, je, j', tu, t', nous, vous) DEXGU = (indicateurs de dialogue) x (100 ÷ 250 mots) = 7,7	24 (! = 12; « = 18; je = 6; tu = 1; t' = 0; nous = 66; nous = 29) 132 total Ramener à une longueur constante de 100 mots : 132 x (100 ÷ 15433) = .85 DEXGU = .85
SCORE	Niveau 11-12 (secondaire supérieur) : **45**

Appendix IX: French ICA Samples and Variables

The following are the texts used in the calculation of the ICA Readability Index. All texts are excerpts of exactly 100 words. Excerpts were taken as systematically as possible from the beginning, middle and end of the compiled non-specialized sub-corpora.

A. Non-translated French non-specialized sub-corpus

Excerpt # 1
CCFST 2
ASL: 58
3 SYLL: 21
4+ SYLL: 4
ICA score: (58 + 4) x 0.4 = 24.8
Le président et Mme Thériault ont rencontré, entre autres, des embres et des représentants des différentes communautés artistiques, du nouveau comité des arts de Moncton, du Centre culturel d'Aberdeen, de l'Association acadienne des artistes professionnel.le.s, de l'organisme DansEncorps, du Conseil des arts du Nouveau-Brunswick et du Théâtre populaire d'Acadie. Cette tournée a aussi été l'occasion, pour le président du Conseil, d'assister aux pièces Laurie ou la vie de galerie d'Herménégilde Chiasson et Les Troisses d'Antonine Maillet, de visiter le Musée Clément-Cormier et le monument Lefebvre, de participer à l'ouverture du 5e Festival de musique de chambre de Baie des Chaleurs et de constater le dynamisme des artistes et des organismes artistiques de la province.

Excerpt #2
DTFST5
ASL: 14.28
3 SYLL: 17
4+ SYLL: 17
ICA score: (14.28 + 17) x 0.4 = 12.51
Ce programme est structuré autour de trois éléments principaux : l'éducation de base, la formation professionnelle et technique et l'appui à l'édition scolaire. C'est l'éducation de base qui est la première des priorités. Les activités de ce secteur visent l'éducation formelle et informelle et comprennent des projet d'alphabétisation. L'appui à l'édition scolaire vise à renforcer la capacité éditoriale des pays du Sud. La Francophonie est un débouché naturel pour l'expertise canadienne en matière d'éducation et de formation professionnelle. Notre expertise en formation à distance est très appréciée, ... L'enseignement supérieur et la recherche, y compris les mesures favorisant la coopération ...

Excerpt # 3
DTFST5
ASL: 26.75
3 SYLL: 14

4+ SYLL: 15
ICA score: (26.75 + 15) x 0.4 = 16.7
C'est dans le cadre de ce programme que les activités de l'AUF sont menées. Ce programme s'intéresse à l'économie, aux entreprises et au développement durable. Les principaux objectifs de la coopération multilatérale dans ce domaine sont la lutte contre la pauvreté, le rééquilibrage du commerce international, notamment au moyen de mesures d'incitation à l'initiative privée, et la mise en place de mesures offrant de meilleures garanties, en particulier juridiques, pour la croissance économique. Dans ce contexte, le Canada a annoncé la création d'un fonds spécial de 700 000 dollars destiné à appuyer l'intégration et la pleine participation des pays les moins avancés au système économique mondial.

Excerpt # 4
PWFST3
ASL: 28.75
3 SYLL: 20
4+ SYLL: 10
ICA score: (28.75 + 10) x 0.4 = 15.5
À titre de directeur général régional, M. Normand Couture représente le ministre et la sous-ministre, et il est le principal porte-parole de Travaux publics et Services gouvernementaux Canada (TPSGC) pour la région du Québec. Il est aussi chargé de régler toutes les questions relatives aux services à la clientèle sur ce territoire. Pendant 21 ans, M. Couture a oeuvré à l'ancien ministère des Approvisionnements et des Services, d'abord, à Ottawa et Hull, puis à Québec et Montréal. De 1972 à 1993, il y a occupé diverses fonctions, notamment comme agent négociateur de contrats, chef des opérations régionales, gestionnaire des services d'imprimerie, gestionnaire des acquisitions, conseiller principal pour les programmes et directeur de l'Ouest du Québec.

Excerpt # 5
DTFST5 & PWFST3
ASL: 25
3 SYLL:14
4+ SYLL: 11
ICA score: (25 + 11) x 0.4 = 14.4
Le Canada est aussi très actif au niveau des projets en agriculture et en environnement : appui aux PME agroalimentaires, gestion d'écosystèmes fluviaux, appui à l'Institut de l'Énergie et de l'environnement de Québec. Il était affecté à ce dernier poste lors de la création de TPSGC en juin 1993. Deux ans plus tard, il a été nommé directeur régional du Centre d'expertise - Services de gestion des locaux à bureaux et Services des biens immobiliers, fonction qu'il a exercée jusqu'au 18 décembre 1998. Il a ensuite agi à titre de directeur général régional par intérim de TPSGC - région du Québec jusqu'à sa nomination officielle le 17 mars 2000.

Excerpt # 6
CCFST13
ASL: 33.33
3 SYLL:20
4+ SYLL: 7
ICA score: (33.3 + 7) x 0.4 = 16.12
S'inspirant des JMF, il a mis sur pied, outre les JMC, l'Orchestre mondial et le Concours national. La passion de Gilles Lefebvre pour la culture a certes été marquée par une volonté pédagogique, mais aussi par un engagement indéniable envers les organismes de soutien aux arts. En plus des trois mandats à la présidence de la Fédération internationale des JM, il a agi à titre de directeur du CCC à Paris, de directeur associé du Conseil des Arts du Canada, de secrétaire général intérimaire de la Commission canadienne pour l'UNESCO et de président du Conseil des arts de la CUM.

Excerpt # 7
DTFST11
ASL: 16.66
3 SYLL: 20

4+ SYLL: 10
ICA score: (16.66 + 10) x 0.4 = 10.66
Le Canada a toujours joué un rôle très actif en Francophonie. Dès les débuts de la création d'une Francophonie structurée, le Canada s'est employé à construire les bases d'une coopération vraiment multilatérale. Cet intérêt du Canada répondait à des objectifs de politique étrangère autant que de politique intérieure. D'une part, le Canada avait et a toujours comme pilier de sa politique étrangère le multilatéralisme. Le texte explique pourquoi la Francophonie en est un bel exemple et quels en sont les bénéfices. En termes de politique intérieure, une participation active du Canada à la Francophonie permettait non...

Excerpt # 8
PWFST 3 & CCFST13
ASL: 50
3 SYLL:9
4+ SYLL: 8
ICA score: (50 + 8) x 0.4 = 23.2
Au cours de sa carrière dans la fonction publique fédérale, M. Couture a planifié et mis en oeuvre de nombreux projets axés, entre autres, sur la qualité de vie au travail, l'intégration des services d'imprimerie, le partenariat interministériel pour la gestion d'un magasin libre-service, la mise sur pied d'un secteur chargé des opérations régionales, la négociation de plusieurs dossiers ...
Jean-Louis Roux, ami de longue date de Gilles Lefebvre, décrivait en ces termes l'ami, le grand homme et son œuvre : « Gilles Lefebvre fut un homme d'action et d'engagement par excellence, et il laisse un vigoureux héritage : grâce à lui, des générations ...

Excerpt # 9
CCFST13 & CCFST14 & DTFST11
ASL: 20
3 SYLL: 12
4+ SYLL: 7
ICA score: (20 + 7) x 0.4 = 10.8
...de citoyens et de citoyennes prendront un plaisir accru à la fréquentation des arts. »
Au cours de sa carrière, Gilles Lefebvre a reçu, à juste titre, plusieurs mentions d'honneur. Nombreux sont les jeunes musiciens pour qui l'oeuvre de Gilles Lefebvre a été un tremplin vers une carrière professionnelle; ...
L'État d'urgence, installation publique montée au centre-ville de Montréal par le groupe multidisciplinaire ATSA (Action terroriste socialement acceptable), en collaboration avec le Musée d'art contemporain de Montréal...
...seulement la promotion de la dualité...

Excerpt # 10
DTFST11
ASL: 16.83
3 SYLL: 16
4+ SYLL: 8
ICA score: (L + M) x 0.4 = 9.93
...linguistique au pays, mais elle permettait aussi au Canada de négocier un rôle approprié pour les provinces désirant s'affirmer sur la scène internationale, notamment le Québec et le Nouveau-Brunswick.

Les structures de la Francophonie sont multiples et complexes. La Francophonie, d'abord un véhicule de coopération, est devenue plus politique par la suite. Le Canada a joué un rôle important tant au niveau de la coopération que sur le plan politique. La présence de deux Canadiens à la tête de l'Agence de coopération culturelle et technique en témoigne. Encore aujourd'hui, le Canada est très influent au sein de...

Excerpt #11
DTFST11 & DTFST1
ASL: 20

3 SYLL: 9
4+ SYLL: 7
ICA score: (L + M) x 0.4 = 10.8
...la Francophonie, autant pour les budgets qu'il y consacre que par ses interventions et ses projets de coopération au niveau des nouvelles priorités.
La Francophonie fait tout de même face à des défis importants, la langue française perdant du terrain à l'échelle mondiale. Certains mécanismes, comme la création d'un poste de Secrétaire ...
La Francophonie ne date pas d'hier! En fait, le terme « francophonie » a été inventé en 1880 par un géographe français, Onésisme Reclus (1837-1916), pour définir l'ensemble des personnes et des pays utilisant le français à des titres divers.
Comme c'était le cas avec les grandes puissances d'autrefois, le passé ...

Excerpt #12
DTFST1 (extrait)
ASL: 20
3 SYLL: 6
4+ SYLL: 6
ICA score: (20 + 6) x 0.4 = 10.4
... colonial de la France a servi de base pour tisser des liens surtout économiques, mais aussi sociaux et culturels entre la France et ses nombreuses colonies au cours des derniers siècles. La langue française est devenu au fil des ans un ciment qui unifiait non seulement la France et ses colonies mais aussi plusieurs colonies entre-elles. Il est vrai qu'au début de la colonisation, le français était surtout parlé par les colons Français et les élites locales mais son utilisation s'est démocratisée avec le temps.
Le 20ième siècle a été marqué par la création de nombreux pays au fur et ...

Excerpt #13
DTFST1
ASL: 20.4
3 SYLL: 15
4+ SYLL: 8
ICA score: (20.4 + 8) x 0.4 = 11.36
...à mesure que ces colonies proclamaient leur indépendance. La langue française continuait cependant d'agir comme trait-d'union entre ces nouveaux pays et facilitait les échanges commerciaux, sociaux et culturels.
Les anciennes puissances coloniales francophones, la France et la Belgique, ont alors mis sur pied des programmes d'aide bilatérale afin d'aider ces nouveaux pays francophones à se structurer tant sur le plan politique, qu'économique ou social. D'autres pays nantis et parlant français, comme le Canada et la Suisse, les ont imités et ont mis sur pied d'importants programmes d'aide bilatérale au cours de la seconde moitié du 20ième siècle.
Cependant, les pays moins développés, ...

Excerpt #14
CCFST9 (extrait)
ASL: 34
3 SYLL: 14
4+ SYLL: 8
ICA score: (34 + 8) x 0.4 = 16.8
Cet événement, qui propose des minispectacles de jazz, une aire d'exposition, des ateliers de jazz, des tables rondes et des séances de réseautage, réunit d'importants diffuseurs canadiens et internationaux, et représente pour les jeunes artistes une occasion unique de faire entendre leurs compositions et leur répertoire. Gratuits, les minispectacles ouvrent aussi toutes grandes les portes de la découverte de la relève du jazz. Bien que les lauréats aient moins de 30 ans, ils possèdent tous une impressionnante et longue expérience du monde du jazz : études auprès des grands noms du jazz, participations à des concerts et à des festivals au Canada et partout dans le monde, et enregistrements déjà acclamés par la critique.

Excerpt #15
CCFST2 (extrait)
ASL: 37.33
3 SYLL:16
4+ SYLL: 9
ICA score: (37.33 + 9) x 0.4 = 18.53
Du 10 au 14 juillet dernier, Jean-Louis Roux, président du Conseil des Arts du Canada, accompagné de Jeannita Thériault, membre du conseil d'administration du Conseil et néo-brunswickoise, ont effectué une tournée qui les a menés à Moncton, Dalhousie, Bouctouche et Caraquet.
Jean-Louis Roux, président du Conseil des Arts, et Bill Wells, conseiller municipal de Regina, procédant au lancement du projet « Moving Write Along ». Déterminé à informer la communauté sur l'importance du financement et du développement régional des arts, Jean-Louis Roux a accordé des entrevues au Telegraph Journal, à la station CHOIX-FM, à l'Acadie Nouvelle, au Moncton Times Transcript et aux radios anglaise et française des différentes chaînes et stations locales de Radio-Canada.

B. Translated French non-specialized sub-corpus

Excerpt #1.
CBFTT1
ASL: 25
3 SYLL:14
4+ SYLL: 8
ICA score: (25 + 8) x 0.4 = 13.2
Les Canadiens ont le droit de savoir pourquoi, comment et où nous investissons leurs cotisations au Régime de pensions du Canada, qui prend les décisions de placement, quels placements sont détenus en leur nom et quel est leur rendement.
Qui nous sommes et notre raison d'être. Notre organisme est une société de placement gérée, indépendamment du Régime de pensions du Canada, par des spécialistes du placement expérimentés provenant du secteur privé.
Notre rôle consiste à investir les cotisations au RPC dont celui-ci n'a pas besoin pour payer les prestations courantes dans le but d'obtenir un rendement maximal tout en assumant ...

Excerpt #2
CBFTT1 & CBFTT3
ASL: 20
3 SYLL: 21
4+ SYLL: 5
ICA score: (20 + 5) x 0.4 = 10
...un minimum de risques. Les fonds que nous investissons aujourd'hui permettront au Régime de pensions du Canada de payer les pensions des travailleurs canadiens qui commenceront à prendre leur retraite dans 20 ans. L'Office d'investissement du RPC a été conçu par suite d'un examen public du Régime de pensions du Canada (RPC) en 1996. Cet examen a été demandé par les ministres des Finances fédéral et provinciaux, soucieux de préserver le Régime et de faire en sorte qu'il demeure abordable pour les Canadiens de la génération actuelle comme des générations futures. En 1996, le Régime de pensions du Canada ne ...

Excerpt #3.
CBFTT3 & SGFTT1
ASL: 33.66
3 SYLL: 17
4+ SYLL: 5
ICA score: (33.66 + 5) x 0.4 = 38.66
...disposait pas d'un actif suffisant pour honorer ses obligations à long terme. À l'issue de l'examen public, les ministres des Finances ont convenu, en 1997, d'augmenter les taux de cotisation de manière à générer des fonds en excédent de ceux qui sont nécessaires pour payer les pensions, au moins jusqu'en 2021. Pour des enfants qui ont l'habitude d'entendre des coups

de feu dans la baie de Miramichi ou qui ressentent de la peur lorsqu'ils voient des hélicoptères voler dans le ciel, travailler côte à côte avec les militaires et parler avec un agent de la GRC de ce qui l'a amené ...

Excerpt #4
SGFTT1
ASL: 25.25
3 SYLL:16
4+ SYLL: 1
ICA score: (25.25 + 1) x 0.4 = 10.5

...à devenir agent peuvent aider à dissiper ces craintes dans une large mesure. Jouer dans l'eau avec Barbara Hall, présidente de la Stratégie nationale sur la sécurité communautaire et la prévention du crime, aide à comprendre qu'une dame adulte de la grande ville est simplement une autre personne qui aime s'amuser tout en essayant de se rafraîchir un jour où il fait chaud. Jeannie nous l'explique ainsi : "La Roue de la médecine nous parle des quatre couleurs de l'homme et comment nous pouvons apprendre les uns des autres. Nous avons chacun un don qui nous a été attribué par le Créateur.

Excerpt #5
SGFTT1
ASL: 32.66
3 SYLL:16
4+ SYLL: 3
ICA score: (L + M) x 0.4 = 14.26

Une fois que nous apprenons que nous avons quelque chose à offrir et que les autres peuvent nous enseigner des choses, nous devenons forts.
Lorsque nous commençons à respecter et à honorer les autres, nous devenons une nation plus puissante." Le rêve qui est né grâce aux camps culturels pour jeunes de Burnt Church continuera d'exister dans le coeur de chacune des personnes qui a été touchée par l'expérience, qu'elle soit jeune ou âgée, autochtone ou non autochtone.
Melony McCarthy, dont les bureaux sont situés au Nouveau-Brunswick, est agente régionale de communications du Centre national de prévention du crime.

Excerpt #6.
CPFTT8
ASL: 24.75
3 SYLL:5
4+ SYLL: 3
ICA score: (L + M) x 0.4 = 11.1

De quoi s'agit-il? Les Services aux régions du Nord ont créé un réseau de livraison pour l'ensemble du Nord qui offre des services postaux reliant les trois océans et qui de plus en plus se conforme à la règle de la plus courte distance qu'appliquent d'autres réseaux. Les plus âgés se rappelleront sans doute des sacs postaux déposés juste à côté du Twin Otter et qui étaient transportés dans le qamutiik tiré par un tracteur de la coopérative, pour ensuite être empilés devant la cabane qui servait de bureau de poste local. Pendant des années, ces sacs étaient bleus.

Excerpt #7
CPFTT8
ASL: 25
3 SYLL:13
4+ SYLL: 4
ICA score: (25 + 4) x 0.4 = 11.6

Et pour les habitants du Nord, cette couleur était devenue symbole de promesse. Ces sacs leur apportaient les articles nécessaires à leur survie, c'est-à-dire les médicaments, la nourriture et les chèques, les vêtements et le nécessaire de chasse ainsi que les pièces de rechange.
La rapidité, la fréquence et la fiabilité des services offerts par Postes Canada aux régions du Nord et sa capacité de desservir chaque personne même dans les plus petites collectivités sont autant de questions qui ont depuis longtemps revêtu une importance première pour les habitants du Nord. Voici un exemple de la contribution que nous avons...

Excerpt #8
CCFTT10
ASL: 25
3 SYLL: 17
4+ SYLL: 12
ICA score: (25 + 12) x 0.4 = 14.8
Le Centre canadien d'architecture (CCA), de concert avec l'Institut royal d'architecture du Canada et le Conseil canadien des écoles universitaires d'architecture, a choisi Melvin Charney pour représenter le Canada. Le Conseil des Arts du Canada et le ministère des Affaires étrangères et du Commerce international avaient chargé le CCA d'organiser le processus de sélection.
La participation du Conseil à la Biennale de Venise témoigne de la volonté du Conseil de renforcer ses relations avec le milieu canadien de l'architecture. En collaboration avec un comité consultatif composé de membres du domaine de l'architecture, le Service des arts visuels du Conseil a...

Excerpt #9
CCFTT10
ASL: 20
3 SYLL:14
4+ SYLL: 10
ICA score: (L + M) x 0.4 = 12
...commencé la révision des programmes et des prix qu'il réserve à l'architecture. Photo: L'architecte canadien Melvin Charney représentera le Canada lors de la Septième Biennale de l'architecture de Venise. Diffusion du message de la Commission du droit de prêt public à l'étranger. Pour donner suite à la Conférence internationale sur le droit de prêt public dont elle a été l'hôte l'automne dernier, la Commission du droit de prêt public organise un séminaire sur le droit de prêt public lors de la réunion de la Fédération internationale des associations de bibliothécaires (FIAB), qui a lieu à Jérusalem en août 2000. Paul Whitney, directeur de la bibliothèque publique...

Excerpt #10
PWFTT1 & CCFTT10
ASL: 33.33
3 SYLL:15
4+ SYLL: 10
ICA score: (33.33 + 10) x 0.4 = 17.33
Un trésor à découvrir afin de permettre aux visiteurs de s'informer des travaux de rénovation en cours, afin de pouvoir faire une visite guidée en trois dimensions des édifices du Parlement. Grâce à l'annuaire électronique des programmes et des services, le public pourrait accéder en ligne aux coordonnées exactes...
...de Burnaby et nouveau représentant de l'Association canadienne des bibliothèques au sein de la Commission, représentera le Canada lors de ce séminaire. Cet événement a pour objectif de forti-fier les relations entre les écrivains et les bibliothèques, qui constituent un élément indispensable à la survie du droit de prêt public.

Excerpt #11
PWFTT1 & PWFTT9
ASL: 16.83
3 SYLL:11
4+ SYLL: 9
ICA score: (16.83 + 9) x 0.4 = 10.33
...et à jour sur les personnes-ressources responsables des différents programmes et services offerts par le gouvernement du Canada. Pour l'instant, la population canadienne peut trouver ces renseignements uniquement dans les pages bleues des annuaires téléphoniques publics.
Vous avez peut-être vu notre logo lorsque vous vous êtes arrêté pour examiner un nouvel immeuble en construction ou vous l'avez peut-être aperçu sur une annonce de biens excédentaires du gouvernement publiée dans les journaux. Vous l'avez aussi peut-être aperçu

lors de votre visite sur la Colline du Parlement. Qui sommes-nous? Nous sommes Travaux publics et Services gouvernementaux Canada et nous sommes là pour vous.

Excerpt #12
PWFTT9
ASL: 20
3 SYLL:11
4+ SYLL: 6
ICA score: (20 + 6) x 0.4 = 10.4
Le Ministère assure la prestation de services communs au gouvernement du Canada et emploie près de 12 000 personnes. Nous offrons des services et des programmes par le biais de bureaux situés dans tout le pays, sans oublier aux États-Unis et en Europe. TPSGC joue de nombreux rôles. Dans cette partie de notre site, vous en découvrirez davantage sur notre travail visant à restaurer les édifices du Parlement. Vous pouvez obtenir de l'information sur ce que nous avons à vendre dans le site Biens de la Couronne, que ce soient des chaises des bureaux, des outils électriques ou des véhicules.

Excerpt #13
CPFTT3
ASL: 33.66
3 SYLL: 11
4+ SYLL: 10
ICA score: (33.66 + 10) x 0.4 = 17.46
S'il est vrai que notre succès a fait notre réputation, c'est sans conteste à notre maîtrise de l'innovation que nous le devons. PCIL offre une panoplie exceptionnelle de services de consultants, de solutions postales axées sur la technologie, notamment en matière de transformation postale, ainsi que des possibilités de commerce électronique basées sur ces mêmes systèmes et ces mêmes services qui ont propulsé la Société canadienne des postes à l'avant-garde dans son domaine. Grâce à notre savoir-faire et à la collaboration incessante que nous ne manquerons pas d'entretenir avec votre personnel et vos autres partenaires locaux, nous nous faisons fort d'accroître ...

Excerpt #14
CPFTT3
ASL: 14.28
3 SYLL:17
4+ SYLL: 11
ICA score: (14.28 + 11) x 0.4 = 10.11
...l'efficacité de votre exploitation postale au-delà de ce que vous auriez cru possible. Un portefeuille à la mesure de vos besoins particuliers! Aujourd'hui, la vaste panoplie des prestations de PCIL renferme des solutions technologiques, des méthodes de gestion, des formules de produits et de services et un savoir-faire postal adaptables à toute situation et entièrement aménageables. Profitez de l'appui des meilleurs dans leur domaine. Son savoir-faire, PCIL le doit au succès de la société-mère, la Société canadienne des postes, dans le réaménagement et l'amélioration de ses propres systèmes et services. Projets Complété. Voici quelques-uns des projets parmi la liste impressionnante ...

Excerpt #15
CPFTT4 & NRFTT4
ASL: 51.5
3 SYLL: 24
4+ SYLL: 9
ICA score: (51.5 + 9) x 0.4 = 24.2
...PCIL a déjà eu l'honneur d'apporter son concours...
Les objectifs de l'Atelier rural national consistaient à fournir aux Canadiens et Canadiennes des régions rurales l'occasion : de valider les observations exprimées dans le cadre du Dialogue rural à ce jour; de discuter des solutions aux questions principales cernées par la population rurale; de suggérer des mesures et recommandations concrètes à propos de la façon dont le gouvernement fédéral et d'autres intervenants peuvent conclure des partenariats afin de

contribuer au développement des collectivités rurales; et de participer à l'élaboration d'éléments et de principes directeurs clés qui pourraient servir de fondements à une politique rurale fédérale.

Appendix X: Tests of Statistical Significance

NORMALIZATION alternative (research) hypothesis:
That p1-p2>0 and that the proportion for the non-translated corpus is larger.

Outcome, English corpus (all texts):

We can reject the null hypothesis and accept the alternative hypothesis with a confidence level of 97%.

Observed using WordSmith 3	EST	ETT
Attested words	**61,664**	**60,415**
Coinages	**35**	**20**
Total	61699	60435
Proportion 1	0.000567	
Proportion 2	0.000331	
Difference	0.000236	
Pooled proportion	0.00045	
Standard error	0.000121	
z	**1.945487**	
p-value	0.025858	<0.03

Outcome, French corpus:

We accept the alternative hypothesis with a confidence level of approximately 90%.

Observed using WordSmith 3	FST	FTT
Attested words	**62,711**	**60,790**
Coinages	**14**	**8**
Total	62725	60798
Proportion 1	0.000223	
Proportion 2	0.000132	
Difference	9.16E-05	
Pooled proportion	0.000178	
Standard error	7.59E-05	
z	**1.206058**	
p-value	0.113897	<0.115

EXPLICITATION Alternative (research) hypothesis:

p1-p2<0, and that the proportion for the translated texts is larger.

Outcome, English corpus (all texts):

We can reject the null hypothesis and accept the alternative with a confidence level of better than 99%.

Reporting Verb + *that*		
Observed	EST	ETT
Reporting verb + 0	**511**	**331**
Reporting verb + *that*	**182**	**204**
Total reporting verbs	693	535
Proportion 1	0.2626263	
Proportion 2	0.3813084	
Difference	-0.118682	
Pooled proportion	0.3143322	
Standard error	0.0267182	
z	**-4.441998**	
p-value	4.456E-06	<0.0001

Outcome, English corpus (non-specialized texts only):

We accept the alternative with a confidence level of better than 99%.

Reporting Verb + *that*		
Observed	EST	ETT
That (all)	**83**	**97**
Reporting verb + THAT	**14**	**35**
Proportion 1	0.168675	
Proportion 2	0.360825	
Difference	-0.19215	
Pooled proportion	0.272222	
Standard error	0.066554	
z	**-2.88714**	
p-value	0.001944	<0.0002

Outcome, English corpus (all texts):

We accept the alternative hypothesis (confidence level of approximately 90%).

***that* ODC**		
Observed	EST	ETT
that (not ODC)	**528**	**712**
that ODC	**11**	**23**
total (all *that*)	539	735
Proportion 1	0.020408	
Proportion 2	0.031293	
Difference	-0.01088	
Pooled proportion	0.026688	
Standard error	0.00914	
z	**-1.1909**	
p-value	0.116847	<0.12

Outcome, English corpus (non-specialized texts only):
We retain the null hypothesis (confidence level of approximately 56%).

***that* ODC**			
Observed	EST		ETT
that (all)	**83**		**97**
that ODC	**3**		**4**
Proportion 1	0.036145		
Proportion 2	0.041237		
Difference	-0.00509		
Pooled proportion	0.038889		
Standard error	0.028908		
z	**-0.17617**		
p-value	0.430082	<0.44	

Outcome, English corpus (all texts):
We accept the alternative hypothesis; confidence level of better than 99%.

***which* ODC**			
Observed	EST		ETT
which zero	**95**		**105**
which ODC	**7**		**46**
total (all *which*)	102		151
Proportion 1	0.068627		
Proportion 2	0.304636		
Difference	-0.23601		
Pooled proportion	0.209486		
Standard error	0.052156		
z	**-4.52505**		
p-value	3.02E-06	<0.0001	

Outcome, English corpus (non-specialized texts only):
We accept the alternative hypothesis with a confidence level of approximately 97%.

***which* ODC**			
Observed	EST		ETT
which (all)	**23**		**27**
which ODC	**0**		**4**
Proportion 1	0		
Proportion 2	0.148148		
Difference	-0.14815		
Pooled proportion	0.08		
Standard error	0.07698		
z	**-1.9245**		
p-value	0.027146	<0.03	

Outcome, French corpus (all texts):
Retain the null hypothesis despite weak evidence supporting research hypothesis, because the confidence level of about 70% is too low.

***Ne* explétif**		
Observed	FST	FTT

Ne zero	**269**	**244**
Ne ODC	**2**	**3**
total (all *ne*)	271	247
Proportion 1	0.00738	
Proportion 2	0.012146	
Difference	-0.00477	
Pooled proportion	0.009653	
Standard error	0.008601	
z	**-0.55409**	
p-value	0.289759	<0.29

Outcome, French corpus (non-specialized texts only):
Retain the null hypothesis despite weak evidence supporting research hypothesis, because the confidence level of about 54% is too low.

***Ne* explétif**		
Observed	FST	FTT
ne all	**9**	**16**
ne explétif	**1**	**2**
Proportion 1	0.111111	
Proportion 2	0.125	
Difference	-0.01389	
Pooled proportion	0.12	
Standard error	0.135401	
z	**-0.10258**	
p-value	0.45915	<0.46

Outcome, French corpus (all texts): Retain the null hypothesis.

L'on		
Observed	FST	FTT
On zero	**41**	**256**
On ODC	**7**	**14**
total (all *on*)	48	270
Proportion 1	0.145833	
Proportion 2	0.051852	
Difference	0.093981	
Pooled proportion	0.066038	
Standard error	0.038902	
z	**2.415853**	
p-value	0.992151	>0.5

Outcome, French corpus (non-specialized texts only):
Retain the null hypothesis.

L'on		
Observed	FST	FTT
on all	**13**	**14**
L'on	**3**	**1**
Proportion 1	0.230769	
Proportion 2	0.071429	
Difference	0.159341	
Pooled proportion	0.148148	

Standard error	0.136828	
z	**1.164529**	
p-value	0.877895	>0.5

SIMPLIFICATION (Type/token ratio and Lexical Density)

Alternative (research) hypothesis: p1-p2<0, and that the proportion for the translated texts is smaller.

Outcome, English corpus (all texts):

Accept the alternative hypothesis with a confidence level of better than 99%.

Type/token ratio			
Observed	EST	ETT	
English	78.59		74.63
Proportion 1	0.7859		
Proportion 2	0.7463		
Difference	0.0396		
Pooled proportion	0.766304916		
Standard error	0.002421924		
z	**16.35063512**		
p-value	0	<0.0001	

Outcome, French corpus (all texts): Retain the null hypothesis.

Type/token ratio			
Observed	FST	FTT	
French	77.32		79.03
Proportion 1	0.7732		
Proportion 2	0.7903		
Difference	-0.0171		
Pooled proportion	0.781755		
Standard error	0.00237		
z	**-7.21586**		
p-value	1	>0.5	

Outcome, English corpus (all texts):

Accept the alternative hypothesis with a confidence level approaching 100%.

Lexical density			
Observed	EST	ETT	
English	61.2		58.56
Proportion 1	0.612		
Proportion 2	0.5856		
Difference	0.0264		
Pooled proportion	0.000981		
Standard error	0.000179		
z	**147.382**		
p-value	0	<0.0001	

Outcome, French corpus (all texts): Retain the null hypothesis.

Lexical density		
Observed	FST	FTT

French	48.66		51.34
Proportion 1	0.4866		
Proportion 2	0.5134		
Difference	-0.0268		
Pooled proportion	0.000823		
Standard error	0.000165		
z	**-162.908**		
p-value	1	>0.5	

SIMPLIFICATION (Mean sentence length)

Alternative (research) hypothesis: mean_1-mean_2>0, and the mean sentence length is shorter for the translated corpus.

Outcome, English corpus (all texts): Retain the null hypothesis.

Mean sentence length			
Observed	EST		ETT
English	21.36		24.73
Sample mean 1	21.36		
Sample mean 2	24.73		
Difference	-3.37		
Standard error	0.433543061		
z-value	-7.773160977		
p-value	1	>0.5	

Outcome, French corpus (all texts):

Accept the alternative hypothesis with a confidence level approaching 100%.

Mean sentence length			
Observed	FST		FTT
French	25.96		22.91
Sample mean 1	25.96		
Sample mean 2	22.91		
Difference	3.05		
Standard error	0.511344		
z-value	5.964675		
p-value	1.23E-09	<0.0001	

LEVELLING-OUT (Variance of readability index scores)

Alternative (research) hypothesis: (variance_1 / variance_2)>1 , and there is lower variance in the scores of the translated corpora.

Outcome, English corpus: For the Fry and Gunning indices, we retain the null hypothesis. For Lix, we accept the alternative hypothesis with a confidence level of 91%.

Variance			F-value	**p-value**
Observed	EST Sd	ETT Sd		
Fry index	3.26	3.19	1.044369	0.467054
Gunning index	2.67	4.25	0.39468	0.959131
Lix index	5.81	4.53	1.644962	0.098032

Outcome, French corpus: For all three indices, we retain the null hypothesis.

Variance			F-value	**p-value**
Observed	FST Sd	FTT Sd		
Henry index	0.94	2.04	0.212322	0.959404
Lix	5.07	10.13	0.250494	0.999759
ICA	4.52	7.32	0.381289	0.964312